

Doctoral thesis in Computational Linguistics

Computational models of language and vision

Studies of neural models
as learners of multi-modal knowledge

Nikolai Ilinykh

June 2024

Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)



UNIVERSITY OF
GOTHENBURG



Doctoral Thesis in Computational Linguistics,
University of Gothenburg, Gothenburg, Sweden, 2024

Computational Models of Language and Vision: Studies of Neural Models as
Learners of Multi-modal Knowledge

© Nikolai Ilinykh, 2024

ISBN 978-91-8069-767-5 (print)

ISBN 978-91-8069-768-2 (pdf)

The research reported in this thesis was supported by a grant from the
Swedish Research Council (VR project 2014-39) for the establishment of the
Centre for Linguistic Theory and Studies in Probability (CLASP) at the
University of Gothenburg.

Cover design: İrfan Meriç

Photographer: Monica Havström

Printed by Stema Specialtryck AB, Borås, Sweden, 2024

Publisher: University of Gothenburg (Dissertations)

Distribution: Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Box 100, SE-405 30, Gothenburg, Sweden

Part I of this thesis is also available in full text at:

<http://hdl.handle.net/2077/80949>

Abstract

This thesis develops and evaluates computational models that generate natural language descriptions of visual content. We build and examine models of language and vision to gain a deeper understanding of how they reflect the relationship between the two modalities. This understanding is crucial for performing computational tasks. The first part of the thesis introduces three studies that inspect the role of self-attention in three different self-attention blocks of the object relation transformer model. We examine attention heatmaps to understand how the model connects different words, objects, and relations within the tasks of image captioning and image paragraph generation. We connect our interpretation of what the model learns in self-attention weights with insights from theories about human cognition, visual perception, and spatial language. The three studies in the second part of the thesis investigate how representations of images and texts can be applied and learned in task-specific models for image paragraph generation, embodied question answering, and variation in human object naming. The last two studies in the third part examine properties of human-generated texts that multi-modal models are expected to acquire in image paragraph generation as well as perceptual category description and interpretation tasks. We analyse discourse structure in image paragraphs produced with different decoding methods. We also inspect whether models of perceptual categories can abstract from visual representations and use this knowledge to generate descriptions that exhibit discriminativity levels important for the task. We show how automatic measures for evaluating text generation behave in a comparison of model-generated and human-generated image descriptions. This thesis presents several contributions. We illustrate that, under specific modelling conditions, self-attention can capture information about the relationship between objects and words. Our results emphasise that the specifics of the task determine the

manner and context in which different modalities are processed, as well as the degree to which each modality contributes to the task. We demonstrate that while favoured by automatic evaluation metrics in different tasks, machine-generated image descriptions lack the discourse complexity and discriminative power that are often important for generating better, human-like image descriptions.

Sammanfattning

Denna avhandling utvecklar och utvärderar datormodeller som genererar beskrivningar i naturligt språk av visuellt innehåll. Vi bygger och undersöker modeller av språk och seende för att få en djupare förståelse för hur de reflekterar relationen mellan de två modaliteterna. Denna förståelse är avgörande vid utförande av olika uppgifter. Avhandlingens första del introducerar tre studier som undersöker vilken roll självuppmärksamhet (self-attention) spelar i tre olika självuppmärksamhetsblock i transformermodellen för objektrelationer. Vi undersöker uppmärksamhetskartor (attention heatmaps) för att förstå hur modellen kopplar samman olika ord, objekt och relationer vid bildtextning och bildparagrafgenerering. Vi kopplar ihop vår tolkning av vad modellen lär sig i självuppmärksamhetsvikterna med insikter från teorier om mänsklig kognition, visuell perception och spatialt språk. De tre studier i avhandlingens andra del undersöker hur multimodala representationer av bilder och texter kan tillämpas och läras i uppgiftsspecifika modeller för bildparagrafgenerering, förkroppsligat frågebesvarande och variation i mänsklig objektbenämning. De två sista studierna i avhandlingens tredje del undersöker egenskaper hos mänskligt framställda texter som multimodala modeller förväntas förvärva vid bildparagrafgenerering samt beskrivning och tolkning av perceptuella kategorier. Vi analyserar diskursstrukturer i bildparagrafer skapade med olika avkodningsmetoder. Vi undersöker också huruvida modeller av perceptuella kategorier kan abstrahera från visuella representationer och använda denna kunskap för att generera beskrivningar som kan diskriminera på nivåer som är viktiga för uppgiften. Vi visar hur automatiska åtgärder för att utvärdera textgenerering beter sig i en jämförelse av modellgenererade och mänskliga genererade bildbeskrivningar. Avhandlingen presenterar flera bidrag. Vi visar att självuppmärksamhet under specifika modelleringsförhållanden kan reflektera information om förhållandet mellan

objekt och ord. Våra resultat indikera att en uppgifts specifika utformning avgör på vilket sätt och i vilken kontext olika modaliteter bearbetas, samt i vilken utsträckning varje modalitet bidrar till uppgiften. Vi demonstrerar att medan datorgenererade bildbeskrivningar favoriseras av automatiska utvärderingsmått vid olika uppgifter, saknar de den diskurskomplexitet och diskriminativa kraft som ofta är viktig för att generera bättre och mer mänskliga bildbeskrivningar.

Acknowledgements

My very first and foremost thanks goes to Simon. *Simon*, thank you for being such a great scientific advisor, supporter, friend. I believe supervising me was not entirely easy and still, you were there for me all the way. Our discussions were long and often heated, but their intensity and your willingness to debate with me have sparked my curiosity even more and shaped me to be a better researcher than I ever was. Thank you for teaching me how to write and think as a scientist. I have more to learn, but with the knowledge and expertise you shared with me, I am confident that I will not be lost.

Asad, I thank you for your support and encouragement along the way. Your scientific rigor combined with plasticity and flexibility is something I always aimed for and hopefully I am on the right track. You might not know this, but I learned so much from you about how to navigate science and how to be a part of the larger scientific community.

Stella, thank you so, so much for being the best opponent during my final seminar. Your feedback and help were invaluable. *Letitia* and *Bill*, thank you for your valuable comments on the versions of this thesis.

I want to thank all the people at CLASP who were always so kind, understanding, and patient with me. Special thanks goes to Shalom and Sharid for making CLASP an amazing research environment. *Bill*, thank you for helping me make sense of the chaos that I might have (accidentally) caused in research or in life. *Aram*, thank you for constantly reminding me, a typical workaholic, that there are so many important things outside of academia. *Jean-Philippe*, thank you for the support with the research project that turned into papers. *Alex*, thank you for helping me with the Swedish translation of the abstract in this thesis. People at FLoV, I am so very grateful to you. I specifically would like to thank the administration at the department for supporting me throughout this journey. *Iines*, thank you for being a mental

and emotional supporter every time (!) I turned to you for help. *Eleni*, since the moment I saw you in Gothenburg for the first time, I knew that I would have great times knowing that you are around. Thank you so much for keeping me in check and, perhaps, unknowingly, making me feel listened to.

I want to thank those who were there when I only had my first baby research steps. *David*, thank you for teaching me so many things about research. Writing papers and chatting over Slack while listening to some nice music has never been better. *Sina*, I am very lucky to know you and to have ever worked with you. Thank you for seeing a human in me even when the times are hard. I also want to thank *Nazia*, *Ronja*, *Sole*, *Simeon*, and other people from Bielefeld University and CITEC who made my time there great and full of experiences.

My Gothenburg friends and Swedish friends, thank you so much! A special thanks goes to friends who helped me feel better when the pandemic hit and I was still new in Gothenburg. *Kostas*, *Kate*, *Tova*, *Saga*, you cannot imagine how much spending time with you helped me in the first years of my doctoral studies. *Hana*, *Pierluigi*, *Hadi*, thank you for being so friendly with me and for bringing the good times. *Irfan*, thank you for being one of my biggest supporters and someone who I can turn to in moments of doubt. You are very special.

Thomas and *Sardana*, thank you for being the best conference buddies ever and for some of the unforgettable times and experiences we had together.

Thank you to my friends in Perm and Kosa. *Polina*, *Christina*, *Lyuba*, *Yulia*, *Olya*, and *Misha*, as time goes on it might be harder to keep in touch, but you should know that you were (and are!) the best. TIII!

Thank you to my friends who I met in Germany, thank you to Prime people! *Anna*, *Vlad*, *Olya*, *Kolya*, *Lesha*, and *Marina*, I am happy to know you and share so many experiences with you. You always took me out of my comfort zone, challenged me, and were there for me during all these years we have known each other.

It is very easy to get overwhelmed and forget to mention someone who helped me during this journey. So, here I want to thank all others who have been there for me.

Мама и Папа, I love you so much. Your unconditional love and support have always been my strongest foundation. I want to say that *we* did this together. We never had the resources, convenient conditions, or money, yet we made it on our own. Я вас очень сильно люблю.

Declaration

I hereby declare that the research presented in this doctoral thesis is the result of my own work and has not been submitted to any other degree at the University of Gothenburg or any other institution.

Contents

1	Introduction	1
2	Research questions	4
3	Motivation	10
3.1	An example of humans performing a multi-modal task	10
3.1.1	World knowledge	10
3.1.2	Perceptual knowledge	13
3.1.3	Knowledge of intents	15
3.2	An example of a model performing a multi-modal task	16
4	Background and methodology	18
4.1	General technical background	18
4.2	Language-and-vision natural language processing	29
5	Summaries of studies	45
5.1	Part I: The role of self-attention in object relation transformer	45
5.1.1	Motivation	45
5.1.2	Study I: How Vision Affects Language	52
5.1.3	Study II: What Does a Language-And-Vision Trans- former See	56
5.1.4	Study III: Attention as Grounding	60
5.2	Part II: Representation learning for language-and-vision tasks	64
5.2.1	Motivation	64
5.2.2	Study IV: When an Image Tells a Story	72
5.2.3	Study V: Look and Answer the Question	76
5.2.4	Study VI: Context matters in object naming	80
5.3	Part III: Task-specific evaluation of model-generated image descriptions	85

5.3.1	Motivation	86
5.3.2	Study VII: Do Decoding Algorithms Capture Dis- course Structure in Multi-Modal Tasks?	89
5.3.3	Study VIII: Describe Me an Auklet	95
6	Studies	102
6.1	How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer	102
6.1.1	Abstract	102
6.1.2	Introduction	102
6.1.3	Model	104
6.1.4	Learning syntactic knowledge	108
6.1.5	Multi-modality and masked self-attention	112
6.1.6	Attention Alignment	117
6.1.7	Conclusion	119
6.2	What does a Language-and-Vision Transformer See: The Im- pact of Semantic Information on Visual Representations	121
6.2.1	Abstract	121
6.2.2	Introduction	122
6.2.3	Materials and Methods	127
6.2.4	Experiments	135
6.2.5	Discussion and Implications	161
6.2.6	Conclusion	168
6.3	Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer	169
6.3.1	Abstract	169
6.3.2	Introduction	169
6.3.3	Experimental Set-Up	171
6.3.4	Methods and Metrics	174
6.3.5	Linking Nouns and Objects	177

6.3.6	Experiments and Results	180
6.3.7	Conclusion	187
6.4	When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions	190
6.4.1	Abstract	190
6.4.2	Introduction	190
6.4.3	Approach	193
6.4.4	Experiments and Evaluation	199
6.4.5	Related Work	205
6.4.6	Conclusion	207
6.5	Look and Answer the Question: On the Role of Vision in Embodied Question Answering . . .	210
6.5.1	Abstract	210
6.5.2	Introduction	210
6.5.3	Task Description	213
6.5.4	Is language really stronger in EQA?	215
6.5.5	“How much” vision is required?	218
6.5.6	EQA: biases and limitations	219
6.5.7	Conclusion	220
6.5.8	Baseline QA Model	221
6.5.9	Image Rendering Problem	221
6.5.10	Colour Problem	224
6.5.11	Example Episode	224
6.6	Context matters: evaluation of target and context features on variation of object naming	227
6.6.1	Abstract	227
6.6.2	Introduction	227
6.6.3	Problem formulation	231
6.6.4	Model	237
6.6.5	Evaluation metrics	238

6.6.6	Results	242
6.6.7	Conclusions	246
6.6.8	Fusing features	248
6.6.9	Representing language for Context-Scene	248
6.7	Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation	250
6.7.1	Abstract	250
6.7.2	Introduction	250
6.7.3	On the importance of decoding	252
6.7.4	Task and model	253
6.7.5	Decoding algorithms	255
6.7.6	Linking	259
6.7.7	Automatic evaluation	260
6.7.8	Human evaluation	263
6.7.9	Non-grounded evaluation	264
6.7.10	Grounded evaluation	266
6.7.11	Attentional structure of discourse	269
6.7.12	Conclusion	269
6.7.13	Limitations	270
6.7.14	Appendix A	270
6.8	Describe Me an Auklet: Generating Grounded Perceptual Category Descriptions	272
6.8.1	Abstract	272
6.8.2	Introduction	272
6.8.3	Background	275
6.8.4	Models	278
6.8.5	Experiments	285
6.8.6	Results	288
6.8.7	Discussion and conclusion	289

6.8.8	Limitations	291
7	Conclusions and discussion	293
7.1	What have we learned from studies?	293
7.2	Discussion	298
7.2.1	Conclusion I: on the role of self-attention	298
7.2.2	Conclusion II: on the role of multi-modal representa- tions	300
7.2.3	Conclusion III: on the quality of generated descriptions	301
7.2.4	General conclusion and future work	302
	Bibliography	304

Chapter 1: Introduction

Developing a computational model that understands the visual world and can use language to describe it or execute actions has been an important challenge for researchers in artificial intelligence and natural language processing. The social need for such systems is clear; for example, they can assist the elderly with physical tasks e.g., bringing a fork from the kitchen¹. However, building such a system is challenging as it must have specific skills to be able to bring the fork from the kitchen to a person. First, the agent has to be embodied, it needs to have sensors and actuators to interact with the environment. Then, it needs to have the ability to recognise objects in the kitchen, understand what a fork is, identify the fork, and navigate towards it. Each of these tasks is a specific *downstream task* and researchers often focus on modelling these tasks individually.

We focus on computational tasks that share a common denominator: they all require models to operate with two modalities, *language* and *vision*, and *the computational processing* of these two is the central backbone of this thesis. There are many other modalities such as sound or speech, but this thesis does not explore them and assumes that “multi-modal” here means a combination of linguistic and visual information. While we do not present a model with all possible modalities, our work offers models and their analysis that work with *two* of them in the context of language-and-vision tasks. There are many computational language-and-vision tasks, for example, – in increasing order of complexity –, image description generation (Bernardi et al., 2016), visual question answering (Antol et al., 2015), and visual dialogue (Das,

¹An argument in favour of building systems that have visual and linguistic abilities has been proposed by us here: <https://spraakbanken.gu.se/en/news-and-events/conferences-and-workshops/sustainable-language-representations/position-statements>

Kottur, et al., 2017; De Vries et al., 2017; Dobnik and Silfversparre, 2021; Haber et al., 2019; Ilinykh, Zarrieß, et al., 2019a). The more complex computational tasks often assume proficiency in easier tasks, albeit each task has its own set of requirements to be met. In image description generation the model must choose how and what to describe in images. When models are answering questions about images, they should focus on relevant parts of the image and be guided by the question that directs their attention.

We investigate the processing of image and text representations by computational models, examine configuration options for these representations, and offer recommendations for constructing systems capable of performing multi-modal tasks which require such representations. In the first part of the thesis we work with a transformer-based model (Vaswani et al., 2017) called object relation transformer (Herdade et al., 2019). Transformer-based models have been adopted in many tasks that are either linguistic or visual. Here we build and use these models in the context of *multi-modal tasks* and study the behaviour of self-attention in them. We focus on two tasks: image captioning and image paragraph generation. The second part of the thesis explores the application of representations from pre-trained models like DenseCap (Johnson et al., 2016) and CLIP (Radford, Kim, et al., 2021) in the context of downstream tasks such as image paragraph generation and variation in human object naming. As visual features we encode information from bounding boxes of objects in images. As textual features we use embeddings of object labels. We also examine how models use representations of images and questions learned from scratch in the context of embodied question answering task. The third part of the thesis inspects the output of transformer-based image description models for image paragraph generation and performance of the models in perceptual category description generation and interpretation tasks. The latter task defines a perceptual category as a combination of features that determine the membership of a specific object in this category, e.g. instances of ravens form a “raven” category. For the image paragraph generation task, we analyse

the discourse structure in texts that are generated by humans and examine if models replicate this structure in their descriptions of images. We also study the extent to which the scores from automatic measures evaluating text generation reflect information about the discourse structure in image paragraphs. We also examine how transformer-based models describe and interpret categories with either visual features of images or instances from these categories or learned category representations. We study the discriminativity in model-generated descriptions of perceptual categories and inspect how descriptions with different levels of discriminativity are useful for the task of perceptual category interpretation.

The studies that we present focus on building and evaluating a range of computational architectures for modelling human language known as “language models”. These models are in part responsible for recent technological developments in natural language processing and AI (Bommasani et al., 2021). Most of the research described in this thesis has been conducted during a period when talking about language models became synonymous with discussing “artificial intelligence” in the eyes of both the general public and research communities. This is not surprising because language models today perform well in many tasks; they are commonly used for editing, searching, summarising, and so on. Our studies are timely because language models are being introduced at such a fast rate and used for purposes that can involve critical decisions. The same goes for the design and use of systems in further research: models are often chosen based on popularity and availability rather than on their performance and *suitability* for the task. Therefore, we need a deeper understanding of the capabilities and shortcomings of these models.

Chapter 2: Research questions

This thesis addresses the following research questions:

1. **Research Question I:** What is the role of self-attention in the multi-modal transformer trained for such image description tasks as image captioning and image paragraph generation? Does such self-attention capture representations and structures which can be linguistically and cognitively interpreted? Three studies in Section 5.1 primarily address this question.
2. **Research Question II:** How can multi-modal representations of objects labels and regions be applied in three different tasks such as image paragraph generation, embodied question answering and variation in human object naming? Do models designed for these three tasks learn from such multi-modal representations? Three studies in Section 5.2 address this question.
3. **Research Question III:** Can multi-modal neural models generate texts with similar discourse structure as human-generated texts in the image paragraph generation task? Can models of perceptual category description and interpretation abstract from visual representations and use this knowledge to generate descriptions that exhibit discriminativity levels that are important for the task? How do model-generated and human-generated texts compare and how do automatic measures for evaluating text generation fare in this comparison? Two studies in Section 5.3 answer these questions.

Chapter 4 introduces the necessary technical background and places this thesis within the current research in multi-modal NLP. Chapter 5 contains

summaries of studies with each of them outlining a motivation for a specific study, key results and questions and implications for future work. We also introduce relevant background that is specific to studies in different parts of this chapter. Chapter 6 includes research papers that have been previously published and which correspond to the core contribution of this thesis. Lastly, Chapter 7 outlines lessons learned from the whole thesis and provides ideas for future work.

Chapter 6 presents published research studies (peer-reviewed) that constitute the core contribution of this thesis. Below I list these studies in the order they appear in this thesis.

- (i) How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer. **Nikolai Ilinykh** and Simon Dobnik. 2021. In Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics. Available at: <https://aclanthology.org/2021.mmsr-1.5/>
- (ii) What Does a Language-And-Vision Transformer See: The Impact of Semantic Information on Visual Representations. **Nikolai Ilinykh** and Simon Dobnik. 2021. *Frontiers in Artificial Intelligence: Identifying, Analyzing, and Overcoming Challenges in Vision and Language Research*, 4, 767971. Available at: <http://dx.doi.org/10.3389/frai.2021.767971>
- (iii) Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. **Nikolai Ilinykh** and Simon Dobnik. 2022. In Findings of the Association for Computational Linguistics: ACL 2022, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics. Available at: <https://aclanthology.org/2022.findings-acl.320/>

- (iv) When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions. **Nikolai Ilinykh** and Simon Dobnik. 2020. In Proceedings of the 13th International Conference on Natural Language Generation, pages 338–348, Dublin, Ireland. Association for Computational Linguistics. Available at: <https://aclanthology.org/2020.inlg-1.40/>
- (v) Look and Answer the Question: On the Role of Vision in Embodied Question Answering. **Nikolai Ilinykh**, Yasmeeen Emampoor, and Simon Dobnik. 2022. In Proceedings of the 15th International Conference on Natural Language Generation, pages 236–245, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics. Available at: <https://aclanthology.org/2022.inlg-main.19/>
- (vi) Context matters: evaluation of target and context features on variation of object naming. **Nikolai Ilinykh** and Simon Dobnik. 2023. In Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing, pages 12–24, Ingolstadt, Germany. Association for Computational Linguistics. Available at: <https://aclanthology.org/2023.limo-1.3/>
- (vii) Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation. **Nikolai Ilinykh** and Simon Dobnik. 2022. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 480–493, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. Available at: <https://aclanthology.org/2022.gem-1.45/>
- (viii) Describe Me an Auklet: Generating Grounded Perceptual Category Descriptions. Bill Noble* and **Nikolai Ilinykh***. 2023. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language

Processing, pages 9330–9347, Singapore. Association for Computational Linguistics. *Equal contribution. Available at: <https://aclanthology.org/2023.emnlp-main.580/>

During my doctoral studies I have been involved in several research projects. These projects are not included in primary contribution of this thesis, but all of them are relevant for the research that the thesis presents. Below I list publications which were created with my colleagues, peer-reviewed and accepted for presentation at various conferences and workshops.

- (i) The VDG Challenge: Response Generation and Evaluation in Collaborative Visual Dialogue. **Nikolai Ilinykh** and Simon Dobnik. 2023. In Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges, pages 23–30, Prague, Czechia. Association for Computational Linguistics. Available at: <https://aclanthology.org/2023.inlg-genchal.4/>
- (ii) Vector Norms as an Approximation of Syntactic Complexity. Adam Ek and **Nikolai Ilinykh**. 2023. In Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), pages 121–131, Tórshavn, the Faroe Islands. Association for Computational Linguistics. Available at: <https://aclanthology.org/2023.resourceful-1.15/>
- (iii) Are Language-and-Vision Transformers Sensitive to Discourse? A Case Study of ViLBERT. Ekaterina Voloshina, **Nikolai Ilinykh**, and Simon Dobnik. 2023. In Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), pages 28–38, Prague, Czech Republic. Association for Computational Linguistics. Available at: <https://aclanthology.org/2023.mmnlg-1.4/>

- (iv) In Search of Meaning and Its Representations for Computational Linguistics. Simon Dobnik, Robin Cooper, Adam Ek, Bill Noble, Staffan Larsson, **Nikolai Ilinykh**, Vladislav Maraev, and Vidya Somashekarappa. 2022. In Proceedings of the 2022 CLASP Conference on (Dis)embodiment, pages 30–44, Gothenburg, Sweden. Association for Computational Linguistics. Available at: <https://aclanthology.org/2022.clasp-1.4/>
- (v) What to refer to and when? Reference and re-reference in two language-and-vision tasks. Simon Dobnik, **Nikolai Ilinykh**, and Aram Karimi. 2022. In Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin. Available at: http://semdial.org/anthology/Z22-Dobnik_semdial_0017.pdf
- (vi) Examining the Effects of Language-and-Vision Data Augmentation for Generation of Descriptions of Human Faces. **Nikolai Ilinykh**, Rafal Černiavski, Eva Elżbieta Sventickaitė, Viktorija Buzaitė, and Simon Dobnik. 2022. In Proceedings of the 2nd Workshop on People in Vision, Language, and the Mind, pages 26–40, Marseille, France. European Language Resources Association. Available at: <https://aclanthology.org/2022.pv1am-1.5/>
- (vii) A General Benchmarking Framework for Text Generation. Diego Mousallem, Paramjot Kaur, Thiago Ferreira, Chris van der Lee, Anastasia Shimorina, Felix Conrads, Michael Röder, René Speck, Claire Gardent, Simon Mille, **Nikolai Ilinykh**, and Axel-Cyrille Ngonga Ngomo. 2020. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), pages 27–33, Dublin, Ireland (Virtual). Association for Computational Linguistics. Available at: <https://aclanthology.org/2020.webnlg-1.3/>

- (viii) The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task: Overview and Evaluation Results (WebNLG+ 2020). Thiago Castro Ferreira, Claire Gardent, **Nikolai Ilinykh**, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics. Available at: <https://aclanthology.org/2020.webnlg-1.7/>

Below I list a couple of position statements and other related research outputs that never made it into full papers but were presented and discussed at the corresponding venues.

- (i) Taking BERT for a walk: on the necessity of grounding, multi-modality and embodiment for impactful NLP. **Nikolai Ilinykh** and Simon Dobnik. 2021. Position statements for Sustainable language representations for a changing world, a workshop at NoDaLiDa 2021. Available at: <https://spraakbanken.gu.se/en/news-and-events/conferences-and-workshops/sustainable-language-representations/position-statements>
- (ii) ChatGPT goes into the physical world: on dangers and future of multi-modal language. **Nikolai Ilinykh**. A talk at “Who’s Responsible for ChatGPT?”. The abstract can be found at: <https://www.philosophyforhumans.com/what-we-ve-done>

Chapter 3: Motivation

3.1. An example of humans performing a multi-modal task

Humans can process information coming either by means of sensory experiences or by means of linguistic communication. Imagine a friend (the describer) who tells you (the listener) what they see in the picture that is displayed in Figure 3.1. They start by saying that they see “a kitchen with cream coloured cabinets”. The second you see this sentence you are immediately prone to mentally visualise the kitchen that they are describing. You are basing this visualisation on both your visual memory and what your friend says. Next you hear that “there is a dog on the floor” and that the floor is “wood coloured”. At this point, if you were given a set of images including the one that is described by your friend and asked to pick that particular image, you know better what to look for. When next your friend says that “there are lot of items on the counters including plants” and “there is a set of cabinet doors open and those have glass panes”, you already have a pretty good idea about the image that your friend observes. Let us say that at this point you make an action and correctly pick the image that your friend talks about. This example illustrates how humans are able to combine linguistic and visual information to discuss concepts and objects in the real world. In its essence, the intricacies of human processing of linguistic and visual information in the example above is what inspires and motivates the **computational research** that this thesis is about.

3.1.1. World knowledge

Generally, the speakers in our example would normally assume that they have a similar knowledge and perception of the world, although they might not



Figure 3.1. Example image described to you by your friend.

necessarily share one. The former is important to understand the common world context of the image, while the latter is necessary to interpret matching between text, image and knowledge, for example, in the case of colours (“cream-coloured”). While both speakers might not be placed in the same physical environment, they can still discuss visual concepts if they have a sufficient level of *common ground* (Clark, 2015; Stalnaker, 2002) Common ground is information that speakers have at any time about what they believe they have agreed on in the interaction so far. Common ground is directly related to a much bigger source of information such as *commonsense thought*, a world knowledge about how “things” are and what can be done with them (Minsky, 2000), Both concepts are related to cognitive aspects of individuals as humans might not necessarily share the same commonsense knowledge or have the same common ground. “Things” that are involved in common-

sense knowledge range in their definition from natural laws of physics to the knowledge about relations between specific objects and the type of actions that can be done to them, e.g. affordances. Such world knowledge can be divided into common knowledge (Cambria et al., 2011) and commonsense knowledge (Davis, 2017). Common knowledge is the knowledge about the world that is often implicitly expressed in human communication. This often includes commonly known facts about the world and scientific knowledge such that the Earth orbits the Sun or that carved pumpkins are used as Halloween decorations in many countries. Also, common ground assumes shared culture. For example, the arrangement of objects in the kitchen in Figure 3.1 is immediately recognisable by many (because of the world history and globalisation processes), but an outdoor kitchen in an Ethiopian village might not be that recognisable. While common knowledge varies between cultures and different regions of the world, commonsense knowledge is assumed to be approximately the same between all humans. This knowledge, contrary to common knowledge, is rarely mentioned in human-human communication, but it is relied on by humans as it helps to achieve common ground (Chai et al., 2018).

Shared world knowledge is pivotal to human-human communication as it is something that gives us a lot of prior information for more efficient communication. For example, cabinet doors are often expected to appear in the kitchens. When cabinets are mentioned in the example in Section 3.1, it is likely to ease mental processing of the information rather than have no effect or complicate it. Neither of the humans in that example had to engage in the conversation about how kitchens and cabinets relate to each other as this is something that is taken care of by the common knowledge that both share. But there is much more knowledge that a simple string of symbols “kitchen” can evoke in us. Specifically, this is the knowledge of how the world is structured and what type of hierarchies exist in our interpretation of the world. For example, speakers might know that cabinets

and plants are typical for kitchens, but dogs are not. However, cabinets and plants might appear in other types of rooms such as bedrooms. On a more general level, bedrooms, in turn, are related to kitchens because both of them commonly appear as parts of the house. Relating concepts to each other and building cognitive structures and hierarchies about the world is crucial for learning about the world (Botvinick, 2008; Cooper, 2023; Tenenbaum et al., 2011). Such type of knowledge has been constructed computationally with resources such as WordNet (Fellbaum, 1998; Miller, 1995), FrameNet (Baker et al., 1998) or SUMO ontology (Niles and Pease, 2001). The primary benefit of these hierarchies in human-human communication is that they allow us to reuse acquired information and structures to learn novel concepts. Such hierarchical organisation of knowledge is yet another important pillar of the world knowledge that humans have and often share.

3.1.2. Perceptual knowledge

Humans can identify real-world objects even when they are represented as images. But how do they interpret pixels into “plants”, “dogs” and “kitchens”? Here we discuss relevant perceptual knowledge that humans use and associate with their knowledge of the world in order to talk about this exact world. Mapping linguistic symbols to perceptual stimuli, be it an object or some of object’s representation (“grounding”) is an important aspect of human-human interaction (Harnad, 1990). Connecting cognitive representations with those of the world is also important for our perception of space and its structuring in our minds (Levinson, 2003; Miller and Johnson-Laird, 1976; Talmy, 1983; Talmy, 2000) along with the evolution and development of our language (Perniss and Vigliocco, 2014). To describe images, we need not only to identify physical objects, relations and events but also associate them with words that we have to express what we want to say about the image. The example in Section 3.1 has descriptions of multiple objects such as “plants” or “a dog” and a few relations, e.g. “a dog *on* the floor”. An interesting detail

here to consider is the choice of words to describe objects. Humans tend to first name objects at their basic categorical level, e.g. dogs and cats, but not mammals (Rosch et al., 1976). Here the concept of basic category is a technical term that refers to a human choice of categorising and naming objects based the most common and shared features that can also be easily understood by others. Unless there is a specific intent or context of the situation in which a different description needs to be produced, humans are likely to pick such basic category descriptions for objects. However, if the our describer is a dog lover and has extensive knowledge about dogs, they might mention the dog's breed.

Relations between objects and spatial relations are recognised in the context of speakers' knowledge about their geometric positions and functional knowledge (Coventry and Garrod, 2004), with different relations relying differently on different sources of information (Garrod et al., 1999). The difference between functional and geometric knowledge here is especially important as it would be definitive in the choice of the words to describe relations. For example, if the speaker decides to talk about the black jar and red tomatoes located to the right of the oven, they would use their knowledge about *function* of the jar to have things inside and say that tomatoes are “in” the jar and not “on top of” the jar despite the fact that geometrically tomatoes are not in the jar. Speakers also might know that jars are used to preserve food in them, thus, this function will lead speakers to say that “tomatoes are in the jar”. Another important characteristic of the human-produced image description is its conformity with the causal interpretation of the events that happen in the image (Lake et al., 2017). For example, it could be incorrect to say that the food is being cooked in the image if there is a gas-based stove with no indications of it being used. This is closely related to the common world knowledge and the knowledge of what different objects afford (Gibson, 1977) and how they interact.

Humans also have preferences for mentioning specific visual elements.

These often include composition (size, location), knowledge about object and scene categories, and contextual factors relating attributes, objects and scenes (Berg et al., 2012). Humans also make individual decisions about the world, for example whether to have dogs in the kitchens or not. Therefore, as a listener, you might (or might not) have a change in the surprisal levels when hearing that there is a dog in the kitchen. This is supported by the existing psycholinguistic studies that demonstrate that less expected words take more efforts to process them (Demberg and Keller, 2008; Hale, 2001).

3.1.3. Knowledge of intents

The image that the describer talks about in our example has many more objects and relations that could be mentioned, but they are never referred to. Partially, it is due to the shared knowledge about the world that does not need to be mentioned, but it is also related to the intents of the describer. A theory of perceptual selection and cognitive control (Lavie, Hirst, et al., 2004) offers a more detailed explanation for this process. As humans describe images, they use different types of knowledge, specifically, perceptual information and a type of cognitive reasoning over this information that defines communicative intents of the describer. For example, the describer chooses to say “a dog on the floor” in their second sentence, while other possible sentences were not produced such as “the cabinets are closed”. In the case of the latter, there is still an intent to simply identify objects, but it is natural for the image description process to happen in the context of the communication, in which communication goals are important (Brennan and Clark, 1996) and referring is often worked on jointly by both speakers (Clark and Wilkes-Gibbs, 1986). Intents and plausible goals depend on the task (Jokinen, 1996). For example, the description of the image in Section 3.1 might mention only the dog and its visual appearance if the purpose of this description is to answer the question about the dog. This description would not contain information about kitchen appliances as this information is unnecessary to answer the question. The

tendency of humans to rely on the task has been shown to lead to drastic changes in how images are described (Ilinykh, Zarrieß, et al., 2018; Mädebach et al., 2022).

3.2. An example of a model performing a multi-modal task

As we have learned, describing images is a complicated process that requires knowledge of many types of information. Developing *a computational model* that can mimic such a process is extremely challenging, and this goal has been a holy grail of research in the intersection of natural language processing, computer vision, and artificial intelligence in general. The current thesis builds on top of many developments that took place in these research areas over the years. Given all these advances, we ask a simple question: *How do models describe images?* For the purpose of illustration, let us use a publicly available model that is designed to describe images. Here we use recently introduced BLIP model (Li, Li, Xiong, et al., 2022) and provide it with the image in Figure 3.1 to generate its description. BLIP is built on top of the transformer architecture (Vaswani et al., 2017). The model produces the following description: “a dog laying on the floor”¹. With minor changes to the model, we can produce other descriptions such as “kitchen with wood floor”² or “cream yellow and white kitchen”³. Although humans and models perform *same tasks*, i.e. image description generation, model-generated captions are very different from the ones produced by a human in Section 3.1. These differences stem from the fact that while models have access to image pixels and texts, humans use a wider set of information types that we have discussed extensively in Section 3.1. Models in this thesis use only visual features of images and linguistic features of texts. We explore the limits and capabilities of such models in multi-modal tasks that require

¹Model: Salesforce/blip2-opt-2.7b, greedy search, accessed on 2024-03-04 17:25 PM

²beam search with width 4

³ancestral sampling

a form of image description, and in our analysis, we often turn to what we know about how humans operate with linguistic and visual information for inspiration.

Chapter 4: Background and methodology

This thesis addresses questions related to construction of computational models that jointly use language and vision for different tasks. Studying language and vision as two modalities that are used by humans is challenging, and a more systematic approach of building **models** of language and vision is used. We need models to conduct studies on a more manageable scale. Consider the example of a weather prediction model. In order to predict the weather for tomorrow, a large amount of data about atmospheric conditions and other factors is collected and analysed. However, meteorologists cannot immediately identify patterns in the raw data, and it would also be very expensive and time-consuming. By feeding raw data to a model that learns a *representation* of this data, researchers are able to make predictions about future weather patterns.

In this chapter we describe the details of computational models of language and vision that are relevant for our work. All of them are neural architectures which learn from a lot of real-world observational data. We also introduce the set of computational tasks that we target. Throughout this section we aim to relate each task and model to various theories about human language and perception. Neural models are known as “black boxes”; that is, we do not necessarily understand how they accomplish tasks. Hypotheses and ideas from studies on human communication and perception then become handy as they provide us with tools for better interpretation of the models.

4.1. General technical background

A multi-layer perceptron The simplest type of a neural network is a multi-layer perceptron, exemplified in Figure 4.1. Such model is designed to learn

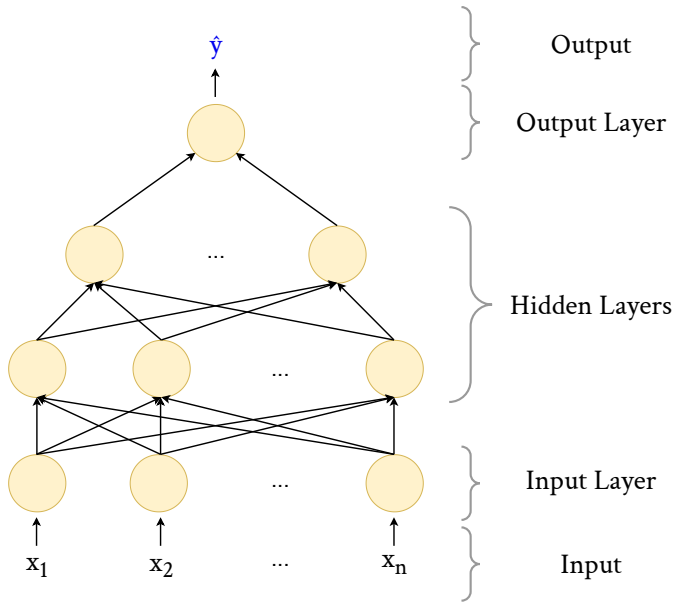


Figure 4.1. A multi-layer perceptron. This type of neural models is often used in classification tasks.

internal representations of inputs that it is provided with (Goodfellow et al., 2016).

The network consists of **layers** of different types such as input layer, hidden layer and output layer. Each layer is a combination of multiple interconnected computations each shown as a coloured circular **node**. Each of these circles can be called a **neuron** and the interaction between these neurons is the core component of a neural network. Every neuron produces an output representation based on the input that it receives. For example, neurons in the input layer *linearly* transform inputs to produce outputs. This output is in turn used by hidden layers, which learn their own representations and produce their own outputs based on the output of the previous layer. Finally, the output layer makes the final prediction. The strength of such networks in the **non-linearity** that is introduced after every linear transformation. Specific non-linear activation functions such as ReLU (Agarap, 2018) allow the model to learn non-trivial connections between different representations. Stacking

more layers and increasing the number of neurons in each layer leads to a **deep** neural network that is capable of learning more complex features.

In practice, each layer is a matrix consisting of rows and columns. If \mathbf{W}_0 stands for the input layer, then each row in this matrix will correspond to one of the coloured circles in the input layer in Figure 4.1. Column values in this matrix can be called **features** of every neuron and their number defines the **dimension size** of the matrix \mathbf{W}_0 . For example, if there are 3 neurons with 5 features each, then $\mathbf{W}_0 \in \mathbb{R}^{3 \times 5}$. Increasing the number of neurons and features often leads to stronger networks, that learn richer representations of data and perform better in tasks.

The whole network can be expressed as the following sequence of computations, starting from the input's transformation:

$$f_0(\mathbf{x}) = \sigma(\mathbf{W}_0 \mathbf{x} + \mathbf{b}_0), \quad (4.1)$$

$$f_1(\mathbf{x}') = \sigma(\mathbf{W}_1 \mathbf{x}' + \mathbf{b}_1), \quad (4.2)$$

$$f_2(\mathbf{x}') = \sigma(\mathbf{W}_2 \mathbf{x}' + \mathbf{b}_2), \quad (4.3)$$

$$f_3(\mathbf{x}') = \sigma(\mathbf{W}_3 \mathbf{x}' + \mathbf{b}_3) \quad (4.4)$$

where f_3 produces the prediction $\hat{\mathbf{y}}$, \mathbf{W}_n are each layer's **weights** that learn representations, \mathbf{b}_n is the bias term of each layer, and σ is the non-linear activation function, which might also differ from layer to layer. The model is trained by the means of the process called backpropagation (Rumelhart et al., 1986), which updates the weights of the model by computing a gradient of a loss function. The loss function computes an error of the model's fit to the data, and this error is used to update the weights by changing them in the direction opposite to the gradient. The most commonly used mechanism to update model's weights is stochastic gradient descent ("Stochastic Estimation of the Maximum of a Regression Function" 1952). Different parameters of the model such as dimension size or number of layers are called model hyperparameters.

The goal of the feed-forward network or a multi-layer perceptron is to learn a mapping between input and output representations. One feature that prevents such networks to be used in language tasks is their non-recurrent nature: each layer and neuron pass the information only once. The concept of continuously feeding the output of the layer to itself is the core of **auto-regressive** neural models, which process the information in a recurrent manner. This is an important feature of more sophisticated types of neural networks such as *recurrent neural networks* (RNNs) or *long-short term memory networks* (LSTMs) as it gives them the ability to provide feedback to itself for every next output. This feature is also what makes these models suited for the task of **language modelling**, in which the model predicts next words from representations of previous words. Therefore, recurrent nature of LSTMs is important for language modelling. Next, we briefly review the history of language modelling task and introduce neural networks which are specifically used for text generation.

Language modelling task To understand how more sophisticated variants of neural networks can be used to generate text, we will first discuss the **next word prediction task** or a language modelling task. This task is now seen as equivalent to the text generation task. However, text generation with pre-neural approaches has often decomposed the task into multiple sub-tasks, connected with one another (Gatt and Krahmer, 2017; Reiter and Dale, 1997; Reiter and Dale, 2000). For example, the model should first decide which parts of the input are important to be described (content selection) and how such elements should be structured in text (document planning). Next, it decides how to realise information as symbols (lexicalisation) and what type of referring expressions to use. Finally, aggregating text with such tools as anaphora and combining lexical elements into a single item (surface realisation) results in the output text. Each step of this paradigm allows for more control and a better understanding of the inner workings of such algorithms

as errors in the output can be traced back to detect which of the sub-modules is ineffective. However, building such generation systems is challenging as each type of input and goal would require an individual approach, requiring researchers to develop many different modelling approaches for many tasks. The neural approach with language modelling as the text generation task offers a very different take. The primary difference is that neural networks are end-to-end systems, in which generation is not conditioned on explicit sub-tasks. Instead, the model learns to represent an input in its continuous representation space and produces an output realised as text.

In language modelling task every next word is predicted depending on what has been produced before, e.g. text history. For example, after seeing “He is turning on his ...” the model could predict “computer” as the next word. In terms of computation, predicting the next word can be expressed through the probability of this word given its previous context, e.g. $p(x_n | (x_1, x_2, \dots, x_{n-1}))$. These probabilities have been traditionally computed based on the frequency of words appearing in specific word contexts. For example, if “computer” is the most frequent continuation of “He is turning on his ...” in the model’s training data, the output of the model which is a probability distribution over all words that the model knows will have “computer” in the head of the distribution. Such an approach is at the foundation of **n-gram language models**. These models restrict the previous history to only n words, reducing the complexity of the generation as now the model is less likely to be affected by the curse of the dimensionality (Bengio and Bengio, 2000), a problem that causes the models to struggle in predicting words from high-dimensional features. Restricting history for predicting the next word in language modelling is related to Markov assumption, which states that the probability of the next event can be assumed to be dependent only on previous n events.

Recurrent models Recurrent neural networks (Elman, 1990) and their extensions such as long-short term memory (Hochreiter and Schmidhuber,

1997) or gated recurrent units (Chung et al., 2014) networks have been used to model time-series data and natural language. These neural network types are more suitable for capturing dependencies between words because, unlike feed-forward networks, they update a single set of weights within each layer of the network, which is computationally efficient. This property allows such networks to not only learn valuable representations at each generation step but also capture information about the order of words and how they make sense. In RNNs, the hidden layers' state at time step t is represented as follows:

$$h_t = \sigma(\mathbf{U} h_{t-1} + \mathbf{W} x_t) \quad (4.5)$$

where σ is the non-linear activation function of choice, \mathbf{U} and \mathbf{W} are weight matrices, h_{t-1} is the previous hidden state representation and x_t is the current word input. RNNs are known to have a problem of vanishing and exploding gradients (Pascanu et al., 2013), a situation when earlier layers of the network are changed less and less with each update from backpropagation meaning that the parts of the network are not learning. The weights can also experience very big updates, leading to an unstable network. These situations occur because of the depth of the network and the number of hidden layers: due to the multiplication of matrices and the nature of backpropagation updates a deeper network might not capture long-range dependencies in sentences.

LSTMs and GRUs learn solve this problem by storing a subset of information and continuously updating it. LSTM, for example, achieves this by learning a separate set of weights for three different gates: input, forget and output. Each of these gates is a matrix \mathbf{W} which is multiplied with the previous hidden state h_{t-1} concatenated with the current input representation x_t . The input gate's goal is to decide how much of the new information to keep, the forget gate decides how much of information from current memory to forget and the output gate captures how much of the information from the cell state should be passed to the next time step. These operations allow LSTMs to

remember information from many time steps back for a very long time, which leads to reasonable gradient changes that do not vanish or explode the weights of the model.

Convolutional models A convolutional neural network is a neural model that is used to process images (LeCun et al., 1989). The structure of this architecture is inspired by how visual cortex is organised. In visual cortex neurons are distributed across different layers and each neuron selectively detects different visual features such as edges, orientation, motion or direction. The most important components of a CNN are convolutional layers, pooling layers and fully connected or feed-forward layers. Convolutional layers use filters to detect image features such as textures or patterns. Pooling layers compress convolutional representations, making them more general and not affected by minor differences in the input. Finally, fully connected layers use the resulting features to make a final prediction, which could be a category of an object in the image. CNNs are commonly used in computer vision tasks such as object detection or image classification.

Encoder-decoder framework The task of image description generation is typically approached with the encoder-decoder modelling scenario, in which both auto-regressive networks (e.g., LSTMs) and convolutional networks play a role. An auto-regressive network is a model that makes predictions based on previous observations. While an encoder (typically a CNN) encodes an image, its representation is used by the decoder (LSTM) to generate text. This model is called sequence-to-sequence model (Sutskever et al., 2014), because *the encoder* compresses visual information into a single representation and sends it to *the decoder* that generates a sequence of words. Studies in this thesis build and analyse models that follows this encoder-decoder framework.

Attention In an encoder-decoder modelling framework, the model is forced to compress input into a fixed-length feature vector. This representation might not be enough to encode all important information from the input, especially when inputs are very large. Cho et al. (2014) observed this problem for machine translation task with the encoder-decoder framework. They have noticed that the performance of the model tends to decrease when the length of the input increases. Instead of packing the whole input into a single representation, one approach could be to learn a mechanism that has access to each element of the input representation and learns to selectively choose elements which are necessary at the specific step. This improvement is called **attention** and was first introduced by Bahdanau et al. (2015) for the task of machine translation. Attention introduces a new set of weights to the decoder that learns to weigh different parts of the encoded input and rely on each of them to a different extent for every generated word. The core idea of the attention is to calculate the **alignment score** between each element in the set of the input feature representations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ and the expected output at a particular time step \mathbf{y}_t . Intuitively, an alignment score is a set of weights, where each weight measures the extent to which a specific element in the input set is important to make an output prediction. For example, when translating “I will borrow a book from a library” into Swedish and producing the next word in “Jag ska låna en ...”, where the next word is “bok”, the model is expected to predict a higher alignment score for mapping “bok” with “book” rather than with “will”.

Once all alignment scores at the particular time step t are computed, each score is multiplied with the corresponding part of the input and a weighted summed input feature vector $\tilde{\mathbf{X}}_t$ is produced:

$$\tilde{\mathbf{X}}_t = \sum_{t=1}^n \alpha_t \mathbf{x}_t. \quad (4.6)$$

This weighted feature vector is part of the input to the LSTM together with embeddings of previously generated words. Once LSTM makes its prediction, the generation process moves to the next time step. Note that every word that is to be predicted requires a whole new set of alignment scores with the input features, thus, the set of scores is computed at every time step t . The score α_t is the result of the following computation:

$$\alpha_t(\mathbf{h}_{t-1}, \mathbf{x}_t) = \text{softmax}(\mathbf{v}_a^\top \tanh(\mathbf{W}_a \mathbf{h}_{t-1} + \mathbf{U}_a \mathbf{x}_t)), \quad (4.7)$$

where \mathbf{W}_a and \mathbf{U}_a are different weight matrices that are used to compute feed-forward computations. \mathbf{v}_a is a context vector that weights input representations taking into account hidden representations at the current timestep. This computation of alignment scores is typically referred to as *additive* (because of addition in the formula). Other mathematical notations to calculate alignment scores have been introduced as well (Luong et al., 2015), but they all share a common idea of learning a separate set of weights trained to represent history and inputs to learn how well these features match with each other.

Transformer model The recurrent nature of neural language models such as LSTMs has often been criticised for its inefficiency in terms of computational resources that are required to build such networks. Although recurrence is intuitively necessary for modelling *natural* language as words in human language follow each other, recurrence also limits the computational power of neural generation models as sentences are not processed in parallel, but word by word. In addition to the computational efficiency bottleneck, past information in recurrent networks is retained only from the previous hidden state, limiting the model’s view of the previous context (Markov property). Different model configurations have been developed to mitigate the aforementioned problems such as bi-directional LSTMs (Peters et al., 2018; Schuster and Paliwal, 1997). The transformer architecture (Vaswani et al., 2017) has performed better than LSTMs and their modifications in many natural language

processing tasks. This architecture does not use any recurrent or convolutional layers, it uses simple linear layers which are matrix multiplications. Below we describe transformer's components which are relevant for our studies.

Self-attention Upon its initial introduction by Vaswani et al. (2017), a transformer is an encoder-decoder language model. In terms of its parts, both encoder and decoder consist of a block with 6 layers. Every next layer is dependent on the previous layer, while the first layer is processing input features all at once, thus, eliminating recurrence on the input level. However, sequential information is still encoded and provided to a model through a different type of input called *positional encoding* represented by sine and cosine functions applied to positions of words in sequences. These functions provide a way to capture the position of a word in a continuous space, possibly allowing the model to learn sequences longer than those observed during training by means of extrapolation.

The transformer model uses **multi-head self-attention**, a mechanism that is capable of learning rich, varied and contextual dependencies between inputs across different layers of each block. Every next layer of the model learns more contextual and richer representations. For each word in the sequence, self-attention in each layer learns a contextualised representation of the word, taking into account other words when constructing the target word's representation. To achieve this, self-attention compares the target word with the other words in the sequence as words that co-occur more often with the target word would inform self-attention that they are the more contextually relevant ones (Cheng et al., 2016). This process might mimic how humans relate different words in text.

In a nutshell, self-attention is learning multiple structures learned over the layers. These transformations are performed by the dot product operation, one of the standard measures to calculate the similarity between two vectors, which gives a scalar value:

$$\text{sim}(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{w}_i \cdot \mathbf{w}_j. \quad (4.8)$$

Applying softmax to the resulting similarity score will transform it to a single value in the range between 0 and 1. These values can be thought of as weights which measure the importance of each intermediate structural representation that are used by self-attention.

The dot product is only part of the picture; the main strength of the transformer is in *how* it learns a whole variety of different modifications of the input with multiple dot product operations. In particular, it uses three weight matrices which learn different transformations of the input:

$$\mathbf{q}_i = \mathbf{x}_i \mathbf{W}^Q; \mathbf{k}_i = \mathbf{x}_i \mathbf{W}^K; \mathbf{v}_i = \mathbf{x}_i \mathbf{W}^V, \quad (4.9)$$

where queries \mathbf{W}^Q , keys \mathbf{W}^K and values \mathbf{W}^V are learned projections of the input representations. Intuitively, the role of the query matrix is to represent the input vector in relation to all other input vectors, given that the current output is at the same position as the input. The key matrix captures a similar set of relations, but this time the weights learn the relation between the current input and all other inputs in terms of the outputs for those other inputs. Finally, once weights have been established, the value matrix is used to compute the resulting output vector by multiplying the weights with the value-transformed input feature. The sequence of these steps is formally defined as follows:

$$\text{score}(\mathbf{w}_i, \mathbf{w}_j) = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}, \quad (4.10)$$

$$\alpha_{i,j} = \text{softmax}(\text{score}(\mathbf{w}_i, \mathbf{w}_j)), \text{ where } \sum_j \alpha_{i,j} = 1 \quad (4.11)$$

$$\mathbf{a}_i = \sum_j \alpha_{i,j} \mathbf{v}_j, \quad (4.12)$$

where \mathbf{a}_i is the output representation vector at the current step i . The scaling factor $\sqrt{d_k}$ is used to scale down the output of the dot product as such multiplication operation can produce either very large or very small values which will have a detrimental effect on the model's training and gradient updates.

Transformers further optimise their use of self-attention by computing it not on the word level, but on the level of sequence of words, leading to a more efficient parallelised set of computations. As words in a sentence can relate to each other in terms of semantic, syntactic and other relations, more sets of parameters are introduced into each layer called *heads*. Each head computes its self-attention with its weights, allowing learning of a variety of relations. Other components are included in the self-attention block such as additional linear layer, residual connections (He et al., 2016) and normalisation layers (Ba et al., 2016).

4.2. Language-and-vision natural language processing

Transformer models for language-and-vision tasks Transformer-based language models have been widely used to learn *general* representations of language, making these models useful for many computational linguistic tasks (Devlin, Chang, et al., 2019; Radford, Wu, et al., 2019). It is not surprising that the field of language-and-vision has followed a similar route and proposed many different modelling architectures that learn general multi-modal representations that are not bounded to the specific task. Several models can be mentioned among the proposed approaches such as UNITER (Chen, Li, Yu, et al., 2020), LXMERT (Tan and Bansal, 2019), ViLBERT (Lu, Batra, et al., 2019), OSCAR (Li, Yin, et al., 2020) and VL-BERT (Su et al., 2020). Such models are typically trained in two steps. First, the models are *pre-trained* with special tasks such as multi-modal masked language modelling. Similar to the standard masked language modelling in BERT (Devlin, Chang, et al., 2019), the multi-modal version of this task requires the model to predict masked

words given other words and regions in the image. Other multi-modal pre-training tasks which allow the model to connect visual and linguistic features such as embeddings of words describing objects in the image represented as visual features, include masked region/object modelling and image-text matching. In the former case, the model learns to predict either features of the masked target region or object label distribution for the target region. Image-text matching is a simple feed-forward network that decides if the text is describing the image based on the combination of object and word representations. Next, the models are *fine-tuned* on a number of downstream tasks with the help of special classifiers that learn task-specific representations and not necessarily the general ones. One example is the image retrieval task: the model needs to identify image that corresponds to the textual description. The task is often performed with data from MSCOCO image captioning dataset (Lin, Maire, et al., 2014) and visual question answering dataset (Antol et al., 2015). The downstream tasks benefit from general knowledge captured in multi-modal transformers as, for example, in order to retrieve the right image corresponding to the text, the model needs to know how words and visual elements come together.

Tasks and datasets This thesis examines multi-modal models in the context of different language-and-vision tasks. One of them is **image captioning** which requires a model that generates a single sentence describing an image. Multiple datasets have been built and collected for this task such as MSCOCO (Lin, Maire, et al., 2014), Flickr30k (Plummer et al., 2015) and, more recently, Conceptual Captions (Sharma et al., 2018). This task has seen a lot of attention in terms of the development of modelling solutions (Donahue et al., 2017; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). A more complex version of the image captioning task is the task of **image paragraph generation** (Krause et al., 2017) which requires a model that can produce multiple sentences about the image, introducing modelling challenges in terms of the discourse

structure and coherence of generated text. Study VIII introduces a novel task for image description generation: a **perceptual category description** in which we explore generation of the description of a category (e.g., “raven”) based on visual features of images from this category or abstract representations learned from many instances from the category.

Most of the studies in this thesis build and evaluate models designed for the three tasks described above. One of the primary question that Studies I, II, and III focus on is the analysis of self-attention in the object relation transformer (Herdade et al., 2019). This transformer generates either one-sentence (captions) or multi-sentence image descriptions (paragraphs). Study IV employs a CNN-LSTM-based model for image paragraph generation. Finally, both Study VII and Study VIII use transformer-based models for image paragraph generation and generation of perceptual categories respectively.

The rest of the studies in this thesis explores other multi-modal tasks. Study V examines models in the context of the task of **embodied question answering** (Das, Datta, et al., 2018). The task of embodied question answering (EQA) is split into two parts: a navigation task and a question answering task. In the navigation task the agent is provided with a question about some target object. For example, “what is the colour of the couch in the living room?”. First, the agent navigates in a 3D virtual world based on House3D environments (Wu, Wu, et al., 2018) and finds the object. Next, the agent answers the question given a recent visual history of image frames with the target object preferably being visible in those frames. We train and test a CNN-LSTM based question answering part of the agent trained for the EQA task. We specifically look at the level of sensitivity of this part of the agent to perturbations of visual features.

Study VI explores the task of human variation in **object naming**. In its fundamental form, the task is very similar to the classic referring expression generation task (Reiter and Dale, 2000): given an image with a target object in it, produce an expression that describes this object. The novelty here is that

the dataset provided by Silberer, Zarri , Westera, et al. (2020) also contains annotations from multiple humans describing a single object. This allows us to study the effects of different contextual factors on the *variation* in human object naming.

Object relation transformer Our studies that examine the tasks of generation of image descriptions (captions, paragraphs) employ a specific transformer-based architecture. Here we motivate the choice of this architecture. The object relation transformer that we use in many studies in this thesis is a two-stream multi-modal image description generation transformer (Herdade et al., 2019). The architecture of the model highly resembles the original transformer architecture (Vaswani et al., 2017). In particular, the architecture centres around three self-attention blocks, each of them operating with different type of modalities and representations. Figure 4.2 illustrates the architecture of the model. Below we focus on the specifics of different self-attention blocks. All other parts of the transformer such as residual connections or feed-forward layers are shown in the Figure 4.2, but not discussed explicitly in the text below.

The **image encoder** block is provided with both visual features and bounding box coordinates of each region. Its task is to encode and combine two complementary types of information about image regions. The primary advantage of this block is the fact that it learns complex representations of spatial information between bounding boxes of objects. This allows the model to utilise **relative geometry between objects**. Relations between objects are often captured by existing multi-modal transformers *independently* from each other. For example, LXMERT is trained with bounding box coordinates as part of its input, and VL-BERT normalises these coordinates by the height and width of the input image. These models do not learn information about relative geometry between objects. In what follows we describe how visual and geometric attention weights are computed and combined.

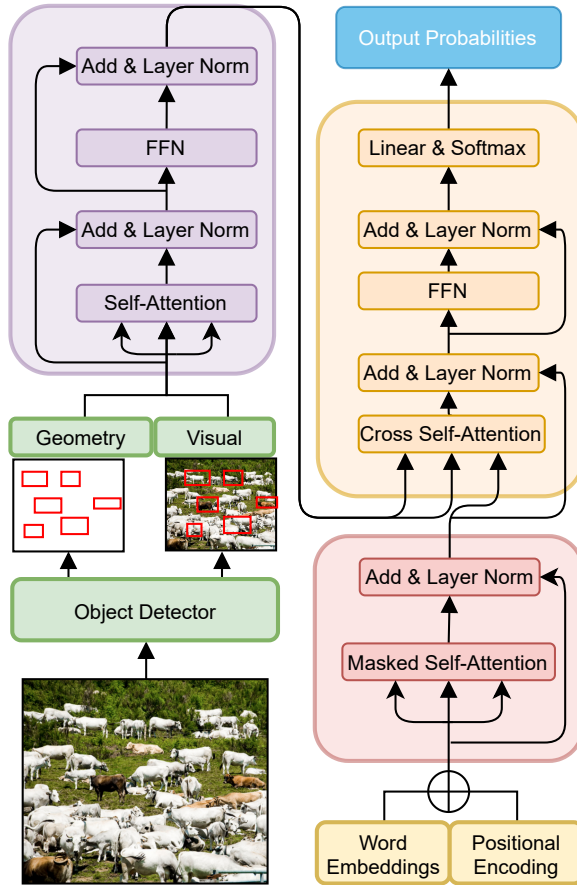


Figure 4.2. Object relation image captioning transformer. The model schema was initially described in Study I, here we duplicate it in order to explain the model’s key components.

The **image encoder** block operates with representations extracted from visual input. We extract these representations from the bottom-up attention model, designed to detect objects in the image (Anderson, He, et al., 2018)¹. The extractor is based on Faster R-CNN (Ren, Kiros, et al., 2015) with ResNet-101 (He et al., 2016) as its visual backbone. It is trained on annotations from Visual Genome (Krishna et al., 2017) to detect and produce information about N image regions, where $N = 36$. Each image region is represented

¹<https://github.com/peteanderson80/bottom-up-attention>

as a feature vector $\mathbf{x}_v \in \mathbb{R}^{1 \times D}$, where $D = 2048$. In addition, each region is supplied with information about position of its bounding box in the image and linguistic description. The bounding box coordinates are relative to the image size. Linguistics descriptions consist of labels and attributes, where labels are typically nouns (“dog”) and attributes are adjectives (“big”). It is important to note that in order to assign a linguistic description to the image region, the feature extractor produces a probability distribution over its vocabulary of labels (1600 overall) and attributes (400 overall). Labels are nouns (e.g., “a couch”) and attributes are adjectives (e.g., “brown”) that describe objects. It then picks the most probable label and attribute akin to multi-class classification. The set of probability values can also be extracted to examine how confidence the detector is in assigning linguistic description to image regions.

A lot of previous modelling approaches for image captioning have represented images as a single vector (Donahue et al., 2017; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). These representations can be viewed as a global top-down representations of the image. Previous research on models for image captioning has shown that a combination of top-down and bottom-up representations results in higher quality image captions (Anderson, He, et al., 2018; Li, Tang, et al., 2017). Such bottom-up representations instead use object-level features to represent an image. In the field of computer vision, a different type of image representations have been used such as 2D image patches, specifically for the task of image classification (Caron et al., 2021; Dosovitskiy et al., 2021). Patch-based representations do not necessarily capture semantic information about images unlike bounding boxes of detected objects which are identified with models like Faster R-CNN (Ren, He, et al., 2015). Such representations are viewed as a top-down signal and have been used to represent images for image classification tasks. In most of our studies we represent images with bottom-up features of image regions. Study II in Part 5.1 also proposes the analysis of interpretation of structures in self-attention when the model is

provided with patch-based features.

Next, we describe how the **image encoder** processes visual and geometric information. First, the dimension size of each visual feature is reduced from 2048 to 512 and non-linearity with dropout layer are applied. The result is the set of inputs fed to the first layer of the multi-layer multi-head self-attention. Every next layer in this block uses output representations of the previous layer. In a standard transformer-based fashion, each attention head computes different projections of the matrix of visual features \mathbf{X} :

$$\mathbf{Q} = \mathbf{XW}^{\mathbf{Q}}, \mathbf{K} = \mathbf{XW}^{\mathbf{K}}, \mathbf{V} = \mathbf{XW}^{\mathbf{V}}. \quad (4.13)$$

The *visual attention weights* are then calculated with a standard multiplication of queries and keys scaled by a factor of d_k :

$$\mathbf{\Omega}^{\mathbf{V}} = \frac{\mathbf{Q} \cdot \mathbf{K}}{\sqrt{d_k}}, \quad (4.14)$$

where $\mathbf{\Omega}^{\mathbf{V}}$ is a matrix that contains attention weights between visual representations of every two detected image regions.

At this stage, **geometric attention weights** are calculated and combined with $\mathbf{\Omega}^{\mathbf{V}}$. We describe what these weights are and how they are computed as follows. In using geometric information about image regions, authors of the object relation transformer are motivated by the object relation module introduced by Hu, Gu, et al. (2018) who show that geometric context is helpful for visual tasks such as object detection (Divvala et al., 2009). A parallel and related type of context that is useful for linguistic tasks such as image captioning is the knowledge of spatial relations between objects (Talmy, 1983). The object relation transformer is expected benefit from both types of information as they are complementary.

As the first step, a 4-dimensional displacement vector between every two objects m and n is computed:

$$\lambda(m, n) = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{|w_n|}{w_m}\right), \log\left(\frac{|h_n|}{h_m}\right) \right), \quad (4.15)$$

Next, each value in the vector is passed through the sinusoid function which is described in (Vaswani et al., 2017). This way the displacement vector is treated as positional encoding in language transformers, although in this case it captures relations geometry between objects. The result is fed to the special linear layer *Emd* which produces embedding in a high-dimensional space. This vector is then multiplied with a learned projection matrix \mathbf{W}_G to produce a scalar value. ReLU non-linearity is applied as well.

The result of previous transformations are geometric attention weights Ω^G , and they are combined with visual attention weights as follows:

$$\Omega = \log(\Omega^G) + \Omega^V. \quad (4.16)$$

Finally, the output of each self-attention head in the green block is computed as a multiplication of the weight matrix with the value matrix of self-attention, e.g. $\text{softmax}(\Omega)\mathbf{V}$. We note that geometric attention weights Ω^G are combined *separately* at each layer and its corresponding input. Therefore, every layer in the model's self-attention is learning from original geometric representations instead of relying on the representations of geometry from previous layers.

The **text decoder** block operates with representations extracted from textual input. This block's role is to generate text in left-to-right fashion given word embeddings and positional encoding (Vaswani et al., 2017). Every next token w_t is conditioned only on previous tokens, e.g. $W_{\setminus t} := (w_1, \dots, w_{t-1})$. This block of self-attention layers contains attention weights between different words in the input. We note that due to auto-regressive nature of generation,

attention weights cannot be constructed between the current word and future words.

The **cross-modal** block of self-attention layers is the last important piece of the architecture. Its role can be described as the task of combining visual and linguistic representations, therefore, this block allows us to examine the attention weights between tokens and image regions. The output of this block is passed through a linear layer to produce a distribution over the vocabulary of the model. This vocabulary is then used by a *decoding method* of choice to choose the next word.

Studies in Part 5.1 inspect structures in self-attention from all three blocks that we have introduced. Study I is analysing the patterns captured in the **text decoder** block and compares them against a pure text-only decoder model. Study II is investigating the attention weights build by the **image encoder** block between image regions or image patches for the task of image captioning. Finally, Study III is examining the weights that are build by the **cross-modal** block between tokens and regions for different types of words such as descriptions of objects (noun phrases) or spatial relations (verbs and adpositions).

One-stream or two-stream multi-modal transformer? An important question to ask is whether there is a particular preference to use two-stream multi-modal transformer over one-stream architecture for image captioning. One-stream transformers such as VL-BERT (Su et al., 2020) have a single self-attention block that is provided with visual and linguistic features of images and texts. Two-stream transformers such as ViLBERT (Lu, Batra, et al., 2019) first process visual and linguistic features by independent blocks of self-attention and then used by a cross-modal self-attention to make a prediction. Existing research on the question of which type of transformers is better (Chen, Li, Yu, et al., 2020; Lu, Batra, et al., 2019) focuses on how these models differ in performance in tasks other than image captioning. Bugliarello

et al. (2021) show that the differences between different architectures are observed only on specific tasks and they are not generalisable across different multi-modal tasks. The main difference appears to be due to the training data and hyper parameters. Future research must investigate better what type of architecture is more suited for image captioning task. However, the nature of the two-stream model allows us to examine self-attention patterns in both visual and textual encoder as well as cross-modal block of self-attention.

Motivation for choosing object relation transformer There are several reasons for us to use object relation transformer in our experiments. This model is specifically designed for image captioning task and it learns complex geometric information between objects. In comparison, existing multi-modal transformers such as LXMERT (Tan and Bansal, 2019) which are not used in left-to-right generation tasks are either not pre-trained or tested for image captioning task, although they might use image captioning data for pre-training on other tasks. It is possible to adopt self-attention in such models for text generation, but it requires introduction of special attention masks that would prevent the model from looking into the future when generating texts (Li, Yin, et al., 2020; Scialom et al., 2020; Zhou, Palangi, et al., 2020). Additionally, multi-modal transformers at that time were not learning a complex geometry between objects, which is necessary for image captioning. A similar two-stream architecture that does not use object features and knowledge of spatial relations has demonstrated lower scores on automatic metrics on the COCO test set (Sharma et al., 2018).

Other multi-modal transformers An important research context that this thesis relates to is how general or task-specific multi-modal transformers should be. The multi-modal NLP has generally seen a push towards building more general-purpose architectures that achieve high performance on many different tasks and benchmarks. These models learn good multi-modal

representations as they perform well on general multi-modal tasks such as discrimination between image-sentence pairs, e.g. image-text matching. Some more recent and noticeable architectures include CLIP (Radford, Kim, et al., 2021) and ALBEF (Li, Selvaraju, et al., 2021). Some work shows that models can learn better visual representations from language-and-vision supervision, e.g. ALIGN (Jia et al., 2021) and VirTex (Desai and Johnson, 2021). However, these models are typically used in general-purpose tasks such as image classification where a fixed number of possible outcomes is pre-defined. When applied to a more specific task such as object counting, general-purpose models such as LXMERT (Tan and Bansal, 2019) do not generalise well to out-of-distribution examples (Parcalabescu, Gatt, et al., 2021). The task of image captioning is more open-ended than counting as there are many viable descriptions for an image. A different line of research has introduced models that are suitable for *both* general-purpose and open-ended tasks. Some of these prominent models are BLIP (Li, Li, Xiong, et al., 2022), BLIP-2 (Li, Li, Savarese, et al., 2023), OFA (Wang, Yang, et al., 2022), Flamingo (Alayrac et al., 2022), and LLaVA (Liu, Li, et al., 2023). A noticeable difference of these models from more general-purpose models is that they can be easily applied in image captioning. They achieve this primarily by converting the image-text matching task into instruction-following format, in which a special prompt (e.g., “What is in the image?”) is appended as input to the model alongside the image.

This thesis contributes to the analysis of general-purpose and task-specific multi-modal transformers. Studies in Part 5.1 use architecture that is specifically designed for text generation. Studies in Part 5.2 and Part 5.3 use representations from general-purpose architecture such as CLIP (Radford, Kim, et al., 2021) or BERT (Devlin, Chang, et al., 2019).

Deterministic and stochastic decoding methods The output of a probabilistic text generator is text consisting of a series of words, with each word generated one after the other. However, what these generators produce is not

a single word, but rather a *probability distribution* across multiple words that the generator has knowledge of. As the input to the generator is incrementally updated, it creates a new probability distribution at each time step. In order to decode text one needs to choose a **decoding method**, which traverses through probability distributions at different time steps and uses a particular heuristics to select words. The sheer size of each distribution coupled with the fact that they vary from one time step to another introduces challenges in finding better methods for word selection.

A number of different deterministic heuristics are commonly used in (multi-modal) text generation. Such methods are stable in producing the same output every time they are used. The simplest deterministic decoding method is to **greedily select the most probable word** at each time step during generation. This method reduces the problem of selecting words within a very vast space of probability distributions. While this often results in a text of high quality on the local word level, which is desirable for some tasks such as machine translation, word choices made greedily might not result in the most optimal sequence of words that is possible under a specific model on a global sentence level (Chen, Li, Cho, et al., 2018). **Beam decoding** offers a partial solution to this problem by maintaining a “beam” of a few other highest probability candidates for every time step and considering multiple alternatives. It calculates conditional probabilities over these alternatives and finds the most likely sequence. Still, beam explores only a few highly probable word candidates and might return a dull and uninteresting text (DeLucia et al., 2021). This is partially because the data is known to follow a Zipfian distribution (Zipf, 1949) and the head of the probability distribution at each generation time step typically consists of more or less the same words.

Sutskever et al. (2014) show that the machine translation model generates most accurate French translations from English texts with beam search. In such tasks as machine translation, deterministic decoding methods have advantage as what matters is generation of accurate texts that correspond to

the ground truth as much as possible. However, because these methods do not explore the probability space to its fullest, they are known to generate candidates that have little to no differences between them (Li and Jurafsky, 2016). Other tasks such as story generation are much more open-ended and they require more focus on **diversity** in generated texts, i.e. accurate texts with many different combinations of words. Generating diverse texts is also relevant for properly capturing variation in referring expression generation (Castro Ferreira et al., 2016).

In order to generate more diverse texts by exploring the probability distribution better, stochastic decoding methods are widely used. By introducing randomness and uncertainty during inference time, such methods relax the maximum likelihood constraint and lead to more diverse texts (Holtzman et al., 2020; Ippolito et al., 2019; Panagiaris et al., 2020). One of the reasons for this is that stochastic methods *sample* from the probability distribution, but they would prefer to take words with similar probabilities which are also semantically similar. Ancestral (random) sampling is the simplest stochastic decoding method which selects a random word from the multinomial distribution at each time step. Other methods such as top- k and top- p decoding algorithms sample from the subset of probability distribution defined either by the number of candidates to consider (k) or the accumulative probability mass (p). Temperature is another metric to either make probabilities sharper or more uniform and sample from them. One disadvantage of stochastic decoding methods is the lack of control over their output: it is hard to find a fitting set of hyper parameters such as k value or p value that would not result in text hallucinations and non-sensical texts.

Evaluation Language-and-vision models are typically evaluated in terms of the quality of texts that they generate. A set of standard automatic evaluation metrics, measuring the accuracy of generated texts is employed, e.g. BLEU (Papineni et al., 2002). Such metrics are typically focused on computing n -

gram matches of different forms between different texts, but they show poor correlation with humans. (Zhang, Kishore, et al., 2020) introduce a metric that uses contextualised embeddings to evaluate text generation. Entropy can also be used to evaluate image description models as it measures the uncertainty of the model in making predictions (Shannon, 1948). Other automatic metrics measure diversity of generated descriptions, their faithfulness to the image (Madhyastha et al., 2019) and evaluate descriptions based on their purpose (Fisch et al., 2020). The problem of generating diverse image descriptions is deeply related to a more general problem of capturing the diversity in how different humans tackle different tasks. An image can be described in many different ways by different humans and this is related to such factors as subjectivity in annotations, multiple plausible answers, and general variation in how humans “label” the world (Pavlick and Kwiatkowski, 2019; Plank, 2022). This is an important question to investigate as current NLP and ML approaches (somewhat mistakenly) assume that there is a ground-truth for a task, developing a dataset and addressing a specific benchmark (Schlangen, 2021).

Evaluation of natural language generation system is inherently difficult task (Reiter and Belz, 2009). Evaluation becomes more challenging in the domain of language-and-vision tasks such as image captioning, where images can be described in many ways depending on the contextual factors. One way to evaluate the quality of automatically generated image descriptions is to compare them against ground-truth human-generated ones which are typically collected as part of the dataset that the model is trained and tested on. In this type of evaluation, researchers are interested in how well two types of texts match each other. Therefore, it is not surprising that this evaluation adopts metrics from other relevant text generation tasks. These metrics are typically BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), or ROUGE (Lin, 2004). Other text-based automatic evaluation metrics were developed specifically for image description evaluation, e.g. CIDEr

(Vedantam, Lawrence Zitnick, et al., 2015), SPICE (Anderson, Fernando, et al., 2016). A significant problem with this type of evaluation is that their scores do not correlate well with human judgements, which are considered to be the gold standard in evaluation of generated texts (Elliott and Keller, 2014; Hodosh et al., 2013; Kulkarni, Premraj, Ordonez, et al., 2013). Such metrics as BLEURT (Sellam et al., 2020) and BERTScore (Zhang, Kishore, et al., 2020) correlate better with human judgements. This has become possible due to the high quality of BERT embeddings and contextualised knowledge stored in them.

Human judgements about properties of generated texts are typically considered the desirable type of evaluation in language generation community. In this setup, humans who can be either linguists (experts) or non-experts (workers from crowd-sourcing platforms) are asked to judge different image descriptions across various criteria. These criteria typically include human-likeness, grammatical correctness, accuracy, relevance (Elliott and Keller, 2013; Kuznetsova et al., 2012; Mitchell, Dodge, et al., 2012). The evaluators are typically asked to rank or rate descriptions on a Likert scale associated with one of the evaluation criteria. Although human evaluation is often viewed as the most reliable type of evaluation, it is often hard to control for, primarily because of the lack of standardised evaluation sheets and replicability of the results (Howcroft et al., 2020).

Text-only evaluation of image descriptions does not take the image itself into account, and a set of other types of metrics have been introduced to account for the visual content. Jiang et al. (2019) have proposed to evaluation not only text-level matching between the texts, but also how well a generated text matches visual content. Hessel et al. (2021) propose a method for measuring compatibility between images and texts without textual ground-truth references. Madhyastha et al. (2019) propose to measure faithfulness of descriptions to images based on the similarity between object descriptions and labels of the objects. These metrics are a more natural choice to the multi-

modal task of image description generation. A subset of metrics is focused on the usefulness of description for a specific task, e.g. what does a description require to be a good generated text in the specific task? Such metrics include, for example, CapWAP (Fisch et al., 2020) that evaluates image descriptions based on their utility for the information needs of the reader.

Chapter 5: Summaries of studies

5.1. Part I: The role of self-attention in object relation transformer

Studies in this part of the thesis examine attention weights in three different blocks of self-attention layers in the object relation transformer (Herdade et al., 2019).

In general, we address the following question:

- How do self-attention weights connect words and objects in the context of two different computational tasks, and can we identify linguistically and cognitively interpretable patterns in these weights?

We use the object relation transformer (Herdade et al., 2019) in the context of two tasks: image captioning and image paragraph generation. Each study analyses weights in one of the three self-attention blocks of the model, described in detail in Section 4.2. In this model a single self-attention block can operate with either linguistic or visual representations or both. We specifically examine the connections built by each block for the corresponding input type, e.g. text, image, image-and-text. Our interpretation of the patterns is based on insights from different theories about human language and visual perception such as the theory of visual routines (Ullman, 1984) and load theory of selective attention and cognitive control (Lavie, Hirst, et al., 2004).

5.1.1. Motivation

The question of whether transformer-based models encode any linguistic or visual information has received a lot of attention in both NLP and computer vision. Transformer models have been shown to “reinvent the NLP pipeline” as they appear to hierarchically encode linguistic information, with local

syntax captured in earlier layers and complex semantics learned in later layers (Tenney, Das, et al., 2019). In fact, the research on the interpretability of neural NLP models has evolved into the field of “BERTology” (Rogers et al., 2020). This field employs a range of analysis methods to interpret models (Belinkov, 2018; Belinkov and Glass, 2019). Next, we describe the methods that we employ in our studies and explain why we chose them.

The first group of interpretability methods is referred to as “*black-box*” methods. These methods make decisions about what the model has learned based on how the model behaves under different dataset conditions. For example, Gardner et al. (2020) introduce contrast data sets that are test sets with minor perturbations designed to test the model’s linguistic capabilities. Shekhar et al. (2017) introduce the dataset of foil captions on which the models are evaluated for their ability to detect or correct incorrect words in image captions. In general, black-box methods develop datasets that are designed to test whether models have a specific type of knowledge without interpreting the internal workings of such models.

The second group of interpretability methods is the “*white-box*” methods that inspect the processes inside the models and their parts. For example, extracting the probability distributions from the models and examining how the change of dataset domain shifts these probabilities can tell if the model acquired useful abstractions from pre-training and was able to apply them in a new domain (Rethmeier et al., 2020). Specific input-output pairs can also lead to different gradient values inside the model, making these values reflective of the knowledge that the model has (Du et al., 2023; Selvaraju et al., 2017).

A more prominent example of the white-box interpretability method is the use of model’s representations of self-attention by a separate probing classifier to perform an auxiliary task. The classifier’s performance on the auxiliary task informs us about the information the model’s representations have about linguistic phenomena of interest (Belinkov, 2022). It is possible

to probe the model's self-attention for many different linguistic properties of texts such as sentence length or syntactic tree depth (Conneau et al., 2018). Building a classifier that is parameter-disjoint from the original model and is agnostic of the original training task makes it challenging to understand how much its performance tells us about the knowledge captured in self-attention representations (Belinkov and Glass, 2019).

A different white-box method is to **visualise** self-attention weights as heatmaps and directly interpret them (Vig, 2019). Instead of assuming that there is a linguistic property that the attention heads might capture as in the probing method, visualising self-attention allows us to first examine the connections built inside the model and then interpret them in terms of the possible linguistic knowledge that these connections might reflect. While visualising self-attention appears to be easier for direct interpretation, it is unclear whether self-attention weights can be interpreted to "explain" the model's internal knowledge.

Relying on self-attention weights for interpreting the knowledge that models learn is a topic that has sparked a lot of debate. Jain and Wallace (2019) show experimentally that it is not reliable to assume that attention weights explain which input feature is responsible for which output feature. Serrano and Smith (2019) demonstrate that higher attention weights do not necessarily correlate with changes in the model's performance. Others argue that using attention as an explanation of the model's learned knowledge depends on the architecture, the task, and the underlying notion of "explanation" (Wiegreffe and Pinter, 2019). While Bastings and Filippova (2020) propose using input saliency methods instead of attention for model interpretation, they also argue that studying the role of attention and the functions it captures in different tasks is a valid research goal. Therefore, it is necessary to make a clear distinction between "attention as an explanation" of what the whole model learns and the role of the attention mechanism itself in the model's learning and the knowledge that attention contains in its weights.

The previous work has examined attention weights in the context of mostly text-only tasks. For example, self-attention has been analysed for knowledge of anaphora resolution and word sense disambiguation in machine translation tasks (Tang et al., 2018; Voita, Serdyukov, et al., 2018). Self-attention heads can also be compared with each other in terms of their importance for making predictions for linguistic tasks and heads that are not useful can be pruned (Voita, Talbot, et al., 2019). The analysis of self-attention weights has been widely used in NLP, allowing researchers, for example, to examine what BERT (Devlin, Chang, et al., 2019) learns about coreferential or syntactic knowledge (Clark, Khandelwal, et al., 2019). Ghader and Monz (2017) have shown that attention weights reflect useful information about alignment in the context of the machine translation task and even capture useful information beyond alignment. Some other work has shown that models can learn knowledge about syntactic dependencies and distribute it between different layers and self-attention heads (Blevins et al., 2018; Tenney, Das, et al., 2019; Tenney, Xia, et al., 2019). Raganato and Tiedemann (2018) and Goldberg (2019) have shown that self-attention heads in different layers of the text-only transformer contain representations that can reflect the knowledge of syntax (e.g., syntactic dependencies) or semantics in texts. Analysis of attention weights and their visualisation for interpretation have also been studied in computer vision community. Visualisation methods for interpretability can often indicate which part of the network is responsible for what type of knowledge (Erhan et al., 2009). Research in computer vision has analysed attention heatmaps and distances between attended image regions built by transformer-based image classification models (Caron et al., 2021; Dosovitskiy et al., 2021). Dosovitskiy et al. (2021), in particular, have demonstrated that attention across layers in such models is processed in a very specific way: heads in earlier layers attend to pixels that are at various distances, while heads in later layers largely focus on pixels that are very distant from each other.

Studies I, II, and III focus on the analysis of the knowledge that attention

captures in the context of two image description tasks which has not been done before. We think that examining the behaviour of self-attention in these tasks brings a lot of useful insights about the role of self-attention in image description transformers because attention in such tasks is also visual, and visual attention can intuitively be associated with the model's attention on different parts of the image. Some recent work has looked at what pre-trained language-and-vision transformers and their self-attention mechanisms learn about specific multi-modal tasks such as visual coreference resolution and visual relation detection (Cao et al., 2020). Here we study both pre-trained models and models trained from scratch. We do not make initial assumptions about the type of linguistic structures and knowledge that self-attention can capture in the context of our computational tasks. Instead, we are looking at attention in its more direct form: whether the object referred to is attended to by the model. Our primary contribution is the analysis showing that self-attention weights can be good predictors of the semantic knowledge captured by the image description models.

Linking objects and words Interpreting self-attention weights is typically done by examining how the weights align with specific linguistic relationships among the attended words. Some of such relations (e.g., syntactic dependencies between words, part-of-speech tags) can be extracted automatically with the help of existing NLP tools such as spaCy (Honnibal et al., 2020). Other types of relations between words such as anaphora resolution are often annotated by humans due to generally higher quality of human annotations compared to machine annotations. In both cases there exists a ground truth that we can compare against the self-attention weights.

Examining self-attention in a language-and-vision context requires ground truth data that would include linking between words and objects. In Studies II and III we need to know which regions in the images are described in texts, and linking between bounding boxes of regions and noun phrases

describing objects is required. These links help us identify self-attention heads that put more weight on objects and words within these links rather than on objects and words outside of them. However, such ground truth linking is not available in the datasets that we use in our studies, which are MSCOCO (Lin, Maire, et al., 2014), the Stanford image paragraph dataset (Krause et al., 2017), and Tell-me-more (Ilinykh, Zarrieß, et al., 2019b). At the same time, existing work that offers automatic tools or manual annotations of links between words and objects has typically been conducted in the context of producing referring expressions for objects and not descriptions for images. Examples of datasets that provide human-annotated links between words and objects are ReferItGame (Kazemzadeh et al., 2014) or instruction-based human annotation of objects in Visual Genome (Krishna et al., 2017). Other work investigates how neural networks can be used to link perceptual features of real-world objects with referring expressions (Schlangen et al., 2016).

Descriptions of objects produced in the referring expression generation task and image captioning task differ in terms of their informativity and length (Coppock et al., 2020). These differences appear due to a more specific communication goal within the referring expression generation task, which requires describers to identify *a specific object* within visual context (Reiter and Dale, 2000), while the image captioning task focuses on describing *images* rather than specific objects (Chen, Fang, et al., 2015). In other words, there is a lack of an explicit need to distinguish referents in the image captioning task. It is unclear if we can directly use methods that connect objects with referring expressions in the context of our tasks, which are focused on the generation of sentences and paragraphs. Here we use an automatic linking method that has been specifically designed to link object descriptions with the corresponding objects by comparing these descriptions with object labels. The method that our studies use for the automatic linking of texts and objects was initially introduced in (Ilinykh, Zarrieß, et al., 2019b). In particular, nouns in texts are linked with object labels available in the corpus by first performing a simple

string-based matching, and if the matches are not found, then the cosine similarity score between word vectors is computed and used to determine which nouns should be linked with which object labels. The method has been further refined and extended in (Dobnik, Ilinykh, et al., 2022), where the attributes and labels of detected objects were determined based on the confidence scores of the object detector model. Study III improves the method even more by explicitly examining what type of pre-trained feature embedding leads to a more accurate linking.

Theories of human visual cognition and attention Human perception is theorised to be hierarchical as neurons learn information of different complexity from the visual input. Such organisation of perceptual knowledge is important for a biological being (Tenenbaum et al., 2011). In Study II we analyse attention weights across layers of the self-attention that operates on the image alone and interpret these weights through the theory of visual routines (Ullman, 1984). This theory introduces a framework that splits human visual cognition into two stages. In the first stage humans construct base visual representations that capture information about general properties of the image such as its colours, edges, their orientation, and motion. When building base representations of the image, humans do not use any high-level knowledge specific to the objects or the task. In the second stage humans apply “visual routines” to the base representations constructed in the previous stage. These routines are used to build a high-level understanding of objects and relations in the image. Study II proposes the interpretation of the self-attention weights in different layers through the theory of visual routines. We show that under specific input feature representations, the model first connects bounding boxes that are thematically related and geometrically close, but do not necessarily correspond to objects. On top of that, the model builds connections between bounding boxes that correspond to different objects in the image, which are ultimately described in the generated caption.

In Study II, we also observe that later layers of the model connect objects that are then described in the generated descriptions. We connect this result with insights from the load theory of selective attention and cognitive control (Lavie, Hirst, et al., 2004). The theory states that humans first perceptually select information they want to describe and then select what and how to describe from this information given the task at hand. Our experiments suggest that the model performs selection of what to describe given the image captioning task as objects it connects in later layers are also described in the caption.

5.1.2. Study I: How Vision Affects Language

- **How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer.** Nikolai Ilinykh and Simon Dobnik. 2021. In Proceedings of the 1st Workshop on Multi-modal Semantic Representations (MMSR), pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.
Link: <https://aclanthology.org/2021.mmsr-1.5/>

5.1.2.1. Overview

In this study we examine a specific type of self-attention commonly referred to as masked self-attention in the context of the image captioning task. This module is illustrated as the **text decoder** in Figure 4.2. The role of this module in the multi-modal image description two-stream transformer is to produce a word representation at every generation step. This module does not have direct access to the image, as the image is handled by a separate self-attention module. However, the model as a whole needs to learn from both modalities, and due to its training regime and back-propagation, different representations are expected to affect each other. We hypothesise that self-attention on previously generated words differs between different task setups, uni-modal and multi-modal. In the study we observe a difference in the structures captured by

masked self-attention weights between words in a description in two modality-different setups.

5.1.2.2. Questions and findings

Question I Do self-attention patterns built by masked self-attention on generated words vary in text-only and language-and-vision task setups?

We compare self-attention weights built between words in texts that are generated either from previous words (uni-modal) or from previous words and images (multi-modal). We use pre-trained GPT-2 (Radford, Wu, et al., 2019) as our language-only model because in our study this model is architecturally close to the part of the image captioning transformer that incorporates masked self-attention (Herdade et al., 2019). We observe that when a word is generated in a multi-modal scenario, masked self-attention in the image object relation transformer demonstrates a higher focus on previously generated nouns and a smaller focus on other parts of the already generated text. In comparison, GPT-2’s masked self-attention relies heavily only on a few immediately generated words, showing a more local pattern, e.g., attention is “neighbouring” the word that is being generated instead of reaching more distant words. Uni-modal masked self-attention also has higher entropy as it appears to be less confident in choosing which word matters the most at a particular time step. Multi-modal masked self-attention, in comparison, has more focus on specific words (e.g., lower entropy), suggesting that by focusing on nouns, the model learns to ground them into the image.

Question II Does masked self-attention in the object relation captioning transformer capture patterns that could have a linguistic interpretation?

As we observe that multi-modality shifts the attention focus of the model to nouns, we also ask whether it has an effect on the knowledge that the model learns about the text. We examine attention on specific part-of-speech tags and syntactic dependency relations, two important sources of knowledge

associated with syntax. Visualising attention targeting words of specific part-of-speech shows that nouns receive attention from many attention heads, which is possibly related to the fact that they can be grounded in visual representations directly; therefore, it is easier to associate them with objects. On the other hand, verbs and adpositions, which can be associated with relations, are not attended to as strongly by the model. In terms of syntactic dependencies, we observe that dependencies that seem to be more important for the task of image captioning are attended to much more strongly across the layers of masked self-attention. The `NUMMOD` (numeral modifier) relation is attended to by many attention heads, possibly because it is important for scene description, e.g., counting and mentioning the number of objects in the image. Dependencies that are often involved in spatial relations such as `POBJ` or `PREP` (“on table”, “bathroom with”) do not receive very strong attention on the words that are in these dependencies. Dependencies which seem to be more relevant for generating grammatically correct descriptions (e.g., `DET`, `COMPOUND`) are generally attended to a lesser degree by the masked self-attention in the multi-modal task setup.

Question III To what extent can we interpret information learned by masked self-attention in relation to other parts of the model, such as cross-modal self-attention? Overall, it appears that masked self-attention in the image captioning transformer learns task-specific semantic knowledge (e.g., grounding of nouns) than a similar attention in text-only transformer. The interpretation of these results should be conducted in the context of the whole model. One explanation is that the focus on nouns is due to cross-modal information fusion happening in other parts of the model. We collect evidence for this hypothesis by inspecting cross-modal self-attention and the connections that it builds between words and objects. We hypothesise that the specific focus on nouns is due to the multi-modal nature of the task and not other factors such as noun frequency in the training data. We first

observe a negative correlation between the frequencies of nouns in image descriptions and masked self-attention on these nouns, while there is a clear positive correlation between the two for the text-only model. We then observe that when a specific noun is about to be generated, the model focuses on the object that is described by this noun. The focus on a specific object changes when a new noun is about to be introduced. When functional words are generated, the model is strongly focused on the noun that can correspond to the last described object. These results suggest that the representations in different self-attention modules of the object relation image captioning transformer are aligned and related to each other, hence the model relies on syntactic cues and learns semantics of nouns because of the cross-modal grounding.

5.1.2.3. Implications for future work

One interesting direction to explore is to gain a better understanding of whether the visual modality interferes and with topic modelling and co-referring that is predicted from sequences. Since the model itself is required to rely on previously generated words in order to produce grammatically and semantically correct continuations, its focus on linguistic information should be preserved in a multi-modal case. Therefore, its attention on nouns can be attributed not necessarily to the multi-modal nature of the task, but to the acquired ability to associate nouns with objects. In other words, the model might either be biased by visual modality to focus on nouns meaning its focus on linguistic modality is less pronounced, or the model might focus on both modalities to the necessary extent and learn a form of grounding. In addition, the model's lack of focus on words involved in spatial relations might be exactly because of a much higher focus on nouns since understanding relations requires access to more information than text alone as is the case with masked self-attention (Ghanimifard and Dobnik, 2019).

5.1.2.4. Author contributions

First version of research questions was developed by Ilinykh who looked at the self-attention in object relation transformer in a course project report. Ilinykh and Dobnik discussed and decided on research questions and experiments for the paper. Ilinykh was responsible for running the experiments and conducting initial analysis. Ilinykh and Dobnik have jointly analysed, discussed and interpreted results of the study. Both authors wrote and approved the final version of the manuscript, where Ilinykh was the main author.

5.1.3. Study II: What Does a Language-And-Vision Transformer See

- **What Does a Language-And-Vision Transformer See: The Impact of Semantic Information on Visual Representations.** Nikolai Ilinykh and Simon Dobnik. 2021. *Frontiers in Artificial Intelligence*, 4. Link: <https://doi.org/10.3389/frai.2021.767971>

5.1.3.1. Overview

The goal of this study is to understand better whether a two-stream object relation image captioning transformer hierarchically learns and structures its self-attention in the language-and-vision context. Here we focus on attention between image objects or regions, which can be extracted from the **image encoder** block in Figure 4.2. This self-attention does not have direct access to textual information. We observe that the task and input feature representations lead to different structures and hierarchical organisation of attended image objects in self-attention. Self-attention patterns can be interpreted to reflect thematic relations and geometric proximity between different objects. In particular, we observe an asymmetry in the type of interpreted knowledge that is distributed differently across earlier and later layers of the self-attention module. We refer to the layers that are closer to the input as “earlier layers”,

while the layers which are closer to the output of the self-attention module are “later layers”. Later layers are also affected by the information about the task of describing an image as such layers appear to learn high-level semantic information between nouns from descriptions and objects in the image. Our results demonstrate that self-attention can be used as basis for learning hierarchically structured knowledge about the objects in the world.

5.1.3.2. Questions and findings

Question I What object-related information is captured in the self-attention weights on the image within the object relation image captioning transformer? Deep neural models excel at detecting patterns and regularities in the inputs they are provided with. As our model is given a set of features of pre-detected objects, we hypothesise that self-attention on the image might learn to relate these objects thematically and semantically. Ultimately, the detector produces objects on different levels of granularity, including whole objects and their parts, such as “cat”, “paw” and “banana” for the image of a cat eating a banana. We need a measure to determine semantic similarity between these bounding boxes. We determine this by thematically clustering labels of objects that are detected with bounding boxes. Then we examine whether self-attention weights in different layers of the model connect bounding boxes that are in the same thematic cluster, which provides us with semantic categories. Based on the empirical results and qualitative analysis of self-attention weights and heatmaps of images, we find that the earlier layers of the model connect objects that are thematically and semantically related. For example, such objects can often be in a part-whole relationship, such as “paw” and “cat”. Later layers of the self-attention that we examine capture thematic relatedness between different objects (e.g., “cat” and “banana”) and not necessarily individual object’s parts. Interestingly, we find that visual features of the objects in the same thematic cluster have a high level of similarity with each other, indicating that these objects are likely in a part-whole relation. We call this

type of knowledge a “thematic bias”. The bias that self-attention learns might not be only thematic but also geometric as semantically similar objects are likely to be visually close to each other. We confirm the presence of such a “geometric bias” by computing the Euclidean distance between pixels of attended objects in the same or different thematic clusters. We find that earlier layers connect thematically similar and geometrically close objects, while later layers connect objects that are geometrically more distant. These results show that self-attention on the image in the object relation image captioning transformer learns more how objects relate to each other in the scene.

Question II How much does the visual feature segmentation affect the structures in self-attention and their interpretation? One important characteristic of the self-attention module we work with is that it operates with features of objects and not patches, which are frequently used in vision tasks with vision transformers (Dosovitskiy et al., 2021). We replace object-level representations with patch-level features and examine whether the hierarchies and structures we have previously identified are still present. Although patches have been used as inputs to image captioning models, the knowledge of image semantics that object detections bring has often resulted in better image descriptions (Anderson, He, et al., 2018). We hypothesise that the pre-defined semantic information that input representations introduce assists the model in hierarchically distributing and organising its self-attention on the visual input. We train and test the model with patch features of the image and analyse distances between the attended objects. Using patches did not result in learning geometric bias since we observe no statistical difference between self-attention weights and the distance between the objects that are connected by these weights. Therefore, semantic information that comes from object detections is indeed an important factor for learning geometric (and, possibly, thematic) hierarchies between image objects as it introduces useful semantic knowledge to the model. This shows that self-attention can learn semantic

connections between visual features given that the input it receives is properly represented to achieve this goal.

Question III What is the effect of the image captioning task on the self-attention on the image? While input representations affect the model's representations in one direction, a different effect might appear in another direction where later layers might be affected by representations from other parts of the model due to backpropagation. The information about text and its grounding in vision could be identified in self-attention on the image. We examine the extent to which attention heads in different layers of the self-attention on the image look at two different objects described by two distinct nouns in the generated caption. These objects are the ones whose labels were linked with noun phrases through the linking mechanism we developed, i.e. the mechanism is described in Section 5.1.1. We find that later layers connect such pairs of objects to a higher degree compared to earlier layers. The result indicates that self-attention on the image captures cross-modal relations between described objects which is consistent with the earlier finding that at that layer attended features are more distant.

5.1.3.3. Implications for future work

Understanding whether knowledge of how the world is structured is captured by the image description transformer is important and introduces many questions. One of them is the role of the computational training task. Does the model learn any structural knowledge about objects if the task were, for example, visual question answering? Analysing weights in a different model trained for other multi-modal tasks is needed to confirm whether our conclusions are generalisable across different architectures.

5.1.3.4. Author contributions

Ilinykh and Dobnik jointly developed research questions. Ilinykh was responsible for running the experiments. Ilinykh and Dobnik have jointly analysed, discussed and interpreted results of the study. Both authors wrote and approved the final version of the manuscript. Ilinykh took the lead in writing the manuscript.

5.1.4. Study III: Attention as Grounding

- **Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer.** Nikolai Ilinykh and Simon Dobnik. 2022. In Findings of the Association for Computational Linguistics: ACL 2022, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.
Link: <https://aclanthology.org/2022.findings-acl.320/>

5.1.4.1. Overview

The last study in this part examines weights of the **cross-modal** self-attention of the object relation image captioning transformer in Figure 4.2. The primary task of this self-attention is to produce a representation that is used to generate image descriptions. This module learns to do so from both linguistic and visual information. Its output is used to generate whole sentences that include two types of words: descriptions of objects and spatial relations. In this study we focus on the image paragraph generation task as longer image descriptions introduce more mentions of entities and relations between them, which allows us to study how this module is responsible for discourse planning.

We inspect self-attention that connects words with image objects and examine how it builds different mappings between descriptions of objects and objects themselves. We additionally analyse masked self-attention on text and its attention on two types of words, i.e., descriptions of objects and

spatial relations. We observe that in later layers self-attention focuses on objects which are described with noun phrases. However, we observe that many attention heads have scattered attention on objects and words when spatial relations are generated. We argue that spatial relations are not only about locating objects (Ghanimifard and Dobnik, 2019), which is supported by the observation that the heatmaps of attention on objects do not have a clear interpretation in terms of knowledge about spatial relations.

5.1.4.2. Questions and findings

Question I What knowledge does cross-modal self-attention acquire about descriptions of objects and relations between them? We examine self-attention heatmaps extracted from the cross-modal self-attention. We investigate whether the patterns formed by attention weights align with our expectation that described objects are attended to by the model, unlike the objects that are not described in the generated paragraph. We first automatically link detected objects and their descriptions, i.e., noun phrases. We also extract triplets of the form “target – relation – landmark” from each description using the spatial relation extractor from (Kolomiyets et al., 2013). The extracted triplets provide us with words that correspond to linked objects (targets, landmarks) that are in a relation. We observe that links between objects and noun phrases are mostly established by attention heads in later layers of this module. We also observe that these heads strongly focus on relating specific word-object pairs. The picture differs when we analyse the focus of self-attention heatmaps on the objects which are in spatial relations. Many different heads are continuously activated when attending to the objects that correspond to either a landmark in a relation (“table”) or a target (“cup” in “cup on the table”). There is no clear structure between earlier or later layers and no specific focus on specific objects.

Question II To what extent can we interpret and explain the behaviour of cross-modal self-attention in generating spatial relations? Examining the heatmaps to identify knowledge about spatial relations in cross-modal self-attention provides us with more insights into what is happening in the model. Interestingly, objects corresponding to targets in spatial relations are often attended to in later layers, while those corresponding to landmarks are also attended to in earlier layers of cross-modal self-attention. This indicates that the model captures some sort of asymmetry about targets and landmarks (and words that describe them) for spatial relation generation (Dobnik, Ghanimifard, et al., 2018). In particular, to describe a target, a good landmark must be chosen first, which is both discourse and visually salient, and then also constrains the set of relations with which they can be related. For example, the spatial relation “on” in “cup on the table” will change if the landmark changes, e.g., “cup next to the phone”. Based on the heatmaps we observe the tendency of the model to focus on the objects that correspond to the desired landmarks earlier, while attending to the target objects later.

Question III What does masked self-attention learn about words corresponding to descriptions of objects and descriptions of object relations? What does the model learn about different semantic categories? We examine patterns captured by the masked self-attention on descriptions of objects and spatial relations. This experiment is different from the one in Study I as the task is image paragraph generation rather than image captioning. By splitting words into two groups – those describing objects (determiners, adjectives, nouns) and relations (verbs, adpositions) – we observe that the earliest layer in masked self-attention strongly focuses on verbs and adpositions, while later layers focus more on nouns, adjectives, and determiners. The focus of masked self-attention in image paragraph generation task differs from the focus of the masked self-attention in image captioning from Study I. In that study the model does not strongly attend to verbs and adpositions but strongly

attends to nouns, determiners, and adjectives, and this result is observed across all layers of the masked self-attention. This difference comes from the distinctions between image captioning and the image paragraph task and the corresponding features provided to the models.

5.1.4.3. Implications for future work

Our results demonstrate how knowledge about object descriptions and spatial relations is learned by cross-modal self-attention in the object relation image captioning transformer. Earlier studies show that if the objects are identified based on their function, i.e., target and landmark objects, and provided to the model, a non-transformer language model identifies them (Ghanimifard and Dobnik, 2019). Overall, our results indicate that grounding spatial relations is not dependent on a particular modality (visually identifying objects) but that information is drawn from several sources. This suggests that information provided to the network in terms of features will play a crucial role in determining what representations are learned, and this question must be investigated in the future.

5.1.4.4. Author contributions

Ilinykh and Dobnik jointly developed research questions. Ilinykh was responsible for developing and running the experiments, including analysis of different linking methods. Ilinykh and Dobnik have jointly analysed, discussed and interpreted results of the study. Both authors wrote and approved the final version of the manuscript, where Ilinykh was the main author.

5.2. Part II: Representation learning for language-and-vision tasks

Studies in this part of the thesis focus on multi-modal representation learning for three tasks: image paragraph generation, embodied question answering and variation in human object naming. The general question that these studies address is the following:

- How are multi-modal representations applied in these tasks and how do task-specific models learn knowledge from linguistic representations of object labels and visual representations of corresponding regions or the scene?

Our models range from a CNN-LSTM image description model to a simple classifier network. These models use visual representations of images or object regions and linguistic representations of descriptions of images or labels of objects. We also test different fusion methods such as max-pooling, attention or concatenation to learn from visual and linguistic representations of objects, labels and images. Studies IV and VI use pre-trained models such as DenseCap (Johnson et al., 2016) or CLIP (Radford, Kim, et al., 2021) to represent images, objects and texts. Study V learns these representations from scratch and the ability of the model to use them is evaluated by performing perturbations to the input representations. We analyse the role of different multi-modal representations of images and texts in three tasks and examine how models for these tasks use these representations.

5.2.1. Motivation

Humans are known to rely on multiple modalities in their understanding of the world. For example, seeing the lips of the speaker helps us distinguish sounds that are very similar (Summerfield, 1992). This ability to operate with many modalities has inspired work that develops cognitively informed models of semantics that are grounded in perception and language (Regier,

1996). Other work has used deep learning methods for multi-modal learning. For example, Ngiam et al. (2011) show that the model learns better features of video modality if it is provided with features of video and audio during learning. Srivastava and Salakhutdinov (2012) introduce a model that learns a single representation from linguistic and visual representations and show that information from this joint representation space is useful for classification and retrieval tasks.

In the domain of natural language processing, attention has been on using different types of deep learning models to learn a multi-modal feature space between language and vision. Work has focused on exploring different model architectures for learning a joint multi-modal feature space. Silberer and Lapata (2014) use stacked auto-encoders to ground representations of texts into images. Kiros et al. (2014) use language models together with convolutional networks to represent texts and images for tasks of text generation and image retrieval. Some methods propose to map images into a text representation space (Frome et al., 2013; Socher et al., 2013).

More recently, the question of learning language-and-vision feature representations has been studied with larger models trained in a multi-task setting (Chen, Li, Yu, et al., 2020; Li, Selvaraju, et al., 2021; Lu, Batra, et al., 2019). These studies propose to first capture general task-agnostic language-and-vision representations through pre-training which are then used in the context of the specific task (fine-tuning). There is also a possibility to initialise models with weights of other pre-trained models as well as encode input representations with existing models. The general process of initialising, pre-training, and fine-tuning multi-modal models can be described as follows:

- Initialisation phase: initialise the weights of the model with pre-trained knowledge of the respective modality. For example, the language-side of a model can be initialised with the weights from a large pre-trained transformer such as BERT (Devlin, Chang, et al., 2019). The input features can also be represented with pre-trained models. For example,

visual features can be extracted from the model pre-trained on image recognition tasks such as Faster R-CNN (Ren, He, et al., 2015).

- Pre-training phase: train the model to learn generic multi-modal representations by training it on tasks such as prediction of masked tokens (masked language modelling), classification of masked object regions or reproduction of their visual features (masked image region modelling), and classifying whether a description and an image match with each other (image-text matching). In this phase, the model is pre-trained in multi-task learning setup, in which it learns generic representations between features of both modalities.
- Fine-tuning phase: the pre-trained model is fine-tuned on a downstream task such as VQA. The downstream task requires the model to learn to apply its general multi-modal knowledge within the context of the specific task. Multi-task learning on many language-and-vision tasks in the pre-training phase has been shown to benefit the performance of models on downstream tasks (Lu, Goswami, et al., 2020).

Bender and Koller (2020) argue that learning from text alone is insufficient for computational modelling of many natural language understanding tasks. One solution is incorporation of modalities other than text, such as sights and sounds, into the modelling paradigm (Bisk et al., 2020), grounding them into one another. However, the concept of multi-modal grounding is challenging to define because there are many tasks, datasets, and modalities and they might all differ in terms of how much of grounded knowledge they require (Chandu et al., 2021). Parcalabescu, Trost, et al. (2021) argue that multi-modality should be understood in the context of the computational task as relevant information and modalities may differ from one task to another.

We highlight two central ideas based on the previous research. First, task-agnostic multi-modal representations improve performance of the models on the downstream tasks. Second, evaluating whether the models learn task-relevant information from different modalities is important. Our stud-

ies contribute to both of these ideas as we look at three different tasks, the corresponding models and multi-modal representations. The primary contributions of Studies IV and VI are about how informative and effective features from pre-trained models (DenseCap (Johnson et al., 2016), CLIP (Radford, Kim, et al., 2021)) are in the context of image paragraph generation and variation in object naming. Study IV contributes to the second idea, evaluating the sensitivity of the question answering system to visual perturbations in the context of the embodied question answering task. Next we describe each task and other contributions of our studies.

Image paragraph generation The task of image paragraph generation was introduced by Krause et al. (2017). The motivation for the task comes from the shortcomings of image captioning and dense captioning tasks. First, one-sentence image descriptions lack details about the image and might not describe all important parts in the image (Karpathy and Fei-Fei, 2015). Second, in the dense captioning task, the region descriptions are highly detailed but, on the other hand, they only describe regions and lack the coherence of image captions (Johnson et al., 2016). The image paragraph generation task addresses both issues because image paragraphs consist of multiple sentences describing images on a fine-grained level and together they form a coherent whole. The paragraphs were produced by human annotators on Amazon Mechanical Turk, and images were taken from MSCOCO (Lin, Maire, et al., 2014) and Visual Genome (Krishna et al., 2017). Krause et al. (2017) generated paragraphs by using the Faster-RCNN-based object detector (Ren, He, et al., 2015) and hierarchical RNN-based model for text generation. First, the detector extracts visual features and object labels of image regions. Next, visual features of detected objects are fed to a sentence-level RNN that makes a classification decision about the number of sentences to generate and also generates a topic vector per sentence. Each topic is given to the word-level RNN that generates the words for the corresponding sentence. By splitting the generation task

between two types of RNNs responsible for different but related tasks, authors ensure that their RNNs learn from sequences of smaller lengths as reasoning over the whole paragraph is challenging for an RNN-based generator.

The multi-modal nature of the image paragraph generation task requires the development of a proper information fusion mechanism that can learn useful information from both language and vision (Baltrusaitis et al., 2019). Linguistic and visual feature vectors can be combined with summation, concatenation, bilinear transformation or other methods (Yang et al., 2019). One of the primary questions is *when* such fusion should take place in the modelling pipeline. Information fusion can occur either early (e.g., at the input feature level) or late (e.g., at the level of the model's output and prediction) (Farnadi et al., 2018). In Study IV we fuse visual features and object label representations early. We first pass each feature through a modality-dependent linear layer and then combine them using one of the two information fusion methods that we also evaluate. The first method is max-pooling, which takes the maximum value from the vector of each modality and concatenates them. Max-pooling can be useful for extracting information about the semantically most important words (Collobert et al., 2011; Kim, 2014). A different pooling method is taking an arithmetic mean of feature vectors of different modalities (Schüz and Zarriß, 2020). However, mean-pooling can smooth out multi-modal features and equalise the effect of each modality on the model's internal representations. Studies have shown that text-based and visual features differ in their relevance for various types of words (Lu, Xiong, et al., 2017). As an alternative to pooling fusion methods we also learn attention on the input features after they are concatenated. We tested a different scenario in which we first attend and then concatenate the resulting features as this has shown improvement on some tasks such as machine translation (Caglayan, Barrault, et al., 2016). However, we observed decrease in performance of the image paragraph generation model.

In Study IV we describe the generation of image paragraphs that are

both accurate and *diverse* in terms of the sentences they include. By diversity we understand the model’s ability to generate many combinations of words in the image description, but that are natural to human interpreters. Otherwise, they are considered noise. Diversity is an important feature of image descriptions produced by humans, and the lack of variability in machine-generated texts is a common issue across many multi-modal computational tasks (van Miltenburg, Elliott, et al., 2018). We hypothesise that using labels of objects alongside their visual features is helpful for the generation of both accurate and diverse paragraphs. Many studies on the related task of image captioning have shown that adding high-level semantic representations of object tags or labels as part of the input to the model helps produce captions that score higher in automatic evaluation (Fang et al., 2015; Gan et al., 2017; Wu, Shen, et al., 2016; You et al., 2016). Image paragraph generation models have been shown to perform better by learning from both the visual features of objects and their individual region-level descriptions (Liang et al., 2017). Existing work has also focused on generating more coherent and consistent image paragraphs (Chatterjee and Schwing, 2018) or learning better topics for individual sentences (Wang and Chan, 2019). We study the question of whether semantic representations of object labels extracted from hidden states of the pre-trained region description model (DenseCap (Johnson et al., 2016)) improve the generation of both accurate and diverse image paragraphs. This question has not been investigated before, and we are specifically interested in transferring knowledge about the semantics of object labels from the model that was specifically trained to capture such representations. These pre-trained semantic representations of object labels can be viewed as general information that the image paragraph generation model can learn from to better describe objects in a multi-sentence text about the image. We also study how uni-modal (vision or language) or multi-modal (vision and language) input to the generator affect the accuracy and diversity of image paragraphs.

Embodied question answering In the embodied question answering (EQA) task (Das, Datta, et al., 2018) a virtual agent is required to answer questions about target object. Both agent and target objects are placed in the visual environment, but in two different locations. The agent first needs to locate the target object by navigating the environment using its perceptual information and history of previous navigation steps. Once the agent decides to stop, it answers the question using the last five image frames in its perceptual history. The EQA dataset that was published in (Das, Datta, et al., 2018) consists of automatically generated questions that ask about the colour, location and place of the target objects. The research on the EQA task has focused on improving the visual capabilities of the agent’s navigation component (Batra et al., 2020; Wijmans et al., 2019). However, the question-answering component of the EQA architecture has not been extensively studied, except for Thomason et al. (2019), who investigated the role of each modality (language or vision) in question answering. The results showed that the question-answering module is capable of answering questions in previously unseen environments using linguistic features alone. In other words, vision is not necessary to correctly answer questions about a target object in the novel environment.

Other tasks have shown that vision is often overlooked by multi-modal models. One prominent example of such task is the Visual Question Answering (VQA) task and corresponding datasets (Antol et al., 2015; Hudson and Manning, 2019; Ren, Kiros, et al., 2015), in which the model is required to answer the question about the image. Zhang, Goyal, et al. (2016) have shown that questions in VQA datasets can be answered correctly without looking at the image due to linguistic biases. By collecting a more balanced dataset where each question is paired with two images that result in two different answers, Goyal et al. (2017) have shown that models struggle to learn from visual information when tested on this dataset. Another example of the task in which vision is important is prediction of colours of common objects (Norlund et al., 2021). Schüz and Zarriß (2020) have additionally shown that the prior infor-

mation about the object itself can help the model to learn predict the colour of the object in the image. Liu, Yin, et al. (2022) and Zhang, Van Durme, et al. (2022) demonstrate that text-only models lack visual commonsense knowledge. Relying on vision is important as it helps overcome the reporting bias in the multi-modal datasets, i.e., humans generally communicate novel information rather than the trivial one (Gordon and Van Durme, 2013), and this might result in unbalanced datasets that the models learn from.

Several studies analysed the behaviour of the VQA models and how they “avoid” looking at the image (Agrawal, Batra, and Parikh, 2016; Agrawal, Batra, Parikh, and Kembhavi, 2018; Kafle and Kanan, 2017; Kafle, Yousefhussien, et al., 2017). Parcalabescu, Gatt, et al. (2021) emphasised the detrimental role of biases in VQA datasets, which prevented models from learning to count objects. More generally, Frank, Bugliarello, et al. (2021) demonstrated that large pre-trained language-and-vision models learn “vision for language” and struggle with properly balancing different modalities required for tasks. Therefore, in Study V we explore the problem of learning from both language and vision to a necessary degree in the context of the EQA task. In particular, we investigate the general role of vision in the EQA task by gradually perturbing visual inputs and examining how much vision is actually used by the question-answering part of the EQA agent. Study V examines the effects that perturbed visual information has on the performance of the EQA model.

Variation in human object naming The third task that we study is the variation in human object naming using the ManyNames dataset (Silberer, Zarrieß, Westera, et al., 2020). Recent work has looked at the contextual factors that affect *human* object naming variation such as the role of visual context (Mädebach et al., 2022) and visual typicality (Gualdoni, Brochhagen, et al., 2023). In their original study Silberer, Zarrieß, Westera, et al. (2020) evaluated the existing pre-trained bottom-up object detector from (Anderson, He, et al., 2018) and whether the labels predicted by this detector are within the

set of possible names for the target object, where the names are provided by the ManyNames dataset. We are interested in computational representations of visual context and linguistic knowledge that can help us to capture variation in human object naming. In Study VI, we use CLIP (Radford, Kim, et al., 2021) to represent labels of objects and their visual features. We study how these representations can be used by a simple classifier to approximate variation in human object naming. While we do not develop a model that predicts object names, we take the first step towards such a model by studying how visual features of objects and semantics of their labels can be computationally represented with the large pre-trained multi-modal model (e.g., CLIP) and used for approximating variation in human object naming.

5.2.2. Study IV: When an Image Tells a Story

- **When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions.** Nikolai Ilinykh and Simon Dobnik. 2020. In Proceedings of the 13th International Conference on Natural Language Generation, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Link: <https://aclanthology.org/2020.inlg-1.40/>

5.2.2.1. Overview

In this study we re-implement and evaluate the image paragraph model proposed by Krause et al. (2017). The model is trained and tested with regard to two aspects: (i) the type of input representations and (ii) the feature fusion mechanism. In the case of the former we provide the model not only with visual features of objects but also with encodings of object labels present in the image. This information is treated as semantic representation of objects. We compare two methods for feature fusion: max-pooling and attention. While max-pooling fuses different vectors by taking the maximum value in each of them and producing a single output, attention learns a type of fusion

by learning to relate different vectors with each other. An example of such multi-modal fusion is the cross-modal self-attention that has been analysed in Study III. The features that we fuse are either (1) multiple visual feature vectors corresponding to objects, (2) multiple feature vectors of object labels, or (3) a concatenation of the two. The role of the fusion method is to learn a compressed and possibly more informative representation from multiple vectors. Fusion is performed on the original feature representations and its result is passed to a sentence-level LSTM.

One important feature of this study is that we explore an LSTM-based image paragraph model, which we re-implement based on Krause et al. (2017). Our contribution is the analysis of the contributions of different input features for the image paragraph generation task and evaluation of different feature fusion mechanisms. We do not focus on the modelling architecture as such. We also use beam search with width 2 to generate paragraphs and do not explore other decoding methods. Our model implementation has one modification to the original model: we do not learn to predict when the paragraph should be finished. Instead we generate as many sentences in the paragraph as found in the ground-truth.

5.2.2.2. Questions and findings

Question I **How do word embeddings of image object labels impact the accuracy and diversity of image paragraphs generated by a CNN-LSTM-based generation model?** We provide our model with both visual features of objects and vector representations of the corresponding labels and train it to generate a paragraph. The results demonstrate that either images or object labels alone are not sufficient to generate image paragraphs of better quality. Automatic evaluation shows that the models generally produce more accurate paragraphs when using a combination of linguistic and visual features. In terms of the evaluation of paragraph diversity, automatic metrics also show

that the model benefits from both word embeddings of object labels and visual features of these objects. However, the results of human evaluation show that humans generally do not favour descriptions generated by the model conditioned on both modalities. They rate paragraphs that are generated by the model that uses embeddings of object labels as its input higher, particularly in terms of sentence structure and text coherence. Also, according to human evaluation, if the fusion method is max-pooling and the model's input is multi-modal, the resulting paragraphs include better word choices and mention salient objects. Overall, embeddings of object labels are useful for generation of paragraphs are similar to human-generated paragraphs in terms of automatic evaluation metrics that focus on accuracy and diversity. Human evaluation suggests that such embeddings are particularly useful for generation of paragraphs with better sentence structure and coherence.

Question II Which of the two fusion mechanisms, max-pooling or attention leads to more natural image paragraphs? While both modalities appear to contribute to the generation of more accurate and diverse paragraphs, the choice of the fusion method is crucial as it determines how the model learns from either uni-modal or multi-modal features. We compare max-pooling and attention as fusion methods. In terms of automatic evaluation, pooling different feature representations by taking the maximum value in every dimension results in more accurate paragraphs, while attention produces more diverse texts. Humans prefer paragraphs generated by the model that uses attention.

Question III How do intrinsic and extrinsic evaluation metrics compare in terms of paragraph model evaluation? We compare the results of automatic and human evaluation. Human language is challenging not only to model but also to evaluate by humans. Automatic evaluation metrics does

not favour texts generated from representations of object labels, instead giving higher scores to texts generated from multi-modal input. Multi-modal input features lead to texts that are closer semantically and syntactically to the ground truth they are compared with. However, humans rank descriptions generated from word embeddings of object labels alone generally much higher than others in terms of word choice, object salience, sentence structure, and paragraph coherence. This becomes especially evident when looking at increases in scores for sentence structure and paragraph coherence, two of the categories in which semantic information is important. This can be explained due to the importance of semantic information captured in embeddings of object labels that is important for such categories as structure of sentences or coherence of texts. This finding demonstrates that automatic and human evaluation assess different aspects of the model's outputs, and hence they should be used together to evaluate different aspects of the generated text.

5.2.2.3. Implications for future work

Our results demonstrate that extracting useful information from linguistic and visual representations is challenging for image paragraph generation. Since humans favoured texts produced from embeddings of objects labels alone, future work should investigate how to better use the visual modality. Different locations and methods for information fusion and their effect on performance need to be investigated as well. The information that is described at the level of a paragraph might not be grounded directly in vision, therefore, representations beyond object labels or visual features should be tested. Decoding methods also impact the quality of the output text and can also be evaluated.

5.2.2.4. Author contributions

Ilinykh and Dobnik have jointly developed research questions. Ilinykh was responsible for re-implementation of the image paragraph model by Krause

et al. (2017). Ilinykh was responsible for implementation of different decoding algorithms. Ilinykh and Dobnik have jointly analysed, discussed and interpreted results of the study. Both authors wrote and approved the final version of the manuscript, where Ilinykh was the main author.

5.2.3. Study V: Look and Answer the Question

- **Look and Answer the Question: On the Role of Vision in Embodied Question Answering.** Nikolai Ilinykh, Yasmeeen Emampoor, and Simon Dobnik. 2022. In Proceedings of the 15th International Conference on Natural Language Generation, pages 236–245, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.
Link: <https://aclanthology.org/2022.inlg-main.19/>

5.2.3.1. Overview

The previous study shows that language-and-vision models are often more biased to language modality and struggle to learn from visual modality. This is very prevalent in the domain of the visual question answering task (Goyal et al., 2017). The current study expands the context of the multi-modal feature representation learning and investigates how vision is used by the question answering module employed for the task of embodied question answering, in which a visual agent has to first navigate to the target object in order to answer a question about it. EQA (Das, Datta, et al., 2018) has two independent sub-tasks of navigation and question answering. The connection between the two is learned by reinforcement learning and there is no guarantee that the agent really sees the object that the question is about. The dataset introduced with the task in Das, Datta, et al. (2018) has image rendering issues, making images incomprehensible to the human eye and, therefore, not suitable for the task. Another problem is that as the questions in the dataset are generated automatically, this introduces biases that raise hallucinations in models. The answers are unnatural because they have been automatically generated from

non-human colour labels and the distribution of colours in the dataset annotations is problematic. We point out that the set of colours used in the dataset was chosen for a different purpose (to emphasise visual contrast) rather than for describing colours in real-world images. One important motivation for our study is to examine in detail how much models can learn from such biased dataset.

Given that images in the EQA dataset have context, content and structure, we remove each of these elements one by one. These perturbations are then used to test how much visual understanding is preserved by the EQA model that is trained on the original images. Our goal is to examine how each of these changes to the visual input affects the performance of the question answering module in the EQA agent to have a better understanding of what type of represented knowledge the model is using. The context of the image is removed by replacing the visual scenes of the environment with a different scene from the dataset. To remove both context and content, we replace the image with a black image (consisting of zeros, thus, keeping some form of structure). Finally, providing the model with random values (noise) eliminates any knowledge from the input, including context, content, and structure.

5.2.3.2. Questions and findings

Question I How does the question answering module of the EQA agent performs with perturbed visual features? We train the question answering model on original data of images-question pairs (**Vis-L**), but test it on data with different visual perturbations. The models are evaluated with accuracy and mean rank across all three types of questions: colour room, colour, location. The model performs the best when it is tested on the original data. The model's performance decreases when it is tested on the scenes from randomly chosen visual environments (**Eval-Shuffled**). Next, the model struggles even more when tested with black images (**Eval-Blind**). Finally, the model's performance

is the lowest when instead of images it's provided with vectors of random noise (**Eval-Random**).

We observe that removing context and content (**Eval-Blind**) is not detrimental as the decrease in performance is much smaller than removing structure alongside context and content (**Eval-Random**). If an image has at least some structure (such as a black image), the model performs well, and its performance will not decrease a lot compared to the performance of the model with original images. For example, **Vis-L** has the mean rank of 10.137 for location questions, **Eval-Blind** has the mean rank of 13.278, and **Eval-Random** has the mean rank of 18.33. The model that uses black images is closer in terms of its performance to the models that use original images. The results suggest that even if the model cannot properly understand the context of the visual scene, it can still use patterns from images that are structurally not varied (zeros for black images). This type of information, together with language, is sufficient for the model to classify for the correct answer. This is possible because the model is using only its internal structures to predict the answer. These structures are learned during training on the automatically generated dataset with three types of questions about visual environment and limited set of answers, where some of the answers are more frequent than others. This behaviour is not desirable as it means that the model is not entirely ineffective in using vision, but it also does not fully understand vision either.

Question II Are question embeddings enough for the question answering module to perform well in the EQA task? We also train and evaluate question answering models on representations in which we gradually remove vision. In terms of the overall accuracy, we observe that the best model uses both images and questions, the second-best model uses only questions and the worst model uses black images and questions. In terms of the mean rank, the picture is different: the best model is the one that uses questions alone, while other two models have worse performance. This question-only model also

achieves the best performance on location questions (e.g., “what room is the chair located in?”). These results suggest that there is a considerable bias in the dataset as the models can successfully hallucinate the answers.

5.2.3.3. Implications for future work

Our work demonstrates that dataset quality is highly impactful on the modelling success of the EQA task. The dataset can be improved by collecting the dataset with natural questions of more than three types and answers which exhibit the variation in human answering. For example, there are multiple shades of brown that could be used by humans to describe a “brown couch”. Modelling such variation from the well-designed dataset is what will improve the EQA task and corresponding models. One interesting research direction that we see is a wider focus on the type of feature representations a question-answering model requires. Perhaps expanding its view of the object or its environment will provide more context to make the right prediction about the target object. Such features (and better dataset quality) could also be helpful in overcoming the dataset bias. Better dataset will result in valid useful information that can be used by the model, unlike the information that is currently provided, i.e., biased answer distribution, problems with navigating to the right room, image rendering problems.

5.2.3.4. Author contributions

The work is based on the master thesis work by Emampoor that Dobnik and Ilinykh co-supervised but the research questions have been expanded and experiments were re-run/re-validated leading to new analyses. The initial codebase has been provided by Emampoor. Ilinykh was responsible for the experiments and re-run of the models. Ilinykh and Dobnik have extensively analysed, discussed and interpreted results of the new analyses. Ilinykh and Dobnik wrote the final version of the manuscript, where Ilinykh was the main

author. Emampoor read the final version and provided comments. All authors approved the final version of the manuscript.

5.2.4. Study VI: Context matters in object naming

- **Context matters: evaluation of target and context features on variation of object naming.** Nikolai Ilinykh and Simon Dobnik. 2023. In Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing, pages 12–24, Ingolstadt, Germany. Association for Computational Linguistics.

Link: <https://aclanthology.org/2023.limo-1.3/>

5.2.4.1. Overview

Yuhas et al. (1989) show that information from different modalities is supplementary rather than complementary in human communication. The interaction of modalities has been explored in studies that show that visual features are used more when there is a certain level of linguistic semantic underspecification (Pezzelle, 2023). In the current study we test different linguistic and visual representations for capturing variation in the human object naming task. The dataset that we use is the ManyNames dataset (Silberer, Zarriß, Westera, et al., 2020) in which humans were shown an image with the target object in red bounding box and were asked to name this object. Each target object has received 36 names from different people. Our task is to capture variation among these names as we test different feature representations extracted with CLIP (Radford, Kim, et al., 2021) model and examine which features results in a better approximation of variation in object naming by a simple classification network.

In our experiments, we represent input to the model in three different conditions. In the first condition (**Target**) the input vector to the classifier includes either a CLIP-based embedding of the label of the target object or its visual feature vector or a combination of both. We take labels of objects

from human annotations in Visual Genome (Krishna et al., 2017) as target objects in ManyNames dataset were chosen based on the annotations in Visual Genome. In the second condition (**Context-as-Objects**) the input vector contains representations of the context objects. We represent context with CLIP-based embeddings of labels of context objects or their visual features. In the third condition (**Context-as-Scene**) we use CLIP to encode whole image as one vector and use representation of a linguistic string describing image as a whole. The string itself consists of a number of relation triplets which are taken from Visual Genome annotations. We feed configurations of features individually to the classifier as its input and learn to predict the name for the object. The classifier produces a probability distribution over all possible labels in the vocabulary. This distribution is used to calculate entropy. The entropy of labels produced by the model and different human describers are correlated to evaluate differences. We compute correlation between entropies with Spearman's rank because entropies predicted by the model and labels assigned by human annotations are different numeric types. We aim to identify such a set of visual features and word embeddings of labels of either target or context objects that allow us to get as close as possible to the variation in human object naming.

We aim to estimate the *variation* among many different humans in object naming. Computing such estimation is challenging as it requires understanding of all the different heuristics that speakers use (Dale and Viethen, 2009), including factors such as the visual typicality of objects (Gualdoni, Brochhagen, et al., 2023) or individual styles of referring (Di Fabbrizio et al., 2008). Cultural background also plays an important role in human object naming. We expect that estimating variation over several speakers will reveal precisely these contextual factors while remove the variable of individual preferences for particular words that would be modelled if we only examine one speaker. In this paper we examine whether the model will produce the same variation as a group of speakers given a particular situational context.

5.2.4.2. Questions and findings

Question I Can CLIP-based image and text embeddings be used to capture variation in human object naming? We find that providing the model with embeddings of labels of target objects (**Target**) results in lower entropy when predicting the target names and higher correlation with variation in human object naming. This result is expected as target object’s labels in Visual Genome are likely to be similar to the names produced by humans in ManyNames dataset. Using CLIP that has a lot of pre-learned general and perceptual knowledge about the target object to encode its label is therefore very informative for the naming classifier. However, a combination of embeddings of target object’s label and its visual features reduces the model’s uncertainty the most and increases correlation with humans in naming. We also find that in the second condition, **Context-as-Objects**, when context objects are represented with embeddings of their labels, the model has higher correlation with humans in naming. Combining these features with visual features of context objects leads to the lowest correlation with humans. This result suggests that CLIP-based representations of context objects’ labels are more informative than their visual features for variation in object naming. Representing context in terms of the whole scene (**Context-as-Scene**) might have its benefits as the model has access to not just object representations, but also to a much broader context, possibly allowing the model to learn more about the scene and object relations. We observe the highest correlation with humans when context in the third condition is represented multi-modally. It is unclear what is the optimal representation of a scene with text, as encoding relation triplets with CLIP to represent the scene produces no correlation with humans. Overall, variation in human object naming can be best approximated with a combination of embeddings of target object’s labels and its visual features encoded with CLIP. Scene-level context representations (third condition) that consist of a combination of a scene visual feature vector and embedding of triplets describing the scene lead to the best model that uses

only knowledge of context objects to approximate variation in human naming.

Examining each feature representation in isolation is useful as it provides a better understanding of what each modality can contribute to the model on each representation level (target, context as objects, context as scene). However, object naming is directly related to the task of referring expression generation, which typically implies that context is essential for reference. Knowledge of context (specifically, visual) allows for the effective identification of the target object with a distinctive referring expression (Reiter and Dale, 2000). Therefore, here we concatenate feature vectors which are informative for capturing variation in their respective feature sets, e.g. target, context as objects, context as image. We concatenate features from such conditions which have shown the best correlation with humans in naming. Therefore, there are three different feature sets to be combined with each other, one per condition. We observe the highest correlation with humans in variation in naming when the model is provided with the concatenation of multi-modal features of the target object (**Target**, condition one) and scene (**Context-as-Scene**, condition three).

5.2.4.3. Implications for future work

Future work should investigate models other than CLIP to encode feature representations. However, studying how CLIP’s own representations contribute to the task is also important in order to understand the contributions that CLIP makes in, for example, **Target** condition: How much does CLIP know about different target objects? Another research direction is to explore more feature combinations as in this study we combined only the best-performing features per condition. We also believe that studying differences within and between different domains (e.g., food, nature, house) is worth investigating, as variation can depend on the domain as well. For example, an object on its own can be named “a cake”, but within a set of other foods it might be named “a dessert”, and we need to capture such shifts computationally in order to

produce more natural names which is useful for a variety of tasks including image captioning and referring expression generation.

5.2.4.4. Author contributions

Ilinykh has developed the first version of the research questions and their relevance for the previous findings. Ilinykh and Dobnik have jointly developed final research questions. Ilinykh was responsible for the experiments. Ilinykh and Dobnik have extensively analysed, discussed and interpreted results of the study. Both authors wrote and approved the final version of the manuscript, where Ilinykh was the main author.

5.3. Part III: Task-specific evaluation of model-generated image descriptions

Studies in this part of the thesis analyse the output of different image description systems. In general, the studies address the following question:

- Do models learn to generate texts that exhibit properties of human-generated texts in tasks such as image paragraph generation and perceptual category description generation?

Study VI in the previous part of the summary has addressed this question for the task of capturing variation in human object naming. We examined how a computational model of object naming learns to replicate a property of human object naming, i.e. variation. Here we explore this question in terms of two different tasks: image paragraph generation and perceptual category description generation and interpretation. We evaluate discourse structure in generated image paragraphs on the text level in Study VII. We focus on noun phrases and their distribution across sentences in the paragraph. We introduce the task of perceptual category description generation and interpretation and develop baseline models in Study VIII. We evaluate generated descriptions of categories for discriminativity and argue that discriminative descriptions of categories are important as in-domain categories might be very similar. In both studies, we use transformer-based architectures based on the standard transformer (Vaswani et al., 2017) or object relation transformer (Herdade et al., 2019). We also employ both deterministic and stochastic decoding methods introduced in Section 4.2. Study VII uses several evaluation metrics, e.g., automatic evaluation with metrics like CIDEr (Vedantam, Lawrence Zitnick, et al., 2015), human evaluation, and evaluation of discourse structure in image paragraphs in noun phrases and their linking with objects in the image. Study VIII uses automatic evaluation metrics but argues that the performance of the model that generates perceptual category descriptions should be evaluated based on the performance of the model that interprets

these descriptions for the category classification task.

5.3.1. Motivation

Discourse structure in image descriptions The topic of discourse coherence in language-and-vision tasks is under-explored. Work like Alikhani, Nag Chowdhury, et al. (2019) and Alikhani, Sharma, et al. (2020) has examined coherent relations in image captions and proposed models for generation of coherent captions. These works were inspired by the insights from discourse coherence theory (Hobbs, 1979). In our analysis we are inspired by the Centering theory (Grosz and Sidner, 1986) with its main proposal that humans generate specific reference patterns to produce a coherent discourse. These reference patterns are represented as a (re)introduction of central entities via referring expressions in different sentences.

The novelty of our research is that we focus on the analysis of discourse structure in the image paragraph generation task, which is a natural test-bed for this analysis because sentences in a paragraph need to form a coherent whole. One of the tasks we study in this thesis is generation of image paragraphs. Study III analyses self-attention in the model that generates image paragraphs. Study IV examines generation of more accurate and diverse image paragraphs. We focus on the *discourse structure* in image paragraphs in terms of choosing and realising a particular set of image entities that will be mentioned across multiple sentences and form a coherent text.

Study VII examines discourse of image paragraphs at the text level. We demonstrate that replicating human-like structure of paragraphs in terms of object referring expressions, noun phrases, their order and attention structure on the image is challenging for models that only representations of visual features of the image and textual features of descriptions. Humans use more information to produce paragraphs, for example, world knowledge (Section 3.1.1). We also show that not every automatic evaluation generation metric is suitable for evaluation of coherence and flow in image paragraphs. But it is not only

the text alone that determines how humans structure the discourse of image descriptions. Images also can have an effect on the structure of a description as they introduce structure of the world. For example, images within the house domain are organised in a particular way and research has shown that when humans describe them, they take the listener “on a tour” (Linde and Goguen, 1980). However, not every house in the world has the same structure and configurations of objects in different room types (kitchens, bathrooms) can vary between communities.

Perceptual category description generation and interpretation People may refer to visual situations that do not directly involve images. Humans can learn to represent concepts as perceptual categories, for example, they might have an idea of how a “penguin” looks like. To talk about penguins we do not need to see one as we can access our concept of a penguin and use it. Rosch et al. (1976) show that human’s conceptual representations of categories depend intra-categorical features and prototypicality effects. Such representations are often used jointly with the knowledge of examples from a category (Blank and Bayer, 2022). Therefore, learning to automatically navigate both category-level representations (more abstract) and exemplar-level representations (grounded in visual information about instances of a category) is essential for natural language generation and interpretation (Silberer, Ferrari, et al., 2017). Study VIII introduces models that use two types of representations of categories: instance-level representations (e.g., visual features of images) or category-level representations which are abstract representations learned per category in the classification task. The study also explores whether texts that are generated from such representations can be used to predict unseen categories.

Evaluation of accuracy and diversity of image descriptions generated by different decoding methods Experiments that analyse texts generated by

language-and-vision models are typically conducted within the task of image captioning (Bernardi et al., 2016). Captions are generated by different decoding methods and we introduce methods that we use in this thesis in Section 4.2. Captions are then evaluated for their **quality** in terms of correspondence to the ground-truth human-generated captions. **Diversity** is another property of human-generated texts and by diverse texts we understand descriptions that consist of combinations of different words, while also being natural to humans. Some methods have been proposed for generation of diverse captions (Ippolito et al., 2019; Lindh et al., 2018; Schüz, Han, et al., 2021). Dai et al. (2017) and van Miltenburg, Elliott, et al. (2018) show that image captioning models struggle to produce varied texts because they explore only the head of the probability distribution due to a maximum likelihood training objective. Generating diverse captions is important for generation of human-like descriptions, because, for example, 99% of image captions in widely used MSCOCO image-caption dataset are unique (Devlin, Cheng, et al., 2015). In addition, more diverse texts that are generated by stochastic decoding methods that rely on sampling are also perceived as more human-like (Meister, Wiher, et al., 2022).

Both quality and diversity are desirable characteristics of texts generated by an image captioning model. However, they are not equally important for all computational tasks that involve text generation (Wiher et al., 2022). Let us look at text-only tasks. In machine translation the objective is to generate a text in the target language that is accurate, grammatical and natural when evaluated against the source language. On the other hand, story generation requires a more open-ended and diverse text to be generated. Decoding methods have a significant effect on the levels of quality and diversity that are exhibited by generated texts and studying the output of decoding methods *across* different tasks is therefore important. This part of the thesis explores what type of texts are generated by common decoding methods when they are employed in image paragraph generation and perceptual category description

generation.

5.3.2. Study VII: Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks?

- **Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation.** Nikolai Ilinykh and Simon Dobnik. 2022. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 480–493, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Link: <https://aclanthology.org/2022.gem-1.45/>

5.3.2.1. Overview

Every image and its description can be interpreted as a story. For example, imagine an image that can be described as follows: “A boy is running to his parents to ask for more ice-cream, because he loves it”. Other descriptions might not tell a story as a sequence of events but instead inform about objects and how they relate to each other by identifying them. For instance, think of an image that is described with the following sequence of sentences: “A car is in front of the office building. There are three pedestrians walking next to it. The street is half-empty”. Both examples can be characterised in terms of the discourse structure that they have: entities and events are organised in a particular order in descriptions and this is our definition of a story. A text with a good discourse structure has patterns in its surface-level organisation, reflected in the choice of words, entities to describe, relations, and, more importantly, their ordering between sentences. By discourse structure we understand a distribution of nouns across sentences in the paragraph. A model-generated paragraph exhibits a good discourse structure when it replicates word choices and order of words in human-generated paragraphs.

In this study we examine two characteristics of such multi-sentence descriptions: structure and discourse in the task of image paragraph generation as in Studies III and IV. Sentences in paragraphs form a discourse and we investigate whether image description models and different decoding methods are capable reproducing such discourse. We train the object relation transformer (Herdade et al., 2019) on the dataset of image paragraphs, Tell-me-more (Ilinykh, Zarri , et al., 2019b), and evaluate the structure of generated texts. We examine how discourse is realised in two types of texts (human-/ and machine-generated) and whether automatically generated paragraphs show the same discourse structure as human-generated ones.

We test several decoding methods that allow to produce different text realisations from the model’s knowledge. We use automatic evaluation metrics such as BLEU (Papineni et al., 2002) to compare generated texts with human texts. Then, human evaluation of content and structure of generated paragraphs is also performed. Automatic and human evaluation results are correlated to examine if any automatic metric evaluates paragraphs on the structural level and not simply on the n-gram level.

We then evaluate structure of discourse in more detail looking at lexical, visual, and attentional characteristics of the texts. In lexical evaluation we compare texts in terms of the noun phrase distribution across different sentences. As noun phrases typically describe objects, this type of evaluation can indirectly highlight a matching or non-matching distribution of nouns between human-/ and machine-generated texts. Visual evaluation is performed in terms of the objects that are described in different sentences. We examine whether noun phrases in sentences can be linked with objects in the image by using the linking method that was used in Studies II, and III. We analyse whether texts generated with different decoding methods contain phrases that can be successfully linked with objects in the image. We also conduct an analysis of how model attends to objects and parts of the scenes when in inference mode interpreting texts generated by itself and humans. This

analysis is focused on the attention structure of humans and models as we examine where in the image they both look at.

5.3.2.2. Questions and findings

Question I Which decoding methods produce image paragraphs most similar to human-generated paragraphs with respect to automatic evaluation metrics? We start by assessing the general quality of generated paragraphs. This type of evaluation shows us how lexically, semantically, and syntactically similar two types of texts are. We use a number of different decoding methods described in Section 4.2. We look at the CIDEr score to determine the decoding method that generates the most human-like descriptions because this metric has been shown to achieve highest correlation with human judgements (Vedantam, Lawrence Zitnick, et al., 2015). The best method is the diverse beam search with width 2, which is one of the deterministic decoding methods. Greedy search also has a very high CIDEr score. Overall, in automatic evaluation, deterministic decoding methods such as greedy, beam, and diverse beam search perform better than sampling-based methods such as ancestral or nucleus sampling. Given that a full paragraph is evaluated as a single item rather than being concatenated from individually generated sentences, the results indicate that the discourse structure of paragraphs generated with, for example, greedy search, is similar to the one in human-generated texts. Ancestral sampling with temperature performs slightly better across all metrics than other sampling-based methods, indicating that temperature can be used to control the randomness in paragraphs.

Deterministic decoding methods such as greedy or beam search consistently generate texts with low diversity across different tasks (Wiher et al., 2022). In our study texts generated by deterministic decoding methods correspond the most to the human ground-truth descriptions based on automatic evaluation. This result means that word combinations in human descriptions are not diverse, thus, images in the dataset are described very similarly to each

other. One of the reasons for this might be the dataset domain: there are limited types of rooms in houses and rooms often share same or similar objects. Therefore, the reason why deterministic decoding methods perform so well in terms of automatic evaluation could be the fact that human descriptions are not diverse. It means that a stochastic algorithm might not be necessary as sampling provides more diversity which would result in paragraphs that deviate from human-generated paragraphs.

We run human evaluation of generated texts for three criteria (relevance, correctness, flow) and compute a correlation with the results of automatic evaluation. Most of the automatic metrics do not correlate with human judgements in relevance and correctness criteria. However, we observe a positive correlation between different variations of BLEU (Papineni et al., 2002) and human judgements of flow in descriptions generated with deterministic decoding methods. This means that the higher the n-gram metric score, the better the flow is in these descriptions according to humans. BLEU analyses texts on the n-gram level and it is not capable of detecting discourse in the context of the whole paragraph. However, as we observe a positive correlation between BLEU and flow as judged by humans, we conclude that the paragraphs that are generated by models **and** humans exhibit flow that can be captured even by such naive evaluation metrics as BLEU. This also hints the reason why stochastic decoding methods do not correlate with human judgements in flow: they over-complicate the task and generate paragraphs with discourse that is more complex than the one found in ground-truth. In addition, top- k sampling with $k = 2$ generates texts for which CIDEr score (Vedantam, Lawrence Zitnick, et al., 2015) shows significant negative correlation. This supports the result that stochastic methods are not suitable for generation of image paragraphs from Tell-me-more dataset (**iliinykh-et al-2019-tell**), possibly because these paragraphs have simple discourse structure that can be better captured with deterministic decoding methods.

Question II How do sentences of model-generated and human-generated paragraphs differ in noun phrases and their distribution across sentences?

We observe that the average number of noun phrases per sentence slowly increases throughout the paragraph in machine-generated texts. In comparison, humans follow an opposite trend: they generate more noun phrases in the first sentence, and the number of these noun phrases decreases with each subsequent sentence. This indicates that decoding methods differ in the way they select noun phrases describing objects compared to how humans sample and place noun phrases in different sentences. This shows that decoding methods do not replicate human object description choice at a lexical level between sentences in a paragraph. Automatic evaluation that was performed to answer the first question has shown that there are decoding methods that generate paragraphs with human-like structure at the paragraph level. But when we analyse the structure on the sentence level in terms of noun phrases and their distribution across sentences, the result is different: all decoding methods tend to generate more noun phrases with each subsequent sentence which is the opposite of what humans do. One possible explanation for this is that decoding methods operate only with sequence probabilities, while humans do take into account other semantic knowledge and context. Another finding is that decoding methods tend to over-generate: they produce more noun phrases per sentence than humans. However, they still might produce acceptable descriptions, different from the structure present in human texts used in the dataset. Next we examine how information described in texts and images relate.

Question III How are descriptions of objects referring to objects in the scene? Do different decoding methods generate texts that refer to those objects generated by humans? We link noun phrases in descriptions with labels of objects in the scene using an automatic algorithm. Then we use Sørensen-Dice coefficient to calculate the overlap of entities referred to in

model and human-generated text. The resulting score is small, indicating that decoding algorithms and humans describe different objects. The highest overlap is observed when greedy and diverse beam searches are used, two of the deterministic-based searches. One reason for this may be that automatically generated texts simply include hallucinations and incorrect object descriptions. Therefore, we also inspect whether the generated texts are grounded in the image. Nearly 50% on average of the generated nouns can be linked with objects in the image, but these objects are different from those described by humans. Sampling-based methods appear to generate noun phrases that are less likely to be linked with labels of objects in the image. This result is expected as sampling is based on random selection of words that are not necessarily grounded in the image.

Question IV Do models and humans focus on the same parts of images?

By automatically linking noun phrases in texts with object labels in images we construct and analyse attention maps which highlight objects mentioned in different texts. The results show that humans typically focus on several objects in the first sentence and then focus on the details about these objects in later sentences, as predicted by the Centering theory (Grosz and Sidner, 1986). In contrast, automatically generated texts often fail to mention objects from different areas of an image in the first sentence. They also struggle to capture potential topic shifts that occur towards the end of human-generated paragraphs. Overall, the attentional discourse structure of automatically generated paragraphs differs from the one observed in human-generated paragraphs.

5.3.2.3. Implications for future work

Our results pave the way for the development of a new evaluation metric that evaluates discourse structure in image paragraphs. Such a metric should focus not only on the evaluation of surface-level patterns in texts but also on the

distribution of different object descriptions across sentences. As Linde and Goguen (1980) show, structure of images can impact discourse structure of corresponding descriptions produced by humans. One possible extension of our work is an automatic examination of whether such effects are observed in machine-generated image descriptions. A different research direction is to examine the complexity of the discourse structure in image paragraphs in the Tell-me-more dataset (Ilinykh, Zarrieß, et al., 2019b). Our study suggests that their discourse structure is easy to predict with deterministic decoding methods, and stochastic decoding methods might not be the most optimal decoding method choice. This result has broader implications for the general question of how complexity of a dataset determines the tools that are suitable to replicate patterns in this dataset.

5.3.2.4. Author contributions

Ilinykh and Dobnik have jointly developed research questions. Ilinykh was responsible for running the experiments and conducting evaluation. Ilinykh and Dobnik have extensively analysed, discussed and interpreted results of the study. Both authors wrote and approved the final version of the manuscript, where Ilinykh was the main author.

5.3.3. Study VIII: Describe Me an Auklet

- **Describe Me an Auklet: Generating Grounded Perceptual Category Descriptions.** Bill Noble* and Nikolai Ilinykh*. 2023. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9330–9347, Singapore. Association for Computational Linguistics. *Equal contribution.
<https://aclanthology.org/2023.emnlp-main.580/>

5.3.3.1. Overview

In this study we construct and evaluate models for the task of perceptual category description generation and interpretation. In the context of our study we define perceptual category as abstract conceptual representations of visual information about objects in the world. Perceptual categories are useful to group objects based on their visual appearance and identify them when compared to different categories. For example, “raven” is one perceptual category which combines knowledge about different instances of raven as a bird. Generating descriptions from either instances or categories is important as humans can produce texts which are either visually grounded or grounded in more abstract representations such as category representations. The latter exemplifies situations where humans do not have immediate access to a visual instance they want to describe and have to rely on their knowledge of the category that this instance belongs to.

We propose a novel task of generating and interpreting descriptions produced from either visual instances or category representations. Our modelling scenario involves two agents. One agent is a generator that produces a description of a category. The generator has knowledge of all perceptual categories, i.e., it has seen at least one instance from each category. The other agent is the interpreter, who lacks knowledge of some categories and is required to predict these unknown categories based on the description from the generator and image of the category. We use descriptions of images of birds from Caltech-UCSH Birds-200 dataset (Wah et al., 2011), which has instances of 200 different bird species. The challenge is to produce descriptions which are useful for the interpreter, who has the knowledge of a subset of categories (seen, 180 categories) and has not seen some of the categories (unseen, 20 categories). We expect the interpreter to rely on similarities between the unknown category that is described and its knowledge of other categories to make a better prediction. The example from the dataset is shown in Figure 5.1.

Both agents are trained separately in a multi-task setup. The generator



Figure 5.1. Example image of the Prairie Warbler category from Caltech-UCSH Birds-200 dataset (Wah et al., 2011). One of the 10 ground-truth descriptions for this image is as follows: “small dark yellow colored bird, with black stripes on his body, with exception of the wings that are brown”.

learns to produce a description, while the interpreter learns to make a category prediction. Each agent also learns to predict labels of categories with two separate classifiers, one per agent. Each classifier takes visual features of images of categories extracted with pre-trained ResNet-101 (Russakovsky et al., 2015). The embeddings of these classifiers are then used as category-level representations for generation of descriptions. The models and tasks that they are trained for are shown in Figure 5.2. We note that during testing both interpreter and generator are provided with the image of an instance from a category, but instances of some of these categories (20 out of 200) were not seen by the interpreter. Therefore, the task can be framed as zero-shot learning.

The generator is based on a transformer architecture (Vaswani et al., 2017). It takes one of three types of representations: (i) visual features of images from a category extracted with pre-trained ResNet-101 (Russakovsky et al., 2015) (instance-level), (ii) embedding of the category represented as a

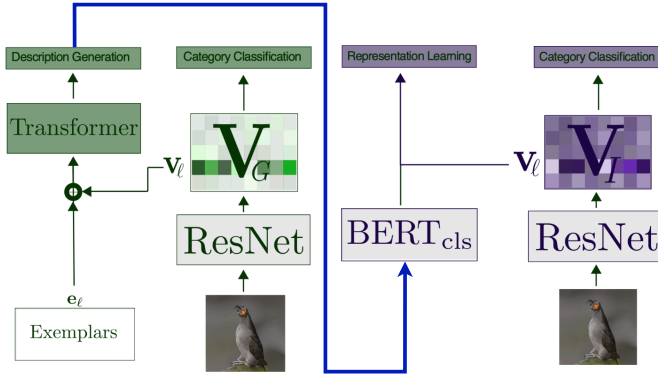


Figure 5.2. The generator model (green) and the interpreter model (purple). The blue arrow shows how descriptions produced by the generator are used by the interpreter.

class embedding from the category classifier (category-level), (iii) or a combination of the two. The interpreter is a classifier that learns category-level representations and predicts category label. It takes visual features of an image of a category to predict the label. It also uses descriptions of the classes produced by the generator as auxiliary information to predict the unseen classes. The descriptions are encoded with BERT (Devlin, Chang, et al., 2019) and we take its CLS token as the representation of description. We fine-tune BERT on descriptions that are seen by both interpreter and generator. The classifier learns category representations which are close to the CLS vector by being trained with cosine embedding loss. This loss prevents the model from learning such representation that is close to the representation of a negative class. Negative classes are classes which do not correspond to the class that the interpreter has to predict.

We emphasise that the task of perceptual category description and generation is useful as a testbed for measuring the communicative success of the generator in terms of the interpretation task. Our contribution is the idea that the performance of the interpreter can be used to measure how optimal the generated category description is. We also focus on generation of descriptions that are discriminative enough to identify a category.

5.3.3.2. Questions and findings

Question I **How is interpretation of perceptual categories affected when descriptions are generated from instance-level features or learned category-level features or the combination of the two?** Our experiments demonstrate that it is challenging to learn category-level features. The interpretation model achieves the highest scores when the generator produces descriptions from instance-level features (e.g., images). Generating descriptions from category-level features has resulted in lower accuracy in the interpreter. The worst interpretation performance is observed when the generator is conditioned on the combination of visual features of images and category-level representations. This result suggests that for interpretation task category-level features are not effective as they do not result in descriptions which are more useful for the interpreter. Instance-level features of images lead to better descriptions which help the interpreter. The performance of the interpreter on the unseen classes is even higher when it is provided with human-generated descriptions of these classes.

Question II **Do discriminativity levels of perceptual category descriptions play a role in the performance of the interpreter, and do automatic evaluation generation metrics favour texts that are also more discriminative?** We evaluate the generated descriptions based on the performance of the interpreter. One of the important properties of natural descriptions of categories is a balance between how accurate and discriminative these descriptions are. Some categories share a lot of similarities and descriptions then need to be more discriminative of these categories. Other categories might be visually distinctive and it might not be necessary to generate highly discriminative category descriptions. The level of *discriminativity* in generated texts matters because class-level descriptions must distinguish the target from others in order for the interpreter to learn about intra-/ and inter-class differences and similarities. Consider “a bird with a long pointy beak and yellow wings” and

“a bird with a beak and wings”. The second description is so general that it can be applied to many categories of birds, while the first description contains mentions of features that are salient for the identification of a specific bird category such as a yellow warbler.

We examine the discriminativity of generated descriptions. We compute the metric by first extracting textual features as noun phrases from descriptions. These noun phrases consists of a noun and one or more adjectives. The discriminativity of a noun phrase is computed as the exponential of the mutual information of how informative this feature is for the category. The resulting score is low if the feature is common among classes and high if the feature is less common, i.e., the feature is unique to specific classes. Comparing evaluation metrics such as BLEU (Papineni et al., 2002) or CIDEr (Vedantam, Lawrence Zitnick, et al., 2015) and our discriminativity metric, we observe the following result. Providing the generation model with category-level representations results in the highest discriminativity scores if texts are generated with nucleus sampling. Using nucleus sampling generally leads to more discriminative texts, which is expected. At the same time, more discriminative texts score low in automatic evaluation. The interpreter performs the best when it is provided with texts which exhibit lower discriminativity. These texts are generated with beam search and conditioned on visual features of categories of instances. We observe a mismatch between the performance of the interpreter and the evaluation scores of the descriptions generated by the generator. Overall, the interpreter benefits from less discriminative descriptions of categories. Image description models that are trained with maximum likelihood objective are known to prioritise certain combinations of words which are present in the training data the most (Dai et al., 2017; Devlin, Gupta, et al., 2015). It means that the models converge on the most common features between image-text pairs in the dataset unlike humans who would prefer more discriminative and more unique, category-/ or instance-specific features to be mentioned in the description.

5.3.3.3. Implications for future work

The results our study indicate that category-level abstractions from images cannot be learned with a simple classification task and a linear layer alone. New experiment using reinforcement learning between two agents similar to Lazaridou et al. (2017) or a more complex variant of the training loss can be used to develop better models for perceptual category description generation and classification. Additionally, more fine-grained category representations such as object-level features instead of image-level features might help the models capture differences and similarities between classes and their instances more effectively. Our study concludes that interpreter model does not require discriminative descriptions of categories in general. More discriminative descriptions might be required for specific categories as some categories are too similar to be identified with a description that contains features that are common among many categories. We emphasise that in order to make a prediction about whether more discriminative descriptions are better for category interpretation, the performance of the model must be compared on a fine-grained category-level and not on the dataset as a whole.

5.3.3.4. Author contributions

Ilinykh trained and evaluated the generation models. Noble trained and evaluated the interpretation models. The task of perceptual category description was developed in close collaboration by both authors. Ilinykh and Noble have extensively discussed and interpreted results of the study. Both authors wrote, read and approved the final version of the manuscript.

Chapter 7: Conclusions and discussion

The studies in this thesis examine several language-and-vision datasets, tasks and deep neural models. We explore how such models process and learn from multi-modal input representations. We also inspect how such representations affect structures that can be extracted and interpreted from the inner mechanisms of models such as self-attention. We look at the textual output of the models, investigating their discourse structure and discriminativity levels important for a task.

7.1. What have we learned from studies?

We outline conclusions from each specific study and describe how they relate to research questions that this thesis answers. The questions were introduced in Chapter 2; we repeat them here:

1. **Research Question I:** What is the role of self-attention in the multi-modal transformer trained for such image description tasks as image captioning and image paragraph generation? Does such self-attention capture representations and structures that match our expectations and findings from research on language and perception? Three studies in Section 5.1 primarily address this question.
2. **Research Question II:** How can multi-modal representations of objects labels and regions be applied in three different tasks such as image paragraph generation, embodied question answering and variation in human object naming? Do models designed for these three tasks learn from such multi-modal representations? Three studies in Section 5.2 address this question.

3. **Research Question III:** What are the properties of human-generated texts that multi-modal models must acquire in the image paragraph generation and perceptual category description and interpretation tasks? Can multi-modal neural models generate texts with similar discourse structure as human-generated texts in the image paragraph generation task? Are models of perceptual categories able to abstract from visual representations and use this knowledge to generate descriptions that exhibit discriminativity levels that are important for the task? Two studies in Section 5.3 answer these questions.

Research Question I is addressed by the following studies. Study I examines structures built by the masked self-attention weights in the **text decoder** of the object relation transformer (Herdade et al., 2019) for image captioning. We find that this self-attention focuses a lot on previously generated nouns. The entropy of multi-modal masked self-attention is low, indicating that it learns the grounding of nouns in images. We compare these patterns with self-attention in the text-only GPT-2 model (Radford, Wu, et al., 2019) which is architecturally similar to the masked multi-modal self-attention. We observe that the model focuses on words that neighbour the word that is being generated. We also find that multi-modal masked self-attention focuses on words in syntactic dependency relations important for describing objects, i.e., the `NUMMOD` relation that can be used to count objects. We observe alignment in self-attention weights between masked self-attention and cross-modal self-attention as, at a particular timestep, cross-modal weights focus on objects which are described by nouns that are attended by masked self-attention.

Study II examines weights in the **image encoder** of the object relation transformer (Herdade et al., 2019) for image captioning. We find that earlier layers of this self-attention connect bounding boxes of thematically similar and geometrically close objects, and later layers relate bounding boxes of objects that are more distant from each other. We also find that providing the model with patch-based representations does not result in the same structures

in self-attention that we observe when the model is provided with bounding boxes of pre-detected objects. We observe that later layers in the self-attention on the image relate objects whose labels can be linked with noun phrases in generated image descriptions.

Study III in the first part of the thesis explores **cross-modal** self-attention in the object relation transformer (Herdade et al., 2019) for image paragraph generation. We observe that attention heads in later layers of cross-modal self-attention learn grounding of nouns into objects. In comparison, we do not find clear patterns of model learning to focus on objects which are described in text in terms of spatial relations. However, we find that the model tends to focus on landmark objects in earlier layers and attends to target objects in later layers, indicating that the model does learn asymmetry about objects which are in spatial relation (Dobnik, Ghanimifard, et al., 2018). We also find that structures in masked self-attention in the object relation transformer trained for image paragraph generation differ from those we observed in Study I, where the model was trained for image captioning. Earlier layers of masked self-attention in the current study focus on verbs and adpositions, and later layers focus on nouns, adjectives, and determiners. In Study I, attention heads in all layers of masked self-attention were mostly focused only on nouns, adjectives, and determiners.

Research question II is addressed by the following studies. Study IV inspects the role of multi-modal representations of objects and their labels in the generation of more accurate and diverse paragraphs by the CNN-LSTM image paragraph model. We find that embeddings of object labels extracted from DenseCap (Johnson et al., 2016) combined with visual representations of objects are helpful for generating paragraphs that are both accurate and diverse as judged by automatic evaluation metrics. Human evaluation suggests that the model generates paragraphs with better sentence structure and coherence from embeddings of object labels alone. Max-pooling as a method to fuse representations from two modalities leads to more accurate

paragraphs. On the other hand, attention generates more diverse texts. We find that humans prefer texts generated by the model that employs attention as an information fusion mechanism. We also find that automatic and human evaluation focus on different aspects of generated paragraphs. Automatic evaluation judges texts generated from multi-modal inputs as semantically and syntactically more similar to the ground truth. Humans rank descriptions generated from embeddings of object labels and attention on them higher across several criteria, i.e., word choice, object salience, sentence structure, paragraph coherence.

Study V examines how much the embodied question answering model learns from visual representations of images of the environment. We test the model's capabilities by performing perturbations to the visual input that the model takes during testing. We observe that any perturbation results in a decrease in the performance of the model. We find that this decrease is smaller when the model is provided with random images of the environment or black images, while random noise leads to the worst performance. We also find that the model does not have to use information from images to answer them as conditioning the model on language-only input results in the best mean answer rank on all questions.

Study VI studies how representations of objects and their labels encoded with CLIP (Radford, Kim, et al., 2021) can be used to build a model that approximates variation in human object naming. We first ask how much each individual feature of target or context objects contributes to learning better approximate variation in human object naming. We find that CLIP-based representations of the label of the target object and its bounding box are the most useful for the model as then the model's entropy scores correlate the most with human variation. If the model is provided only with information about the context, the highest correlation between the model's entropy scores with human variation is achieved when the context is represented by CLIP-based representation of the visual scene and triplets that describe objects in

relations in images. We then ask whether we can better approximate variation by concatenating individual features of target and context objects. We observe that overall the highest correlation with humans in variation in naming is achieved when the model is given multi-modal CLIP-based features of the target object and image as a whole (including the target object).

Research question III is addressed by the following studies. Study VII examines quality and discourse structure of paragraphs generated by humans and models. We conduct several types of evaluation. Automatic evaluation shows that texts generated with deterministic decoding methods (greedy, beam, diverse beam search) better correspond to the ground truth texts than texts generated with stochastic decoding methods (ancestral sampling, nucleus sampling, sampling with temperature). We also observe that scores of the automatic evaluation metrics such as BLEU (Papineni et al., 2002) have significant positive correlation with human judgements about the flow of the information in image paragraphs when deterministic decoding methods are used. This is an indication that stochastic decoding methods introduce too much random variation in image paragraphs and affect their discourse structure. Deterministic decoding methods better capture the discourse structure in paragraphs generated by the model that learns from the Tell-me-more dataset (Ilinykh, Zarrieß, et al., 2019b).

We find that models generate more noun phrases in every next sentence in the paragraph, while humans generate fewer noun phrases in every next sentence. Models also generate more noun phrases than humans per sentence. We find that around half of the nouns that are generated by the model can be grounded in the image. We also observe that models and humans focus on different parts of images in sentences in the paragraphs.

Study VIII introduces the task of perceptual category description generation and interpretation alongside baseline transformer-based models. We build two models: the generator who produces descriptions of categories of birds from either abstract representations learned by classifying birds or visual

features of particular instances of birds. The interpreter learns to use generated descriptions to predict the category label of instances from this category in a zero-shot fashion. We find that the interpreter performs the best when it uses descriptions which are generated from visual features of category instances. We also find that choosing texts which are fed to the interpreter based on the results of automatic generation evaluation leads to the highest performance in interpretation. At the same time, we show that evaluating descriptions for their inter-class discriminativity levels and taking descriptions which are more discriminative results in the decrease in the performance of the interpreter.

7.2. Discussion

Below we provide our interpretation and discussion of the results of studies in this thesis. Overall, there are **three** general outcomes of our research.

7.2.1. Conclusion I: on the role of self-attention

Examining self-attention in the object relation transformer used in image captioning and image paragraph generation shows us that self-attention can capture knowledge of object grounding specific to the multi-modal task (Studies I, II, and III). We complement previous studies on self-attention in text-only tasks and show that self-attention behaves differently in multi-modal tasks (Study I). We also observe that self-attention can learn knowledge about thematic relatedness and visual proximity between objects in the image (Study II); a knowledge type that has not been identified before in multi-modal self-attention. We also see that self-attention corresponds well to human intuitions about spatial language without explicit information about spatial language (Study III), while previous studies focused on LSTM-based models and curated input features that provide the model with spatial knowledge of the scene.

In terms of a broader discussion, self-attention is one of the mechanisms that neural models use to map input modalities with the desired output. The inspiration for building such computational mapping between input and output comes from human performance. For example, we first see an image and then we produce a sentence about it. Based on these assumptions we introduce *inductive biases* in models of language and vision. A convolutional network takes an image and an LSTM network generates a description. Each of these networks is biased to fit modality-specific knowledge because of *inductive biases* incorporated in the mechanisms inside these models. The inductive bias of self-attention inside transformer models, especially a multi-modal one, is less clear. We show that self-attention is a useful mechanism for information fusion as it allows models to identify matching patterns in the data and link information. Our results offer a step forward towards better understanding of self-attention and its role in modelling information from text and images, especially if this is of different sorts and ranges of continuous values from pixels and labels of concepts and words.

My thoughts on this (which Anna agrees with) is that emergent properties really boil down to generalisations and knowledge captured by machine learning algorithms which are biased to do so anyway. These generalise within the biases and hypothesis space they are given but cannot extend beyond and discover new knowledge. (i.e. that has not been in the data, it may be un-transparent to us though but 5at means hidden and not emergent to me!). Hence, attention allows us to provide bias in discovery of such knowledge by providing extra guidance of what patterns connect the modalities.

“Emergent properties” and self-attention Research narrative around the analysis of self-attention has involved the notion of “emergent properties” specifically within vision transformers (Caron et al., 2021). Work in natural language processing has shown that desired knowledge can be controlled for, for example, by analysing the effect of norm growth on the representations

learned in self-attention (Merrill et al., 2021) inside the T5 generative text-only transformer (Raffel et al., 2020). Recent position paper by Luccioni and Rogers (2023) argues that it is important to be clear and precise about definitions, especially in the context of public and scientific narrative including mentions of models acquiring “emergent properties”. The debate has been there due to different definitions of what researchers mean by emergent properties. Our position is that self-attention discovers emergent properties in the sense of uncovering hidden and latent information and associations from patterns in the data. These patterns may not otherwise be directly identifiable by humans. Self-attention might also build intermediate hidden/latent structural representations from the data, but they do not extend beyond such patterns. Hence, the discovery of emergent properties is what is *normally* expected of machine learning. This knowledge is then identified by us because self-attention provides biases for its discovery and guides us towards patterns it builds between different modalities.

7.2.2. Conclusion II: on the role of multi-modal representations

Multi-modal representation learning is highly specific. The work conducted in this thesis demonstrates that tasks often define the type of inputs and their representations that are required. The range of knowledge that humans use is broad as indicated by the example in Section 3.1. Here we test how linguistic and visual representations are processed by models for three multi-modal tasks. We show that using both embeddings of objects labels and their visual features is useful for generation of not only more accurate, but also more diverse image paragraphs (Study IV). We observe that agent that performs embodied question answering task does not use visual features of images when answering questions about objects in virtual environment (Study V). We also find that the agent performs well when conditioned on black images. We

conclude that the EQA task is not well-defined and the corresponding dataset has biases because such non-informative features as black images are useful for the model. We demonstrate that encoding multi-modal information about target object and its image-level context with CLIP (Radford, Kim, et al., 2021) is the best way to computationally model variation in human object naming. Hence, very different features work for individual tasks and a representation that works well for one might not work well for the other. This raises questions about training a single multi-modal language model that performs many tasks since there are multiple ways of doing this, i.e. even the model's training objective will affect representations that it learns.

7.2.3. Conclusion III: on the quality of generated descriptions

Our work analyses different characteristics of automatically generated image descriptions important for two multi-modal tasks (Part 5.3). We demonstrate that different decoding methods for image descriptions fail to replicate structure and organisation of human texts (Study VII). We also illustrate the challenge of generating descriptions for perceptual categories that are both accurate, precise and discriminative at the same time (Study VIII). Overall, we highlight the importance of using a diverse set of evaluation methods, particularly task-specific ones.

Our research on the properties of human-/ and machine-generated texts connects with the question of how to identify a text that is generated by a machine. Recent research on large language models and texts they produce has shown that they often generate disrespectful and hateful language (Bender, Gebru, et al., 2021) and even fake information (Weidinger et al., 2022). These texts are important to identify, and this can also be studied in multi-modal contexts. We propose to identify and examine if models generate texts that exhibit structural properties of texts generated by humans in two multi-modal tasks. One important question to consider is how to identify descriptions that are viable alternatives, but are not ranked high by automatic measures as they

are not identical to ground-truth image descriptions. We need to develop new evaluation tools and methodology that will allow us to distinguish between valid image descriptions and hallucinations.

7.2.4. General conclusion and future work

As novel tasks, datasets and models are being developed, the general trajectory in the multi-modal natural language processing is to build a single model that can perform multiple tasks and learn from multiple data sources. In Section 3.1 we identify different sources of knowledge that are relevant for computational tasks that we study in this thesis. These types of information include world knowledge, perceptual knowledge, and knowledge of intents. While the variety of information sources is important to achieve a more general understanding of language and the world, simply providing more data to a computational multi-modal model is not enough. In our studies described in Chapter 5 and Chapter 6 we examine how models capture such knowledge from computational representations of these types knowledge in embeddings of images, objects in them and their labels. We also look at how models capture such properties of human-generated image descriptions which are relevant for tasks and intents that tasks introduce.

Performance of neural models on multi-modal text generation depends on how the learning is structured and optimised. The importance of input feature representation is hard to neglect as not every task would need the same type of input information. While describing images requires, well, images, discussing penguins in Antarctica might not require their image. On the other hand, navigating in the house in order to find a fork requires a deeper understanding of immediately available visual information, unless you are familiar with the house layout and internal arrangements and can predict where the fork would be without your vision. In the end, what we need are *specialist models* that are not trying to learn *everything* about language and the world, but excel at specific tasks such as generation of longer texts describing

objects and relations in the visual world such as image paragraph generation. The ability to describe the world in detail by means of language is useful in situations where, for example, a machine is placed in the area that suffers from a flood and it is not safe for humans to be there. A precise linguistic description of the environment and if there are any people around and what is their physical state might even save lives in this situation. Finally, let us not diminish the importance of the *general* knowledge of the world as this information is often a foundation for learning a more specific information.

This thesis shows that computational modelling of each language-and-vision task has to be approached with special care. On the surface, computational tasks are identical to human tasks: humans and models generate paragraphs, both also can answer questions about environment and predict a name for an object. However, what makes the crucial difference is how these tasks are performed internally in humans and models. This thesis is inspired by these differences. We offer insights into how multi-modal models can be built, interpreted and analysed based on what we know about human perception and language. Future research should consider these questions and further study them without falsely claiming that current natural language processing and computer vision models are achieving human-level performance across diverse tasks.

Bibliography

Agarap, A. F. (2018). “Deep Learning using Rectified Linear Units (ReLU)”.

In: *CoRR* abs/1803.08375. URL: <http://arxiv.org/abs/1803.08375>.

Agrawal, A., D. Batra, and D. Parikh (2016). “Analyzing the Behavior of Visual Question Answering Models”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by J. Su, K. Duh, and X. Carreras. Austin, Texas: Association for Computational Linguistics, pp. 1955–1960. URL: <https://aclanthology.org/D16-1203>.

Agrawal, A., D. Batra, D. Parikh, and A. Kembhavi (2018). “Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, pp. 4971–4980. URL: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Agrawal%5C_Dont%5C_Just%5C_Assume%5C_CVPR%5C_2018%5C_paper.html.

Alayrac, J.-B. et al. (2022). *Flamingo: A Visual Language Model for Few-Shot Learning*.

Alikhani, M., S. Nag Chowdhury, G. de Melo, and M. Stone (2019). “CITE: A Corpus of Image-Text Discourse Relations”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 570–575. URL: <https://aclanthology.org/N19-1056>.

- Alikhani, M., P. Sharma, et al. (2020). “Cross-modal Coherence Modeling for Caption Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schlueter, and J. Tetreault. Online: Association for Computational Linguistics, pp. 6525–6535. URL: <https://aclanthology.org/2020.acl-main.583>.
- Alikhani, M. and M. Stone (2019). ““Caption” as a Coherence Relation: Evidence and Implications”. In: *Proceedings of the Second Workshop on Shortcomings in Vision and Language*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 58–67. URL: <https://aclanthology.org/W19-1806>.
- Anand, A. et al. (2018). “Blindfold Baselines for Embodied QA”. In: *CoRR* abs/1811.05013. URL: <http://arxiv.org/abs/1811.05013>.
- Anderson, P., B. Fernando, M. Johnson, and S. Gould (2016). “SPICE: Semantic Propositional Image Caption Evaluation”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Vol. 9909. Lecture Notes in Computer Science. Springer, pp. 382–398. URL: https://doi.org/10.1007/978-3-319-46454-1_5C_24.
- Anderson, P., X. He, et al. (2018). “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086.
- Anderson, P., Q. Wu, et al. (2018). “Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.pdf.

- Andreas, J. (2022). “Language Models as Agent Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5769–5779. URL: <https://aclanthology.org/2022.findings-emnlp.423>.
- Antol, S. et al. (2015). “VQA: Visual Question Answering”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433. URL: https://openaccess.thecvf.com/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf.
- Artstein, R. and M. Poesio (2005). *Kappa³ = Alpha (or Beta)*. Tech. rep. Available at: <http://ron.artstein.org/publications/kappa3.pdf>. University of Essex Department of Computer Science.
- Ba, L. J., J. R. Kiros, and G. E. Hinton (2016). “Layer Normalization”. In: *CoRR* abs/1607.06450. URL: <http://arxiv.org/abs/1607.06450>.
- Bahdanau, D., K. Cho, and Y. Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. URL: <http://arxiv.org/abs/1409.0473>.
- Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). “The Berkeley FrameNet Project”. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. Montreal, Quebec, Canada: Association for Computational Linguistics, pp. 86–90. URL: <https://aclanthology.org/P98-1013>.
- Balakrishnan, A. et al. (2019). “Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Lin-*

- guistics*. Florence, Italy: Association for Computational Linguistics, pp. 831–844. URL: <https://aclanthology.org/P19-1080>.
- Baltaretu, A., E. Krahmer, and A. Maes (2019). “Producing Referring Expressions in Identification Tasks and Route Directions: What’s the Difference?” In: *Discourse Processes* 56.2, pp. 136–154. URL: <https://doi.org/10.1080/0163853X.2017.1386522>.
- Baltrusaitis, T., C. Ahuja, and L. Morency (2019). “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.2, pp. 423–443. URL: <https://doi.org/10.1109/TPAMI.2018.2798607>.
- Barzilay, R. and M. Lapata (2008). “Modeling local coherence: An entity-based approach”. In: *Computational Linguistics* 34.1, pp. 1–34.
- Bastings, J. and K. Filippova (2020). “The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?” In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by A. Alishahi et al. Online: Association for Computational Linguistics, pp. 149–155. URL: <https://aclanthology.org/2020.blackboxnlp-1.14>.
- Batra, D. et al. (2020). “ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects”. In: *CoRR* abs/2006.13171. URL: <https://arxiv.org/abs/2006.13171>.
- Beinborn, L., T. Botschen, and I. Gurevych (2018). “Multimodal Grounding for Language Processing”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2325–2339. URL: <https://aclanthology.org/C18-1197>.

- Belinkov, Y. (2018). *On internal language representations in deep learning: an analysis of machine translation and speech recognition*. PhD thesis. Massachusetts Institute of Technology. URL: https://groups.csail.mit.edu/sls/publications/2018/Belinkov_PhD-Thesis_2018.pdf.
- Belinkov, Y. (2022). “Probing Classifiers: Promises, Shortcomings, and Advances”. In: *Computational Linguistics*, pp. 1–13. URL: https://doi.org/10.1162/coli%5C_a%5C_00422.
- Belinkov, Y. and J. Glass (2019). “Analysis Methods in Neural Language Processing: A Survey”. In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. URL: https://doi.org/10.1162/tacl%5C_a%5C_00254.
- Ben-Yosef, G. and S. Ullman (2018). “Image interpretation above and below the object level”. In: *Interface Focus* 8.
- Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. URL: <https://doi.org/10.1145/3442188.3445922>.
- Bender, E. M. and A. Koller (2020). “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. URL: <https://aclanthology.org/2020.acl-main.463>.
- Bengio, S. and Y. Bengio (2000). “Taking on the curse of dimensionality in joint distributions using neural networks”. In: *IEEE Trans. Neural Networks*

- Learn. Syst.* 11.3, pp. 550–557. URL: <https://doi.org/10.1109/72.846725>.
- Berg, A. C. et al. (2012). “Understanding and predicting importance in images”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3562–3569.
- Bernardi, R. et al. (2016). “Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures”. In: *J. Artif. Intell. Res.* 55, pp. 409–442. URL: <https://doi.org/10.1613/jair.4900>.
- Bisk, Y. et al. (2020). “Experience Grounds Language”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8718–8735. URL: <https://aclanthology.org/2020.emnlp-main.703>.
- Blank, H. and J. Bayer (2022). “Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task”. English. In: *COMMUN BIOL* 5.1. © 2022. The Author(s).
- Blevins, T., O. Levy, and L. Zettlemoyer (2018). “Deep RNNs Encode Soft Hierarchical Syntax”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 14–19. URL: <https://aclanthology.org/P18-2003>.
- Bommasani, R. et al. (2021). “On the Opportunities and Risks of Foundation Models”. In: *ArXiv*. URL: <https://crfm.stanford.edu/assets/report.pdf>.
- Botvinick, M. M. (2008). “Hierarchical models of behavior and prefrontal function”. In: *Trends in Cognitive Sciences* 12.5, pp. 201–208. URL: <https://>

[//www.sciencedirect.com/science/article/pii/S1364661308000880](http://www.sciencedirect.com/science/article/pii/S1364661308000880).

- Brennan, S. and H. Clark (1996). “Conceptual Pacts and Lexical Choice in Conversation”. In: *Learning, Memory* 22.6, pp. 1482–1493.
- Bugliarello, E., R. Cotterell, N. Okazaki, and D. Elliott (2021). “Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs”. In: *Transactions of the Association for Computational Linguistics* 9. Ed. by B. Roark and A. Nenkova, pp. 978–994. URL: <https://aclanthology.org/2021.tacl-1.58>.
- Bujwid, S. and J. Sullivan (2021). “Large-Scale Zero-Shot Image Classification from Rich and Diverse Textual Descriptions”. In: *Proceedings of the Third Workshop on Beyond Vision and LANGUAGE: inTEgrating Real-world kNOWLEDge (LANTERN)*. Kyiv, Ukraine: Association for Computational Linguistics, pp. 38–52.
- Caccia, M. et al. (2020). “Language GANs Falling Short”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=BJgza6VtPB>.
- Caglayan, O., L. Barrault, and F. Bougares (2016). “Multimodal Attention for Neural Machine Translation”. In: *arXiv arXiv:1609.03976 [cs.CL]*. URL: <https://arxiv.org/abs/1609.03976>.
- Caglayan, O., P. Madhyastha, L. Specia, and L. Barrault (2019). “Probing the Need for Visual Context in Multimodal Machine Translation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association

- for Computational Linguistics, pp. 4159–4170. URL: <https://aclanthology.org/N19-1422>.
- Cambria, E., Y. Song, H. Wang, and A. Hussain (2011). “Isanette: A Common and Common Sense Knowledge Base for Opinion Mining”. In: *ICDM Workshops*. URL: <https://www.microsoft.com/en-us/research/publication/isanette-a-common-and-common-sense-knowledge-base-for-opinion-mining/>.
- Cao, J. et al. (2020). “Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models”. In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, pp. 565–580. URL: https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123510562.pdf.
- Carey, S. (1981). “The Child as Word Learner”. In: *Linguistic Theory and Psychological Reality*. Ed. by M. Halle, J. Bresnan, and G. A. Miller. First Paperback Edition. Cambridge, Mass.: The MIT Press, pp. 264–293.
- Caron, M. et al. (2021). “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660.
- Castro Ferreira, T., E. Kraemer, and S. Wubben (2016). “Towards more variation in text generation: Developing and evaluating variation models for choice of referential form”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by K. Erk and N. A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 568–577. URL: <https://aclanthology.org/P16-1054>.
- Chai, J. Y. et al. (2018). “Language to Action: Towards Interactive Task Learning with Physical Agents”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint

- Conferences on Artificial Intelligence Organization, pp. 2–9. URL: <https://doi.org/10.24963/ijcai.2018/1>.
- Chandu, K. R., Y. Bisk, and A. W. Black (2021). “Grounding ‘Grounding’ in NLP”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 4283–4305. URL: <https://aclanthology.org/2021.findings-acl.375>.
- Chang, A. et al. (2017). *Matterport3D: Learning from RGB-D Data in Indoor Environments*. cite arxiv:1709.06158. URL: <http://arxiv.org/abs/1709.06158>.
- Chatterjee, M. and A. G. Schwing (2018). “Diverse and Coherent Paragraph Generation from Images”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. URL: https://openaccess.thecvf.com/content_ECCV_2018/papers/Moitreya_Chatterjee_Diverse_and_Coherent_ECCV_2018_paper.pdf.
- Chen, F., R. Ji, et al. (2018). “GroupCap: Group-Based Image Captioning With Structured Relevance and Diversity Constraints”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1345–1353. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Chen_GroupCap_Group-Based_Image_CVPR_2018_paper.pdf.
- Chen, X., H. Fang, et al. (2015). “Microsoft COCO Captions: Data Collection and Evaluation Server”. In: *CoRR* abs/1504.00325. URL: <http://arxiv.org/abs/1504.00325>.
- Chen, Y.-C., L. Li, L. Yu, et al. (2020). “UNITER: UNiversal Image-TEXT Representation Learning”. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. Glas-

- gow, United Kingdom: Springer-Verlag, pp. 104–120. URL: https://doi.org/10.1007/978-3-030-58577-8_7.
- Chen, Y., V. O. Li, K. Cho, and S. Bowman (2018). “A Stable and Effective Learning Strategy for Trainable Greedy Decoding”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 380–390. URL: <https://aclanthology.org/D18-1035>.
- Cheng, J., L. Dong, and M. Lapata (2016). “Long Short-Term Memory-Networks for Machine Reading”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by J. Su, K. Duh, and X. Carreras. Austin, Texas: Association for Computational Linguistics, pp. 551–561. URL: <https://aclanthology.org/D16-1053>.
- Cho, K., B. van Merriënboer, D. Bahdanau, and Y. Bengio (2014). “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Ed. by D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi. Doha, Qatar: Association for Computational Linguistics, pp. 103–111. URL: <https://aclanthology.org/W14-4012>.
- Choi, E., A. Lazaridou, and N. de Freitas (2018). “Compositional Obverter Communication Learning from Raw Visual Input”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=rknt2Be0->.
- Chung, J., Ç. Gülçehre, K. Cho, and Y. Bengio (2014). “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”. In: *CoRR* abs/1412.3555. URL: <http://arxiv.org/abs/1412.3555>.

- Clark, E. V. (2015). “Common Ground”. In: *The Handbook of Language Emergence*. John Wiley and Sons, Ltd. Chap. 15, pp. 328–353. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118346136.ch15>.
- Clark, H. H. and D. Wilkes-Gibbs (1986). “Referring as a collaborative process”. In: *Cognition* 22.1, pp. 1–39. URL: <https://www.sciencedirect.com/science/article/pii/0010027786900107>.
- Clark, K., U. Khandelwal, O. Levy, and C. D. Manning (2019). “What Does BERT Look at? An Analysis of BERT’s Attention”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 276–286. URL: <https://www.aclweb.org/anthology/W19-4828>.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.
- Cohn-Gordon, R., N. Goodman, and C. Potts (2018). “Pragmatically Informative Image Captioning with Character-Level Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 439–443. URL: <https://aclanthology.org/N18-2070>.
- Collobert, R. et al. (2011). “Natural Language Processing (Almost) from Scratch”. In: *J. Mach. Learn. Res.* 12, pp. 2493–2537. URL: <https://dl.acm.org/doi/10.5555/1953048.2078186>.
- Conneau, A. et al. (2018). “What you can cram into a single $\&!#^*$ vector: Probing sentence embeddings for linguistic properties”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne,

- Australia: Association for Computational Linguistics, pp. 2126–2136. URL: <https://aclanthology.org/P18-1198>.
- Cooper, R. (2023). *From Perception to Communication: A Theory of Types for Action and Meaning*. Oxford University Press. URL: <https://doi.org/10.1093/oso/9780192871312.001.0001>.
- Coppock, E. et al. (2020). “Informativity in Image Captions vs. Referring Expressions”. In: *Proceedings of the Probability and Meaning Conference (PaM 2020)*. Ed. by C. Howes, S. Chatzikyriakidis, A. Ek, and V. Somashekarappa. Gothenburg: Association for Computational Linguistics, pp. 104–108. URL: <https://aclanthology.org/2020.pam-1.14>.
- Coventry, K., A. Cangelosi, et al. (2005). “Spatial prepositions and vague quantifiers:: Implementing the functional geometric framework”. English. In: *Spatial Cognition IV*. Ed. by C. Freksa et al. Vol. IV. Error 1 : ISSN or ISBN parsed from 0302-9743 but is invalid for outputType A which is a Book. United States: Springer Nature, pp. 98–110. URL: https://www.researchgate.net/profile/Kenny-Coventry/publication/221104131_Spatial_Prepositions_and_Vague_Quantifiers_Implementing_the_Functional_Geometric_Framework/links/0deec539f3b494d09e000000/Spatial-Prepositions-and-Vague-Quantifiers-Implementing-the-Functional-Geometric-Framework.pdf.
- Coventry, K. and S. Garrod (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Essays in Cognitive Psychology. Taylor & Francis. URL: <https://books.google.se/books?id=rBtJDZFRNU8C>.
- Cun, Y. L. et al. (1990). “Handwritten Digit Recognition with a Back-Propagation Network”. In: *Advances in Neural Information Processing Systems 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 396–404.

- Dai, B., S. Fidler, R. Urtasun, and D. Lin (2017). “Towards Diverse and Natural Image Descriptions via a Conditional GAN”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 2989–2998. URL: <https://doi.org/10.1109/ICCV.2017.323>.
- Dale, R. and J. Viethen (2009). “Referring Expression Generation through Attribute-Based Heuristics”. In: *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 58–65. URL: <https://aclanthology.org/W09-0609>.
- Dale, R. and M. White (2007). “Shared Tasks and Comparative Evaluation in Natural Language Generation”. In: *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*. Ed. by R. Dale and M. White. URL: <https://www.ling.ohio-state.edu/nlgeval07/NLGEval07-Report.pdf>.
- Das, A., S. Datta, et al. (2018). “Embodied Question Answering”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, pp. 1–10. URL: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Das%5C_Embodied%5C_Question%5C_Answering%5C_CVPR%5C_2018%5C_paper.html.
- Das, A., S. Kottur, et al. (2017). “Visual Dialog”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 1080–1089. URL: <https://doi.org/10.1109/CVPR.2017.121>.
- Davis, E. (2017). “Logical formalizations of commonsense reasoning: A survey”. English (US). In: *Journal of Artificial Intelligence Research* 59. Pub-

lisher Copyright: © 2017 AI Access Foundation. All rights reserved., pp. 651–723.

- De Vries, H. et al. (2017). “GuessWhat?! Visual object discovery through multi-modal dialogue”. In: *Conference on Computer Vision and Pattern Recognition*. Honolulu, United States. URL: <https://hal.inria.fr/hal-01549641>.
- Deemter, K. v. (2016). *Computational models of referring: a study in cognitive science*. Cambridge, Massachusetts and London, England: The MIT Press.
- Deerwester, S. C. et al. (1990). “Indexing by Latent Semantic Analysis”. In: *Journal of the American Society of Information Science* 41.6, pp. 391–407.
- DeLucia, A., A. Mueller, X. L. Li, and J. Sedoc (2021). “Decoding Methods for Neural Narrative Generation”. In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Ed. by A. Bosselut et al. Online: Association for Computational Linguistics, pp. 166–185. URL: <https://aclanthology.org/2021.gem-1.16>.
- Demberg, V. and F. Keller (2008). “Data from eye-tracking corpora as evidence for theories of syntactic processing complexity”. In: *Cognition* 109.2, pp. 193–210. URL: <https://www.sciencedirect.com/science/article/pii/S0010027708001741>.
- Denkowski, M. and A. Lavie (2014). “Meteor Universal: Language Specific Translation Evaluation for Any Target Language”. In: *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.
- Desai, K. and J. Johnson (2021). “VirTex: Learning Visual Representations From Textual Annotations”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, pp. 11162–11173. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Desai%5C_VirTex%5C_Learning%5

C_Visual%5C_Representations%5C_From%5C_Textual%5C_Annotations%5C_CVPR%5C_2021%5C_paper.html.

- Deselaers, T. and V. Ferrari (2011). “Visual and semantic similarity in ImageNet”. In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, pp. 1777–1784. URL: <https://doi.org/10.1109/CVPR.2011.5995474>.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423>.
- Devlin, J., H. Cheng, et al. (2015). “Language Models for Image Captioning: The Quirks and What Works”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by C. Zong and M. Strube. Beijing, China: Association for Computational Linguistics, pp. 100–105. URL: <https://aclanthology.org/P15-2017>.
- Devlin, J., S. Gupta, et al. (2015). “Exploring Nearest Neighbor Approaches for Image Captioning”. In: *CoRR abs/1505.04467*. URL: <http://arxiv.org/abs/1505.04467>.
- Di Fabrizio, G., A. J. Stent, and S. Bangalore (2008). “Referring Expression Generation Using Speaker-based Attribute Selection and Trainable Realization (ATTR)”. In: *Proceedings of the Fifth International Natural Language Generation Conference*. Ed. by M. White, C. Nakatsu, and D. McDonald.

- Salt Fork, Ohio, USA: Association for Computational Linguistics, pp. 211–214. URL: <https://aclanthology.org/W08-1133>.
- Divvala, S. K. et al. (2009). “An empirical study of context in object detection”. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, pp. 1271–1278. URL: <https://doi.org/10.1109/CVPR.2009.5206532>.
- Dobnik, S., M. Ghanimifard, and J. Kelleher (2018). “Exploring the Functional and Geometric Bias of Spatial Relations Using Neural Language Models”. In: *Proceedings of the First International Workshop on Spatial Language Understanding*. New Orleans: Association for Computational Linguistics, pp. 1–11. URL: <https://www.aclweb.org/anthology/W18-1401>.
- Dobnik, S., N. Ilinykh, and A. Karimi (2022). “What to refer to and when? Reference and re-reference in two language-and-vision tasks”. In: *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. Dublin, Ireland: SEMDIAL, pp. 146–159. URL: https://semdial2022.github.io/includes/DubDial_Proceedings.pdf.
- Dobnik, S. and J. D. Kelleher (2016). “A Model for Attention-Driven Judgments in Type Theory with Records”. In: *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. New Brunswick, NJ: SEMDIAL, pp. 25–34. URL: http://semdial.org/anthology/Z16-Dobnik_semdial_0007.pdf.
- Dobnik, S. and V. Silfversparre (2021). “The red cup on the left: Reference, coreference and attention in visual dialogue”. In: *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. Potsdam, Germany: SEMDIAL. URL: http://semdial.org/anthology/Z21-Dobnik_semdial_0008.pdf.

- Donahue, J. et al. (2017). “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4, pp. 677–691. URL: <https://doi.org/10.1109/TPAMI.2016.2599174>.
- Dosovitskiy, A. et al. (2021). “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- Du, K. et al. (2023). “Generalizing Backpropagation for Gradient-Based Interpretability”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 11979–11995. URL: <https://aclanthology.org/2023.acl-long.669>.
- Elhoseiny, M., B. Saleh, and A. Elgammal (2013). “Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions”. In: *2013 IEEE International Conference on Computer Vision*. 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, Australia: IEEE, pp. 2584–2591. URL: <http://ieeexplore.ieee.org/document/6751432/> (visited on 04/06/2022).
- Elhoseiny, M., Y. Zhu, H. Zhang, and A. Elgammal (2017). “Link the Head to the “Beak”: Zero Shot Learning from Noisy Text Description at Part Precision”. In: *arXiv:1709.01148 [cs]*.
- Elliott, D. (2018). “Adversarial Evaluation of Multimodal Machine Translation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational

- Linguistics, pp. 2974–2978. URL: <https://aclanthology.org/D18-1329>.
- Elliott, D. and F. Keller (2013). “Image Description using Visual Dependency Representations”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Ed. by D. Yarowsky et al. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1292–1302. URL: <https://aclanthology.org/D13-1128>.
- Elliott, D. and F. Keller (2014). “Comparing Automatic Evaluation Measures for Image Description”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 452–457. URL: <https://aclanthology.org/P14-2074>.
- Elman, J. L. (1990). “Finding Structure In Time”. In: *Cognitive Science* 14, pp. 179–211.
- Erhan, D., Y. Bengio, A. C. Courville, and P. Vincent (2009). “Visualizing Higher-Layer Features of a Deep Network”. In: URL: <https://api.semanticscholar.org/CorpusID:15127402>.
- Fan, A., M. Lewis, and Y. Dauphin (2018). “Hierarchical Neural Story Generation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 889–898. URL: <https://aclanthology.org/P18-1082>.
- Fang, H. et al. (2015). “From captions to visual concepts and back”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 1473–1482. URL: <https://doi.org/10.1109/CVPR.2015.7298754>.

- Farnadi, G., J. Tang, M. De Cock, and M.-F. Moens (2018). “User Profiling through Deep Multimodal Fusion”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, pp. 171–179. URL: <https://doi.org/10.1145/3159652.3159691>.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press. URL: <https://doi.org/10.7551/mitpress/7287.001.0001>.
- Fisch, A. et al. (2020). “CapWAP: Image Captioning with a Purpose”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8755–8768. URL: <https://aclanthology.org/2020.emnlp-main.705>.
- Frank, M. C. and N. D. Goodman (2012). “Predicting Pragmatic Reasoning in Language Games”. In: *Science* 336.6084, pp. 998–998. URL: <https://www.science.org/doi/abs/10.1126/science.1218633>.
- Frank, S., E. Bugliarello, and D. Elliott (2021). “Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 9847–9857. URL: <https://aclanthology.org/2021.emnlp-main.775>.
- Freitag, M. and Y. Al-Onaizan (2017). “Beam Search Strategies for Neural Machine Translation”. In: *Proceedings of the First Workshop on Neural Machine Translation*. URL: <http://dx.doi.org/10.18653/v1/W17-32.07>.

- Fried, D. et al. (2023). *Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches*.
- Frisson, S. (2009). “Semantic Underspecification in Language Processing”. In: *Lang. Linguistics Compass* 3.1, pp. 111–127. URL: <https://doi.org/10.1111/j.1749-818X.2008.00104.x>.
- Frome, A. et al. (2013). “DeViSE: A Deep Visual-Semantic Embedding Model”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.
- Gan, Z. et al. (2017). “Semantic Compositional Networks for Visual Captioning”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 1141–1150. URL: <https://doi.org/10.1109/CVPR.2017.127>.
- Gardner, M. et al. (2020). “Evaluating Models’ Local Decision Boundaries via Contrast Sets”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, pp. 1307–1323. URL: <https://aclanthology.org/2020.findings-emnlp.117>.
- Garrod, S., G. Ferrier, and S. Campbell (1999). “In and On: Investigating the Functional Geometry of Spatial Prepositions”. In: *Cognition* 72.2, pp. 167–189.
- Gatt, A. and E. Krahmer (2017). “Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation”. In: *Journal of AI Research (JAIR)* 61, pp. 75–170. URL: <https://arxiv.org/abs/1703.09902>.

- Gelman, S. A. and A. C. Brandone (2010). “Fast-Mapping Placeholders: Using Words to Talk about Kinds”. In: *Language learning and development : the official journal of the Society for Language Development* 6.3, pp. 223–240.
- Geman, S., D. Potter, and Z. Chi (2002). “Composition systems”. In: *Quarterly of Applied Mathematics* 60.
- Ghader, H. and C. Monz (2017). “What does Attention in Neural Machine Translation Pay Attention to?” In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 30–39. URL: <https://www.aclweb.org/anthology/I17-1004>.
- Ghanimifard, M. and S. Dobnik (2018). “Knowing When to Look For What and Where: Evaluating Generation of Spatial Descriptions with Adaptive Attention”. In: *Computer Vision – ECCV 2018 Workshops. ECCV 2018*. Ed. by L. Leal-Taixé and S. Roth. Vol. 11132. Lecture Notes in Computer Science (LNCS). Proceedings of the Workshop on Shortcomings in Vision and Language (SiVL), ECCV 2018, Munich, Germany: Springer, Cham, pp. 1–9. URL: <https://gup.ub.gu.se/publication/274350?lang=en>.
- Ghanimifard, M. and S. Dobnik (2019). “What goes into a word: generating image descriptions with top-down spatial knowledge”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 540–551. URL: <https://www.aclweb.org/anthology/W19-8668>.
- Gibson, E. A. F. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. UMI Order No. GAX91-26944. PhD thesis. USA: Carnegie Mellon University. URL: https://tedlab.mit.edu/tedlab_website/researchpapers/Gibson_1991_PhDthesis.pdf.

- Gibson, J. J. (1977). "The theory of affordances". In: *Perceiving, acting, and knowing: toward an ecological psychology*. Ed. by J. B. Robert E Shaw. Hillsdale, N.J. : Lawrence Erlbaum Associates, pp.67–82. URL: <https://hal.science/hal-00692033>.
- Giulianelli, M. (2022). "Towards Pragmatic Production Strategies for Natural Language Generation Tasks". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7978–7984. URL: <https://aclanthology.org/2022.emnlp-main.544>.
- Goldberg, Y. (2019). *Assessing BERT's Syntactic Abilities*.
- Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press.
- Gordon, D., A. Kembhavi, et al. (2018). "IQA: Visual Question Answering in Interactive Environments". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4089–4098. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Gordon_IQA_Visual_Question_CVPR_2018_paper.pdf.
- Gordon, J. and B. Van Durme (2013). "Reporting bias and knowledge acquisition". In: *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*. AKBC '13. San Francisco, California, USA: Association for Computing Machinery, pp. 25–30. URL: <https://doi.org/10.1145/2509558.2509563>.
- Goyal, Y. et al. (2017). "Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913. URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Goyal_Making_the_v_CVPR_2017_paper.pdf.

- Graf, C., J. Degen, R. X. D. Hawkins, and N. D. Goodman (2016). “Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions”. In: *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Recognizing and Representing Events, CogSci 2016, Philadelphia, PA, USA, August 10-13, 2016*. Ed. by A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell. cognitivesciencesociety.org. URL: <https://mindmodeling.org/cogsci2016/papers/0392/index.html>.
- Grosz, B. J. and C. L. Sidner (1986). “Attention, intentions, and the structure of discourse”. In: *Computational linguistics* 12.3, pp. 175–204. URL: <http://www.aclweb.org/anthology/J86-3001>.
- Grosz, B. J., A. K. Joshi, and S. Weinstein (1995). “Centering: A Framework for Modeling the Local Coherence of Discourse”. In: *Computational Linguistics* 21.2, pp. 203–225. URL: <https://aclanthology.org/J95-2003>.
- Gu, J., K. Cho, and V. O. Li (2017). “Trainable Greedy Decoding for Neural Machine Translation”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1968–1978. URL: <https://aclanthology.org/D17-1210>.
- Gualdoni, E., T. Brochhagen, A. Mädebach, and G. Boleda (2022). *Woman or tennis player? Visual typicality and lexical frequency affect variation in object naming*. URL: psyarxiv.com/34ckf.
- Gualdoni, E., T. Brochhagen, A. Mädebach, and G. Boleda (2023). “What’s in a name? A large-scale computational study on how competition between names affects naming variation”. In: *Journal of Memory and Language* 133, p. 104459. URL: <https://www.sciencedirect.com/science/article/pii/S0749596X2300058X>.

- Gualdoni, E., A. Madebach, T. Brochhagen, and G. Boleda (2022). “Horse or pony? Visual typicality and lexical frequency affect variability in object naming”. In: *Proceedings of the Society for Computation in Linguistics 2022*. online: Association for Computational Linguistics, pp. 241–243. URL: <https://aclanthology.org/2022.scil-1.25>.
- Gurari, D. et al. (2018). “VizWiz Grand Challenge: Answering Visual Questions From Blind People”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.pdf.
- Haber, J. et al. (2019). “The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1895–1910. URL: <https://aclanthology.org/P19-1184>.
- Hale, J. (2001). “A Probabilistic Earley Parser as a Psycholinguistic Model”. In: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. URL: <https://aclanthology.org/N01-1021>.
- Harnad, S. (1990). “The symbol grounding problem”. In: *Physica D: Nonlinear Phenomena* 42.1, pp. 335–346. URL: <https://www.sciencedirect.com/science/article/pii/0167278990900876>.
- He, K., X. Zhang, S. Ren, and J. Sun (2016). “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
- Herdade, S., A. Kappeler, K. Boakye, and J. Soares (2019). “Image Captioning: Transforming Objects into Words”. In: *Advances in Neural Information*

- Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2019/file/680390c55bbd9ce416d1d69a9ab4760d-Paper.pdf>.
- Hessel, J. et al. (2021). “CLIPScore: A Reference-free Evaluation Metric for Image Captioning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7514–7528. URL: <https://aclanthology.org/2021.emnlp-main.595>.
- Hill, F., K. Cho, and A. Korhonen (2016). “Learning Distributed Representations of Sentences from Unlabelled Data”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 1367–1377. URL: <https://aclanthology.org/N16-1162>.
- Hirota, Y., Y. Nakashima, and N. Garcia (2022). “Gender and Racial Bias in Visual Question Answering Datasets”. In: *FACCT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, pp. 1280–1292. URL: <https://doi.org/10.1145/3531146.3533184>.
- Hobbs, J. R. (1979). “Coherence and Coreference*”. In: *Cognitive Science* 3.1, pp. 67–90. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0301_4.
- Hochreiter, S. and J. Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.
- Hodosh, M., P. Young, and J. Hockenmaier (2013). “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”. In: *J. Artif. Int. Res.* 47.1, pp. 853–899.

- Holtzman, A. et al. (2020). “The Curious Case of Neural Text Degeneration”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=rygGQyrFvH>.
- Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. URL: <https://doi.org/10.5281/zenodo.1212303>.
- Hoover, B., H. Strobel, and S. Gehrmann (2020). “exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 187–196. URL: <https://www.aclweb.org/anthology/2020.acl-demos.22>.
- Hopkins, J. and D. Kiela (2017). “Automatically Generating Rhythmic Verse with Neural Networks”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 168–178. URL: <https://aclanthology.org/P17-1016>.
- Howcroft, D. M. et al. (2020). “Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions”. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Ed. by B. Davis, Y. Graham, J. Kelleher, and Y. Sripada. Dublin, Ireland: Association for Computational Linguistics, pp. 169–182. URL: <https://aclanthology.org/2020.inlg-1.23>.
- Hu, H., J. Gu, et al. (2018). “Relation Networks for Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3588–3597.

- Hu, R., D. Fried, et al. (2019). “Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6551–6557. URL: <https://aclanthology.org/P19-1655>.
- Huang, T.-H. K. et al. (2016). “Visual Storytelling”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 1233–1239. URL: <https://aclanthology.org/N16-1147>.
- Hubel, D. H. and T. N. Wiesel (1959). “Receptive Fields of Single Neurons in the Cat’s Striate Cortex”. In: *Journal of Physiology* 148, pp. 574–591.
- Hudson, D. A. and C. D. Manning (2019). “GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp. 6700–6709. URL: http://openaccess.thecvf.com/content%5C_CVPR%5C_2019/html/Hudson%5C_GQA%5C_A%5C_New%5C_Data%5C_for%5C_Real-World%5C_Visual%5C_Reasoning%5C_and%5C_Compositional%5C_CVPR%5C_2019%5C_paper.html.
- Hupkes, D. et al. (2023). “A taxonomy and review of generalization research in NLP”. In: *Nature Machine Intelligence* 5.10, pp. 1161–1174. URL: <https://doi.org/10.1038/s42256-023-00729-y>.
- Ilinykh, N. and S. Dobnik (2020). “When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions”. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics,

- pp. 338–348. URL: <https://www.aclweb.org/anthology/2020.inlg-1.40>.
- Ilinykh, N. and S. Dobnik (2021). “How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer”. In: *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*. Groningen, Netherlands (Online): Association for Computational Linguistics, pp. 45–55. URL: <https://www.aclweb.org/anthology/2021.mmsr-1.5>.
- Ilinykh, N. and S. Dobnik (2022). “Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 4062–4073. URL: <https://aclanthology.org/2022.findings-acl.320>.
- Ilinykh, N., Y. Emampoor, and S. Dobnik (2022). “Look and Answer the Question: On the Role of Vision in Embodied Question Answering”. In: *Proceedings of the 15th International Conference on Natural Language Generation*. Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics, pp. 236–245. URL: <https://aclanthology.org/2022.inlg-main.19>.
- Ilinykh, N., S. Zarrieß, and D. Schlangen (2018). “The Task Matters: Comparing Image Captioning and Task-Based Dialogical Image Description”. In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics, pp. 397–402. URL: <https://aclanthology.org/W18-6547>.
- Ilinykh, N., S. Zarrieß, and D. Schlangen (2019a). “Meet Up! A Corpus of Joint Activity Dialogues in a Visual Environment”. In: *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

London, United Kingdom: SEMDIAL. URL: http://semdial.org/anthology/Z19-Ilinykh_semdial_0006.pdf.

Ilinykh, N., S. Zarrieß, and D. Schlangen (2019b). “Tell Me More: A Dataset of Visual Scene Description Sequences”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 152–157. URL: <https://aclanthology.org/W19-8621>.

Inan, M. et al. (2021). “COSMic: A Coherence-Aware Generation Metric for Image Descriptions”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3419–3430. URL: <https://aclanthology.org/2021.findings-emnlp.291>.

Ippolito, D. et al. (2019). “Comparison of Diverse Decoding Methods from Conditional Language Models”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3752–3762. URL: <https://aclanthology.org/P19-1365>.

Jain, S. and B. C. Wallace (2019). “Attention is not Explanation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. URL: <https://www.aclweb.org/anthology/N19-1357>.

Jawahar, G., B. Sagot, and D. Seddah (2019). “What Does BERT Learn about the Structure of Language?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association

- for Computational Linguistics, pp. 3651–3657. URL: <https://aclanthology.org/P19-1356>.
- Jia, C. et al. (2021). “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 4904–4916. URL: <https://proceedings.mlr.press/v139/jia21b.html>.
- Jiang, M. et al. (2019). “TIGer: Text-to-Image Grounding for Image Caption Evaluation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2141–2152. URL: <https://aclanthology.org/D19-1220>.
- Johnson, J., A. Karpathy, and L. Fei-Fei (2016). “DenseCap: Fully Convolutional Localization Networks for Dense Captioning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Jokinen, K. (1996). “Goal Formulation based on Communicative Principles”. In: *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. URL: <https://aclanthology.org/C96-2101>.
- Kafle, K. and C. Kanan (2017). “An Analysis of Visual Question Answering Algorithms”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 1983–1991. URL: <https://doi.org/10.1109/ICCV.2017.217>.
- Kafle, K., M. Yousefhussien, and C. Kanan (2017). “Data Augmentation for Visual Question Answering”. In: *Proceedings of the 10th International Conference on Natural Language Generation*. Santiago de Compostela,

- Spain: Association for Computational Linguistics, pp. 198–202. URL: <https://aclanthology.org/W17-3529>.
- Karpathy, A. and L. Fei-Fei (2015). “Deep visual-semantic alignments for generating image descriptions”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 3128–3137. URL: <https://doi.org/10.1109/CVPR.2015.7298932>.
- Kazemzadeh, S., V. Ordonez, M. Matten, and T. Berg (2014). “ReferItGame: Referring to Objects in Photographs of Natural Scenes”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 787–798. URL: <https://aclanthology.org/D14-1086>.
- Kelleher, J. D. and S. Dobnik (2019). “Referring to the recently seen: reference and perceptual memory in situated dialogue”. In: *CLASP Papers in Computational Linguistics: Dialogue and Perception – Extended papers from DaP-2018 Gothenburg*, pp. 41–50. URL: <http://hdl.handle.net/2077/63998>.
- Kelleher, J. D. and S. Dobnik (n.d.). “What is not where: the challenge of integrating spatial representations into deep learning architectures”. In: *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12–13 June*. CLASP Papers in Computational Linguistics, pp. 41–52. URL: <https://gup.ub.gu.se/publication/262970?lang=en>.
- Kelly, K. L. (1965). “Twenty-two colors of maximum contrast”. In: *Color Engineering* 3, pp. 26–27. URL: http://www.iscc-archive.org/pdf/PC54_1724_001.pdf.

- Kennington, C. and D. Schlangen (2015). “Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 292–301. URL: <https://aclanthology.org/P15-1029>.
- Kiddon, C., L. Zettlemoyer, and Y. Choi (2016). “Globally Coherent Text Generation with Neural Checklist Models”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 329–339. URL: <https://aclanthology.org/D16-1032>.
- Kilickaya, M., A. Erdem, N. Ikizler-Cinbis, and E. Erdem (2017). “Re-evaluating Automatic Metrics for Image Captioning”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 199–209. URL: <https://www.aclweb.org/anthology/E17-1019>.
- Kim, Y. (2014). “Convolutional Neural Networks for Sentence Classification”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, and W. Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1746–1751. URL: <https://aclanthology.org/D14-1181>.
- Kingma, D. P. and J. Ba (2015). “Adam: A Method for Stochastic Optimization.” In: *ICLR (Poster)*. Ed. by Y. Bengio and Y. LeCun. URL: <http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14>.
- Kiros, R., R. Salakhutdinov, and R. Zemel (2014). “Multimodal Neural Language Models”. In: *Proceedings of the 31st International Conference on*

- Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research. Beijing, China: PMLR, pp. 595–603. URL: <http://proceedings.mlr.press/v32/kiros14.html>.
- Klein, G. et al. (2017). “OpenNMT: Open-Source Toolkit for Neural Machine Translation”. In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, pp. 67–72. URL: <https://www.aclweb.org/anthology/P17-4012>.
- Kobayashi, G., T. Kuribayashi, S. Yokoi, and K. Inui (2020). “Attention is Not Only a Weight: Analyzing Transformers with Vector Norms”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7057–7075. URL: <https://aclanthology.org/2020.emnlp-main.574>.
- Kolomiyets, O., P. Kordjamshidi, M.-F. Moens, and S. Bethard (2013). “SemEval-2013 Task 3: Spatial Role Labeling”. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 255–262. URL: <https://www.aclweb.org/anthology/S13-2044>.
- Kong, C. et al. (2014). “What are You Talking About? Text-to-Image Coreference”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3565. URL: https://openaccess.thecvf.com/content_cvpr_2014/papers/Kong_What_are_You_2014_CVPR_paper.pdf.
- Kousha, S. and M. A. Brubaker (2021). *Zero-shot Learning with Class Description Regularization*. URL: <https://arxiv.org/abs/2106.16108>.

- Krahmer, E. and K. van Deemter (2012). “Computational Generation of Referring Expressions: A Survey”. In: *Computational Linguistics* 38.1, pp. 173–218. URL: <https://aclanthology.org/J12-1006>.
- Krause, J., J. Johnson, R. Krishna, and L. Fei-Fei (2017). “A Hierarchical Approach for Generating Descriptive Image Paragraphs”. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 317–325. URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Krause_A_Hierarchical_Approach_CVPR_2017_paper.pdf.
- Kreiss, E. et al. (2022). “Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 4685–4697. URL: <https://aclanthology.org/2022.emnlp-main.309>.
- Krishna, R. et al. (2017). “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations”. In: *Int. J. Comput. Vision* 123.1, pp. 32–73. URL: <https://doi.org/10.1007/s11263-016-0981-7>.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., pp. 1097–1105. URL: <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
- Kruijff-Korbayová, I. et al. (2015). “TRADR Project: Long-Term Human-Robot Teaming for Robot Assisted Disaster Response”. In: *KI - Künstliche Intelligenz* 29.2, pp. 193–201. URL: <https://hal.archives-ouvertes.fr/hal-01143484>.

- Kulikov, I., A. Miller, K. Cho, and J. Weston (2019). “Importance of Search and Evaluation Strategies in Neural Dialogue Modeling”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 76–87. URL: <https://aclanthology.org/W19-8609>.
- Kulkarni, G., V. Premraj, S. Dhar, et al. (2011). “Baby talk: Understanding and generating simple image descriptions”. In: *CVPR 2011*, pp. 1601–1608.
- Kulkarni, G., V. Premraj, V. Ordonez, et al. (2013). “BabyTalk: Understanding and Generating Simple Image Descriptions”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.12, pp. 2891–2903. URL: <https://doi.org/10.1109/TPAMI.2012.162>.
- Kusner, M. J., Y. Sun, N. I. Kolkin, and K. Q. Weinberger (2015). “From Word Embeddings To Document Distances”. In: *ICML*.
- Kuznetsova, P. et al. (2012). “Collective Generation of Natural Image Descriptions”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by H. Li et al. Jeju Island, Korea: Association for Computational Linguistics, pp. 359–368. URL: <https://aclanthology.org/P12-1038>.
- Lake, B. M., T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman (2017). “Building machines that learn and think like people”. In: *Behavioral and Brain Sciences* 40, e253.
- Lample, G., A. Conneau, L. Denoyer, and M. Ranzato (2018). “Unsupervised Machine Translation Using Monolingual Corpora Only”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=rkYTTf-AZ>.

- Larsson, S. (2013). “Formal Semantics for Perceptual Classification”. In: *Journal of Logic and Computation* 25.2, pp. 335–369.
- Larsson, S. (2018). “Grounding as a Side-Effect of Grounding”. In: *Topics in Cognitive Science* 10.2, pp. 389–408.
- Lavie, A. and A. Agarwal (2007). “METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments”. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, pp. 228–231. URL: <https://aclanthology.org/W07-0734>.
- Lavie, N., A. Hirst, J. Fockert, and E. Viding (2004). “Load Theory of Selective Attention and Cognitive Control”. In: *Journal of experimental psychology. General* 133, pp. 339–54.
- Lazaridou, A., A. Peysakhovich, and M. Baroni (2017). “Multi-Agent Cooperation and the Emergence of (Natural) Language”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=Hk8N3ScIlg>.
- LeCun, Y. et al. (1989). “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1, pp. 541–551.
- Levinson, S. C. (2003). *Space in language and cognition: explorations in cognitive diversity*. Cambridge: Cambridge University Press.
- Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell.
- Li, G., L. Zhu, P. Liu, and Y. Yang (2019). “Entangled Transformer for Image Captioning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8928–8937. URL: <https://openaccess.th>

ecvf.com/content_ICCV_2019/papers/Li_Entangled_Transformer_for_Image_Captioning_ICCV_2019_paper.pdf.

- Li, J., M. Galley, et al. (2016). “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 110–119. URL: <https://aclanthology.org/N16-1014>.
- Li, J. and D. Jurafsky (2016). “Mutual Information and Diverse Decoding Improve Neural Machine Translation”. In: *CoRR* abs/1601.00372. URL: <http://arxiv.org/abs/1601.00372>.
- Li, J., D. Li, S. Savarese, and S. C. H. Hoi (2023). “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”. In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by A. Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 19730–19742. URL: <https://proceedings.mlr.press/v202/li23q.html>.
- Li, J., D. Li, C. Xiong, and S. C. H. Hoi (2022). “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by K. Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 12888–12900. URL: <https://proceedings.mlr.press/v162/li22n.html>.
- Li, J., R. Selvaraju, et al. (2021). “Align before Fuse: Vision and Language Representation Learning with Momentum Distillation”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 9694–9705. URL: <https://proceedings.n>

- eurips.cc/paper_files/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.
- Li, L., S. Tang, et al. (2017). “Image Caption with Global-Local Attention”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11236>.
- Li, X., X. Yin, et al. (2020). “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks”. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J. Frahm. Vol. 12375. Lecture Notes in Computer Science. Springer, pp. 121–137. URL: https://doi.org/10.1007/978-3-030-58577-8%5C_8.
- Liang, X. et al. (2017). “Recurrent Topic-Transition GAN for Visual Paragraph Generation”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Lin, C.-Y. (2004). “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: <https://aclanthology.org/W04-1013>.
- Lin, D., C. Kong, S. Fidler, and R. Urtasun (2015). “Generating Multi-Sentence Lingual Descriptions of Indoor Scenes”. In: *arXiv arXiv:1503.00064 [cs.CV]*. URL: <https://arxiv.org/abs/1503.00064>.
- Lin, M., H. Lucas, and G. Shmueli (2013). “Too big to fail: Large samples and the p-value problem”. In: *Information Systems Research* 24.4, pp. 906–917.
- Lin, T., M. Maire, et al. (2014). “Microsoft COCO: Common Objects in Context”. In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Ed. by D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Vol. 8693. Lecture Notes in Computer

- Science. Springer, pp. 740–755. URL: https://doi.org/10.1007/978-3-319-10602-1%5C_48.
- Linde, C. and J. Goguen (1980). “On the Independence of Discourse Structure and Semantic Domain”. In: *18th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 35–37. URL: <https://aclanthology.org/P80-1010>.
- Lindh, A. et al. (2018). “Generating Diverse and Meaningful Captions - Unsupervised Specificity Optimization for Image Captioning”. In: *Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I*. Ed. by V. Kurková et al. Vol. 11139. Lecture Notes in Computer Science. Springer, pp. 176–187. URL: https://doi.org/10.1007/978-3-030-01418-6%5C_18.
- Liu, H., C. Li, Q. Wu, and Y. J. Lee (2023). “Visual Instruction Tuning”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 34892–34916. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.
- Liu, X., D. Yin, Y. Feng, and D. Zhao (2022). “Things not Written in Text: Exploring Spatial Commonsense from Visual Signals”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 2365–2376. URL: <https://aclanthology.org/2022.acl-long.168>.
- Lloyd, S. (1982). “Least squares quantization in PCM”. In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137.

- Lu, J., D. Batra, D. Parikh, and S. Lee (2019). “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf>.
- Lu, J., V. Goswami, et al. (2020). “12-in-1: Multi-Task Vision and Language Representation Learning”. In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lu, J., C. Xiong, D. Parikh, and R. Socher (2017). “Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 375–383. URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Lu_Knowing_When_to_CVPR_2017_paper.pdf.
- Luccioni, A. and A. Rogers (2023). *Mind Your Language (Model): Fact-Checking LLMs and Their Role in NLP Research and Practice*. English. Other.
- Luo, R., B. L. Price, S. Cohen, and G. Shakhnarovich (2018). “Discriminability Objective for Training Descriptive Captions”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6964–6974. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Luo_Discriminability_Objective_for_CVPR_2018_paper.pdf.
- Luo, Y., P. Banerjee, et al. (2022). “To Find Waldo You Need Contextual Cues: Debiasing Who’s Waldo”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 355–361. URL: <https://aclanthology.org/2022.acl-short.39>.

- Luong, T., H. Pham, and C. D. Manning (2015). “Effective Approaches to Attention-based Neural Machine Translation”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by L. Màrquez, C. Callison-Burch, and J. Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421. URL: <https://aclanthology.org/D15-1166>.
- Mädebach, A., E. Torubarova, E. Gualdoni, and G. Boleda (2022). *Effects of task and visual context on referring expressions using natural scenes*. URL: psyarxiv.com/fyzsk.
- Madhyastha, P., J. Wang, and L. Specia (2019). “VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6539–6550. URL: <https://aclanthology.org/P19-1654>.
- Malt, B. C. (1989). “An On-Line Investigation of Prototype and Exemplar Strategies in Classification”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15, pp. 539–555.
- Mann, H. B. and D. R. Whitney (1947). “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1, pp. 50–60. URL: <https://doi.org/10.1214/aoms/1177730491>.
- Mareček, D. and R. Rosa (2019). “From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 263–275. URL: <https://www.aclweb.org/anthology/W19-4827>.

- Massarelli, L. et al. (2020). “How Decoding Strategies Affect the Verifiability of Generated Text”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 223–235. URL: <https://aclanthology.org/2020.findings-emnlp.22>.
- Medin, D. L. and M. M. Schaffer (1978). “Context Theory of Classification Learning”. In: *Psychological Review* 85, pp. 207–238.
- Meister, C., M. Forster, and R. Cotterell (2021). “Determinantal Beam Search”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6551–6562. URL: <https://aclanthology.org/2021.acl-long.512>.
- Meister, C., G. Wiher, T. Pimentel, and R. Cotterell (2022). “On the probability-quality paradox in language generation”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 36–45. URL: <https://aclanthology.org/2022.acl-short.5>.
- Melas-Kyriazi, L., A. Rush, and G. Han (2018). “Training for Diversity in Image Paragraph Captioning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 757–761. URL: <https://www.aclweb.org/anthology/D18-1084>.
- Merrill, W. et al. (2021). “Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Online

- and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1766–1781. URL: <https://aclanthology.org/2021.emnlp-main.133>.
- Mikolov, T., I. Sutskever, et al. (2013). “Distributed representations of words and phrases and their compositionality”. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. URL: <http://arxiv.org/abs/1301.3781>.
- Miller, G. A. (1995). “WordNet: A Lexical Database for English”. In: *Commun. ACM* 38.11, pp. 39–41. URL: <https://doi.org/10.1145/219717.219748>.
- Miller, G. A. and P. N. Johnson-Laird (1976). *Language and perception*. Cambridge: Cambridge University Press.
- Minsky, M. (2000). “Commonsense-based interfaces”. In: *Commun. ACM* 43.8, pp. 66–73. URL: <https://doi.org/10.1145/345124.345145>.
- Mitchell, J. and M. Lapata (2008). “Vector-based Models of Semantic Composition”. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, pp. 236–244. URL: <https://aclanthology.org/P08-1028>.
- Mitchell, M., J. Dodge, et al. (2012). “Midge: Generating Image Descriptions From Computer Vision Detections”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by W. Daelemans. Avignon, France: Association for Computational Linguistics, pp. 747–756. URL: <https://aclanthology.org/E12-1076>.

- Mitchell, M., E. Reiter, and K. van Deemter (2013). “Typicality and Object Reference”. In: *Cognitive Science* 35, pp. 3062–3067.
- Narayan, S. et al. (2022). “A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1319–1339. URL: <https://aclanthology.org/2022.acl-long.94>.
- Ngiam, J. et al. (2011). “Multimodal deep learning”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning. ICML’11*. Bellevue, Washington, USA: Omnipress, pp. 689–696.
- Niles, I. and A. Pease (2001). “Towards a standard upper ontology”. In: *2nd International Conference on Formal Ontology in Information Systems, FOIS 2001, Ogunquit, Maine, USA, October 17-19, 2001, Proceedings*. ACM, pp. 2–9. URL: <https://doi.org/10.1145/505168.505170>.
- Nishimura, T., A. Hashimoto, and S. Mori (2019). “Procedural Text Generation from a Photo Sequence”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 409–414. URL: <https://aclanthology.org/W19-8650>.
- Norlund, T., L. Hagström, and R. Johansson (2021). “Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?” In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by J. Bastings et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 149–162. URL: <https://aclanthology.org/2021.blackboxnlp-1.10>.

- Nosofsky, R. M. (1984). "Choice, Similarity, and the Context Theory of Classification". In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 10.1, pp. 104–114.
- Oroojlooy, A. and D. Hajinezhad (2021). "A Review of Cooperative Multi-Agent Deep Reinforcement Learning". In: *arXiv:1908.03963 [cs, math, stat]*.
- Panagiaris, N., E. Hart, and D. Gkatzia (2020). "Improving the Naturalness and Diversity of Referring Expression Generation models using Minimum Risk Training". In: *Proceedings of the 13th International Conference on Natural Language Generation*. Ed. by B. Davis, Y. Graham, J. Kelleher, and Y. Sripada. Dublin, Ireland: Association for Computational Linguistics, pp. 41–51. URL: <https://aclanthology.org/2020.inlg-1.7>.
- Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. URL: <https://www.aclweb.org/anthology/P02-1040>.
- Parcalabescu, L., A. Gatt, A. Frank, and I. Calixto (2021). "Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks". In: *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*. Ed. by L. Donatelli, N. Krishnaswamy, K. Lai, and J. Pustejovsky. Groningen, Netherlands (Online): Association for Computational Linguistics, pp. 32–44. URL: <https://aclanthology.org/2021.mmsr-1.4>.
- Parcalabescu, L., N. Trost, and A. Frank (2021). "What is Multimodality?" In: *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*. Ed. by L. Donatelli, N. Krishnaswamy, K. Lai, and J. Pustejovsky. Groningen, Netherlands (Online): Association for Computational Linguistics, pp. 1–10. URL: <https://aclanthology.org/2021.mmsr-1.1>.

- Parmar, N. et al. (2018). “Image Transformer”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4055–4064. URL: <http://proceedings.mlr.press/v80/parmar18a.html>.
- Pascanu, R., T. Mikolov, and Y. Bengio (2013). “On the difficulty of training recurrent neural networks”. In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1310–1318. URL: <http://proceedings.mlr.press/v28/pascanu13.html>.
- Patel, R. and E. Pavlick (2022). “Mapping Language Models to Grounded Conceptual Spaces”. In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. URL: <https://openreview.net/forum?id=gJcEM8sxHK>.
- Paulus, R., C. Xiong, and R. Socher (2017). *A Deep Reinforced Model for Abstractive Summarization*.
- Pavlick, E. and T. Kwiatkowski (2019). “Inherent Disagreements in Human Textual Inferences”. In: *Transactions of the Association for Computational Linguistics* 7, pp. 677–694. URL: https://doi.org/10.1162/tac1%5C_a%5C_00293.
- Paz-Argaman, T., R. Tsarfaty, G. Chechik, and Y. Atzmon (2020). “ZEST: Zero-shot Learning from Text Descriptions using Textual Similarity and Visual Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 569–579. URL: <https://aclanthology.org/2020.findings-emnlp.50>.
- Pennington, J., R. Socher, and C. Manning (2014). “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical*

- Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>.
- Perniss, P. and G. Vigliocco (2014). “The bridge of iconicity: from a world of experience to the experience of language”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1651, p. 20130300. URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2013.0300>.
- Peters, M. E. et al. (2018). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by M. Walker, H. Ji, and A. Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. URL: <https://aclanthology.org/N18-1202>.
- Pezzelle, S. (2023). “Dealing with Semantic Underspecification in Multimodal NLP”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 12098–12112. URL: <https://aclanthology.org/2023.acl-long.675>.
- Plank, B. (2022). “The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10671–10682. URL: <https://aclanthology.org/2022.emnlp-main.731>.
- Plummer, B. A. et al. (2015). “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile*,

- December 7-13, 2015. IEEE Computer Society, pp. 2641–2649. URL: <https://doi.org/10.1109/ICCV.2015.303>.
- Poesio, M. (2004). “Discourse Annotation and Semantic Annotation in the GNOME corpus”. In: *Proceedings of the Workshop on Discourse Annotation*. Barcelona, Spain: Association for Computational Linguistics, pp. 72–79. URL: <https://aclanthology.org/W04-0210>.
- Poesio, M., R. Stevenson, B. Di Eugenio, and J. Hitzeman (2004). “Centering: A Parametric Theory and Its Instantiations”. In: *Computational Linguistics* 30.3, pp. 309–363. URL: <https://aclanthology.org/J04-3003>.
- Radford, A., J. W. Kim, et al. (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8748–8763. URL: <http://proceedings.mlr.press/v139/radford21a.html>.
- Radford, A., J. Wu, et al. (2019). *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI. URL: https://d4mucfpksyw.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Raffel, C. et al. (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *J. Mach. Learn. Res.* 21, 140:1–140:67. URL: <http://jmlr.org/papers/v21/20-074.html>.
- Raganato, A. and J. Tiedemann (2018). “An Analysis of Encoder Representations in Transformer-Based Machine Translation”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Lin-

- guistics, pp. 287–297. URL: <https://www.aclweb.org/anthology/W18-5431>.
- Raghu, M. et al. (2021). *Do Vision Transformers See Like Convolutional Neural Networks?*
- Raunak, V. et al. (2019). “On Leveraging the Visual Modality for Neural Machine Translation”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 147–151. URL: <https://aclanthology.org/W19-8620>.
- Ravishankar, V. et al. (2021). “Attention Can Reflect Syntactic Structure (If You Let It)”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3031–3045. URL: <https://www.aclweb.org/anthology/2021.eacl-main.264>.
- Reed, S., Z. Akata, H. Lee, and B. Schiele (2016). “Learning Deep Representations of Fine-Grained Visual Descriptions”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, pp. 49–58.
- Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Bradford book. MIT Press. URL: <https://books.google.se/books?id=ZS9s4X6PJ1oC>.
- Řehůřek, R. and P. Sojka (2010). “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, pp. 45–50.
- Reimers, N. and I. Gurevych (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

- Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410>.
- Reiter, E. and A. Belz (2009). “An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems”. In: *Computational Linguistics* 35.4, pp. 529–558. URL: <https://aclanthology.org/J09-4008>.
- Reiter, E. and R. Dale (1997). “Building applied natural language generation systems”. In: *Natural Language Engineering* 3.1, pp. 57–87.
- Reiter, E. and R. Dale (2000). *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.
- Ren, M., R. Kiros, and R. S. Zemel (2015). “Exploring models and data for image question answering”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’15. Montreal, Canada: MIT Press, pp. 2953–2961.
- Ren, S., K. He, R. Girshick, and J. Sun (2015). “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf>.
- Rennie, S. J. et al. (2017). “Self-Critical Sequence Training for Image Captioning”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1195.
- Rethmeier, N., V. Kumar Saxena, and I. Augenstein (2020). “TX-Ray: Quantifying and Explaining Model-Knowledge Transfer in (Un-)Supervised NLP”. In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by J. Peters and D. Sontag. Vol. 124. Proceedings of

- Machine Learning Research. PMLR, pp. 440–449. URL: <http://proceedings.mlr.press/v124/rethmeier20a.html>.
- Rogers, A., O. Kovaleva, and A. Rumshisky (2020). “A Primer in BERTology: What We Know About How BERT Works”. In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866. URL: <https://www.aclweb.org/anthology/2020.tacl-1.54>.
- Rohrbach, A. et al. (2018). “Object Hallucination in Image Captioning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4035–4045.
- Rosch, E. (1975a). “Cognitive Reference Points”. In: *Cognitive Psychology* 7.4, pp. 532–547.
- Rosch, E. (1975b). “Cognitive Representations of Semantic Categories”. In: *Journal of Experimental Psychology: General* 104, pp. 192–233.
- Rosch, E. (1978). “Principles of Categorization”. In: *Cognition and Categorization*. Ed. by E. Rosch and B. B. Lloyd. Hillsdale, NJ: Erlbaum, pp. 27–48.
- Rosch, E. et al. (1976). “Basic objects in natural categories”. In: *Cognitive Psychology* 8.3, pp. 382–439. URL: <https://www.sciencedirect.com/science/article/pii/001002857690013X>.
- Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). “Learning representations by back-propagating errors”. In: *nature* 323.6088, pp. 533–536.
- Russakovsky, O. et al. (2015). “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision (IJCV)* 115.3, pp. 211–252.

- Savva, M. et al. (2019). “Habitat: A Platform for Embodied AI Research”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347. URL: https://openaccess.thecvf.com/content_ICCV_2019/papers/Savva_Habitat_A_Platform_for_Embodied_AI_Research_ICCV_2019_paper.pdf.
- Schlangen, D. (2021). “Targeting the Benchmark: On Methodology in Current Natural Language Processing Research”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, pp. 670–674. URL: <https://aclanthology.org/2021.acl-short.85>.
- Schlangen, D. (2022). “Norm Participation Grounds Language”. In: *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 62–69. URL: <https://aclanthology.org/2022.clasp-1.7>.
- Schlangen, D., S. Zarriß, and C. Kennington (2016). “Resolving References to Objects in Photographs using the Words-As-Classifiers Model”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1213–1223. URL: <https://aclanthology.org/P16-1115>.
- Schönfeld, E. et al. (2019). *Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders*.
- Schumann, R. and S. Riezler (2022). “Analyzing Generalization of Vision and Language Navigation to Unseen Outdoor Areas”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational

- Linguistics, pp. 7519–7532. URL: <https://aclanthology.org/2022.acl-long.518>.
- Schuster, M. and K. Paliwal (1997). “Bidirectional recurrent neural networks”. In: *IEEE Transactions on Signal Processing* 45.11, pp. 2673–2681.
- Schüz, S., A. Gatt, and S. Zarrieß (2023). “Rethinking symbolic and visual context in Referring Expression Generation”. In: *Frontiers in Artificial Intelligence* 6. URL: <https://www.frontiersin.org/articles/10.3389/frai.2023.1067125>.
- Schüz, S., T. Han, and S. Zarrieß (2021). “Diversity as a By-Product: Goal-oriented Language Generation Leads to Linguistic Variation”. In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by H. Li et al. Singapore and Online: Association for Computational Linguistics, pp. 411–422. URL: <https://aclanthology.org/2021.sigdial-1.43>.
- Schüz, S. and S. Zarrieß (2020). “Knowledge Supports Visual Language Grounding: A Case Study on Colour Terms”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics, pp. 6536–6542. URL: <https://aclanthology.org/2020.acl-main.584>.
- Scialom, T. et al. (2020). “What BERT Sees: Cross-Modal Transfer for Visual Question Generation”. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Ed. by B. Davis, Y. Graham, J. Kelleher, and Y. Sripada. Dublin, Ireland: Association for Computational Linguistics, pp. 327–337. URL: <https://aclanthology.org/2020.inlg-1.39>.
- Sellam, T., D. Das, and A. Parikh (2020). “BLEURT: Learning Robust Metrics for Text Generation”. In: *Proceedings of the 58th Annual Meeting of the*

- Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics, pp. 7881–7892. URL: <https://aclanthology.org/2020.acl-main.704>.
- Selvaraju, R. R. et al. (2017). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 618–626. URL: <https://doi.org/10.1109/ICCV.2017.74>.
- Serrano, S. and N. A. Smith (2019). “Is Attention Interpretable?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 2931–2951. URL: <https://aclanthology.org/P19-1282>.
- Shannon, C. E. (1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3, pp. 379–423. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x>.
- Sharma, P., N. Ding, S. Goodman, and R. Soricut (2018). “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 2556–2565. URL: <https://aclanthology.org/P18-1238>.
- Shekhar, R. et al. (2017). “FOIL it! Find One mismatch between Image and Language caption”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by R. Barzilay and M.-Y. Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 255–265. URL: <https://aclanthology.org/P17-1024>.

- Shetty, R. et al. (2017). “Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4155–4164.
- Silberer, C., V. Ferrari, and M. Lapata (2017). “Visually Grounded Meaning Representations”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.11, pp. 2284–2297.
- Silberer, C. and M. Lapata (2014). “Learning Grounded Meaning Representations with Autoencoders”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by K. Toutanova and H. Wu. Baltimore, Maryland: Association for Computational Linguistics, pp. 721–732. URL: <https://aclanthology.org/P14-1068>.
- Silberer, C., S. Zarrieß, and G. Boleda (2020). “Object Naming in Language and Vision: A Survey and a New Dataset”. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 5792–5801. URL: <https://www.aclweb.org/anthology/2020.lrec-1.710>.
- Silberer, C., S. Zarrieß, M. Westera, and G. Boleda (2020). “Humans Meet Models on Object Naming: A New Dataset and Analysis”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by D. Scott, N. Bel, and C. Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1893–1905. URL: <https://aclanthology.org/2020.coling-main.172>.
- Simonyan, K. and A. Zisserman (2015). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. URL: <http://arxiv.org/abs/1409.1556>.

- Skantze, G. and B. Willemsen (2022). “CoLLIE: Continual Learning of Language Grounding from Language-Image Embeddings”. In: *J. Artif. Int. Res.* 74. URL: <https://doi.org/10.1613/jair.1.13689>.
- Socher, R., M. Ganjoo, C. D. Manning, and A. Ng (2013). “Zero-Shot Learning Through Cross-Modal Transfer”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/2d6cc4b2d139a53512fb8cbb3086ae2e-Paper.pdf.
- Spearman, C. (1904). “The Proof and Measurement of Association Between Two Things”. In: *The American Journal of Psychology* 15.1, pp. 72–101.
- Srivastava, N. and R. R. Salakhutdinov (2012). “Multimodal Learning with Deep Boltzmann Machines”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.
- Stalnaker, R. (1978). “Assertion”. In: *Syntax and Semantics (New York Academic Press)* 9, pp. 315–332.
- Stalnaker, R. (2002). “Common Ground”. In: *Linguistics and Philosophy* 25.5-6, pp. 701–721.
- “Stochastic Estimation of the Maximum of a Regression Function” (1952). In: *Annals of Mathematical Statistics* 23.3, pp. 462–466.
- Su, W. et al. (2020). “VL-BERT: Pre-training of Generic Visual-Linguistic Representations”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=SygXPaEYvH>.

- Suglia, A. et al. (2020). “Imagining Grounded Conceptual Representations from Perceptual Information in Situated Guessing Games”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1090–1102. URL: <https://aclanthology.org/2020.coling-main.95>.
- Summerfield, Q. (1992). “Lipreading and Audio-Visual Speech Perception”. In: *Philosophical Transactions: Biological Sciences* 335.1273, pp. 71–78. URL: <http://www.jstor.org/stable/55477> (visited on 04/27/2024).
- Sutskever, I., O. Vinyals, and Q. V. Le (2014). “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.
- Takmaz, E. et al. (2020). “Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4350–4368. URL: <https://aclanthology.org/2020.emnlp-main.353>.
- Talmy, L. (1983). “How language structures space”. In: *Spatial orientation: theory, research, and application*. Ed. by H. L. Pick Jr. and L. P. Acredolo. Based on the proceedings of a Conference on Spatial Orientation and Perception held on July 14–16, 1980, at the University of Minnesota, Minneapolis, Minnesota. New York: Plenum Press, pp. 225–282.
- Talmy, L. (2000). *Toward a cognitive semantics: concept structuring systems*. Vol. 1 and 2. Cambridge, Massachusetts: MIT Press.

- Tan, H. and M. Bansal (2019). “LXMERT: Learning Cross-Modality Encoder Representations from Transformers”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics, pp. 5100–5111. URL: <https://aclanthology.org/D19-1514>.
- Tang, G., R. Sennrich, and J. Nivre (2018). “An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation”. In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Ed. by O. Bojar et al. Brussels, Belgium: Association for Computational Linguistics, pp. 26–35. URL: <https://aclanthology.org/W18-6304>.
- Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman (2011). “How to Grow a Mind: Statistics, Structure, and Abstraction”. In: *Science* 331.6022, pp. 1279–1285. URL: <https://www.science.org/doi/abs/10.1126/science.1192788>.
- Tenney, I., D. Das, and E. Pavlick (2019). “BERT Rediscovered the Classical NLP Pipeline”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4593–4601. URL: <https://www.aclweb.org/anthology/P19-1452>.
- Tenney, I., P. Xia, et al. (2019). “What do you learn from context? Probing for sentence structure in contextualized word representations”. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. URL: <https://openreview.net/forum?id=SJzSgnRcKX>.

- Thomason, J., D. Gordon, and Y. Bisk (2019). "Shifting the Baseline: Single Modality Performance on Visual Navigation & QA". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1977–1983. URL: <https://aclanthology.org/N19-1197>.
- Ullman, S. (1984). "Visual Routines". In: *Cognition* 18.1-3, pp. 97–159.
- Van der Lee, C. et al. (2019). "Best practices for the human evaluation of automatically generated text". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 355–368. URL: <https://www.aclweb.org/anthology/W19-8643>.
- Van Deemter, K., A. Gatt, R. P. van Gompel, and E. Krahmer (2012). "Toward a Computational Psycholinguistics of Reference Production". In: *Topics in Cognitive Science* 4.2, pp. 166–183. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2012.01187.x>.
- Van Deemter, K., I. van der Sluis, and A. Gatt (2006). "Building a Semantically Transparent Corpus for the Generation of Referring Expressions." In: *Proceedings of the Fourth International Natural Language Generation Conference*. Sydney, Australia: Association for Computational Linguistics, pp. 130–132. URL: <https://aclanthology.org/W06-1420>.
- Van Miltenburg, E. (2017). "Pragmatic descriptions of perceptual stimuli". In: *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, pp. 1–10. URL: <https://aclanthology.org/E17-4001>.

- Van Miltenburg, E., D. Elliott, and P. Vossen (2017). “Cross-linguistic differences and similarities in image descriptions”. In: *Proceedings of the 10th International Conference on Natural Language Generation*. Santiago de Compostela, Spain: Association for Computational Linguistics, pp. 21–30. URL: <https://www.aclweb.org/anthology/W17-3503>.
- Van Miltenburg, E., D. Elliott, and P. Vossen (2018). “Measuring the Diversity of Automatic Image Descriptions”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by E. M. Bender, L. Derczynski, and P. Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1730–1741. URL: <https://aclanthology.org/C18-1147>.
- Vaswani, A. et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon et al., pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Vedantam, R., S. Bengio, et al. (2017). “Context-Aware Captions From Context-Agnostic Supervision”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 251–260.
- Vedantam, R., C. Lawrence Zitnick, and D. Parikh (2015). “CIDEr: Consensus-Based Image Description Evaluation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf.
- Vig, J. (2019). “A Multiscale Visualization of Attention in the Transformer Model”. In: *Proceedings of the 57th Annual Meeting of the Association for*

- Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 37–42. URL: <https://www.aclweb.org/anthology/P19-3007>.
- Vig, J. and Y. Belinkov (2019). “Analyzing the Structure of Attention in a Transformer Language Model”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 63–76. URL: <https://www.aclweb.org/anthology/W19-4808>.
- Vijayakumar, A. et al. (2018). “Diverse Beam Search for Improved Description of Complex Scenes”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/12340>.
- Vijayakumar, A. K. et al. (2016). “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models”. In: *ArXiv abs/1610.02424*.
- Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2015). “Show and tell: A neural image caption generator”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 3156–3164. URL: <https://doi.org/10.1109/CVPR.2015.7298935>.
- Voita, E., P. Serdyukov, R. Sennrich, and I. Titov (2018). “Context-Aware Neural Machine Translation Learns Anaphora Resolution”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 1264–1274. URL: <https://aclanthology.org/P18-1117>.
- Voita, E., D. Talbot, et al. (2019). “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned”. In: *Pro-*

- ceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5797–5808. URL: <https://www.aclweb.org/anthology/P19-1580>.
- Wah, C. et al. (2011). *Caltech-UCSD Birds 200*. Tech. rep. CNS-TR-2011-001. California Institute of Technology.
- Wallace, E. et al. (2019). “AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, pp. 7–12. URL: <https://aclanthology.org/D19-3002>.
- Wang, J., Y. Pan, et al. (2019). *Convolutional Auto-encoding of Sentence Topics for Image Paragraph Generation*.
- Wang, J., J. Tuyls, E. Wallace, and S. Singh (2020). “Gradient-based Analysis of NLP Models is Manipulable”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 247–258. URL: <https://aclanthology.org/2020.findings-emnlp.24>.
- Wang, P., A. Yang, et al. (2022). “OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework”. In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by K. Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 23318–23340. URL: <https://proceedings.mlr.press/v162/wang22a1.html>.
- Wang, Q. and A. B. Chan (2019). *Describing like humans: on diversity in image captioning*.

- Wang, S., Z. Yao, et al. (2021). “FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14045–14054.
- Weidinger, L. et al. (2022). “Taxonomy of Risks posed by Language Models”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. <conf-loc>, <city>Seoul</city>, <country>Republic of Korea</country>, </conf-loc>: Association for Computing Machinery, pp. 214–229. URL: <https://doi.org/10.1145/3531146.3533088>.
- Wiegrefe, S. and Y. Pinter (2019). “Attention is not not Explanation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics, pp. 11–20. URL: <https://aclanthology.org/D19-1002>.
- Wiher, G., C. Meister, and R. Cotterell (2022). “On Decoding Strategies for Neural Text Generators”. In: *Transactions of the Association for Computational Linguistics* 10. Ed. by B. Roark and A. Nenkova, pp. 997–1012. URL: <https://aclanthology.org/2022.tacl-1.58>.
- Wijmans, E. et al. (2019). “Embodied Question Answering in Photorealistic Environments With Point Cloud Perception”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp. 6659–6668. URL: http://openaccess.thecvf.com/content%5C_CVPR%5C_2019/html/Wijmans%5C_Embodied%5C_Question%5C_Answering%5C_in%5C_Photorealistic%5C_Environments%5C_With%5C_Point%5C_Cloud%5C_Perception%5C_CVPR%5C_2019%5C_paper.html.

- Wu, Q., C. Shen, et al. (2016). “What Value Do Explicit High Level Concepts Have in Vision to Language Problems?” In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 203–212. URL: <https://doi.org/10.1109/CVPR.2016.29>.
- Wu, Y., Y. Wu, G. Gkioxari, and Y. Tian (2018). “Building Generalizable Agents with a Realistic and Rich 3D Environment”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=r1b06vyDG>.
- Wu, Y., M. Schuster, et al. (2016). “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR abs/1609.08144*. URL: <http://arxiv.org/abs/1609.08144>.
- Xian, Y., C. H. Lampert, B. Schiele, and Z. Akata (2020). *Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly*.
- Xu, G., P. Kordjamshidi, and J. Chai (2021). “Zero-Shot Compositional Concept Learning”. In: *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*. Online: Association for Computational Linguistics, pp. 19–27. URL: <https://aclanthology.org/2021.metanlp-1.3>.
- Xu, K., J. Ba, et al. (2015). “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 2048–2057. URL: <https://proceedings.mlr.press/v37/xuc15.html>.
- Yang, F. et al. (2019). “Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification”. In: *Proceedings of the Third Workshop on*

- Abusive Language Online*. Ed. by S. T. Roberts, J. Tetreault, V. Prabhakaran, and Z. Waseem. Florence, Italy: Association for Computational Linguistics, pp. 11–18. URL: <https://aclanthology.org/W19-3502>.
- Yosinski, J., J. Clune, Y. Bengio, and H. Lipson (2014). “How transferable are features in deep neural networks?” In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., pp. 3320–3328. URL: <https://proceedings.neurips.cc/paper/2014/file/375c71349b295f2dcda9206f20a06-Paper.pdf>.
- You, Q. et al. (2016). “Image Captioning with Semantic Attention”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 4651–4659. URL: <https://doi.org/10.1109/CVPR.2016.503>.
- Young, P., A. Lai, M. Hodosh, and J. Hockenmaier (2014). “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics 2*. Ed. by D. Lin, M. Collins, and L. Lee, pp. 67–78. URL: <https://aclanthology.org/Q14-1006>.
- Yu, L., X. Chen, et al. (2019). “Multi-Target Embodied Question Answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6309–6318. URL: https://openaccess.thecvf.com/content_CVPR_2019/papers/Yu_Multi-Target_Embodied_Question_Answering_CVPR_2019_paper.pdf.
- Yu, L., P. Poirson, et al. (2016). “Modeling Context in Referring Expressions”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Vol. 9906. Lecture Notes in Computer Science. Springer, pp. 69–85. URL: https://doi.org/10.1007/978-3-319-46475-6_5C_5.

- Yuhas, B. P., M. H. Goldstein, and T. J. Sejnowski (1989). "Integration of acoustic and visual speech signals using neural networks". In: *Comm. Mag.* 27.11, pp. 65–71. URL: <https://doi.org/10.1109/35.41402>.
- Zarriß, S. and D. Schlangen (2018). "Decoding Strategies for Neural Referring Expression Generation". In: *Proceedings of the 11th International Conference on Natural Language Generation*. Ed. by E. Krahmer, A. Gatt, and M. Goudbeek. Tilburg University, The Netherlands: Association for Computational Linguistics, pp. 503–512. URL: <https://aclanthology.org/W18-6563>.
- Zarriß, S., H. Voigt, and S. Schüz (2021). "Decoding Methods in Neural Language Generation: A Survey". In: *Information* 12.9. URL: <https://www.mdpi.com/2078-2489/12/9/355>.
- Zhang, C., B. Van Durme, Z. Li, and E. Stengel-Eskin (2022). "Visual Commonsense in Pretrained Unimodal and Multimodal Models". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 5321–5335. URL: <https://aclanthology.org/2022.naacl-main.390>.
- Zhang, H., D. Duckworth, D. Ippolito, and A. Neelakantan (2021). "Trading Off Diversity and Quality in Natural Language Generation". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Ed. by A. Belz et al. Online: Association for Computational Linguistics, pp. 25–33. URL: <https://aclanthology.org/2021.humeval-1.3>.
- Zhang, P., Y. Goyal, et al. (2016). "Yin and Yang: Balancing and Answering Binary Visual Questions". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 5014–5022. URL: <https://doi.org/10.1109/CVPR.2016.542>.

- Zhang, T., V. Kishore, et al. (2020). “BERTScore: Evaluating Text Generation with BERT”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- Zhang, Y., J. S. Hare, and A. Prügel-Bennett (2018). “Learning to Count Objects in Natural Images for Visual Question Answering”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=B12Js%5C_yRb.
- Zheng, C., Q. Guo, and P. Kordjamshidi (2020). “Cross-Modality Relevance for Reasoning on Language and Vision”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7642–7651. URL: <https://aclanthology.org/2020.acl-main.683>.
- Zhou, B., H. Zhao, et al. (2017). “Scene Parsing Through ADE20K Dataset”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. URL: <https://people.csail.mit.edu/bzhou/publication/scene-parse-camera-ready.pdf>.
- Zhou, D., B. Kang, et al. (2021). “DeepViT: Towards Deeper Vision Transformer”. In: *ArXiv abs/2103.11886*.
- Zhou, L., H. Palangi, et al. (2020). “Unified Vision-Language Pre-Training for Image Captioning and VQA”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 13041–13049. URL: <https://doi.org/10.1609/aaai.v34i07.7005>.

- Zhu, Y. et al. (2018). “Texygen: A Benchmarking Platform for Text Generation Models”. In: *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, pp. 1097–1100. URL: <https://doi.org/10.1145/3209978.3210080>.
- Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.