

**Department of Philosophy, Linguistics and
Theory of Science**

Computational models of language and vision

Studies of neural models as learners of multi-modal knowledge

Nikolai Ilinykh

Thesis submitted for the Degree of Doctor of Philosophy in Computational Linguistics, to be publicly defended, by due permission of the dean of the Faculty of Humanities at University of Gothenburg, on June 11, 2024, at 13:00, in J222 Hörsalen, Humanisten, Renströmsgatan 6, Gothenburg, Sweden.

opponent · Assistant Professor Carina Silberer, Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany.



UNIVERSITY OF
GOTHENBURG

<i>title</i>	· Computational models of language and vision Studies of neural models as learners of multi-modal knowledge
<i>author</i>	· Nikolai Ilinykh
<i>supervisors</i>	· Simon Dobnik, Asad Sayeed
<i>language</i>	· English
<i>department</i>	· Department of Philosophy, Linguistics, and Theory of Science
<i>ISBN</i>	· 978-91-8069-767-5 (print), 978-91-8069-768-2 (PDF)
<i>key words</i>	· language and vision, self-attention, multi-modal representation learning, evaluation of language models, computational linguistics, machine learning

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.