

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 uppmärksamhetskort (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, förkroppslikat 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 datogenererade 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. *Lines*, 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. ТПЛ!

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.

Mama и Пана, 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 Transformer 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 representations	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://sprakbanken.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 of 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**, Yasmeen Emamipoor, 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 Somashekharappa. 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://semodial.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 Elžbieta 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.pvlam-1.5/>
- (vii) A General Benchmarking Framework for Text Generation. Diego Mous-sallem, 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 on 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 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 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 model 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

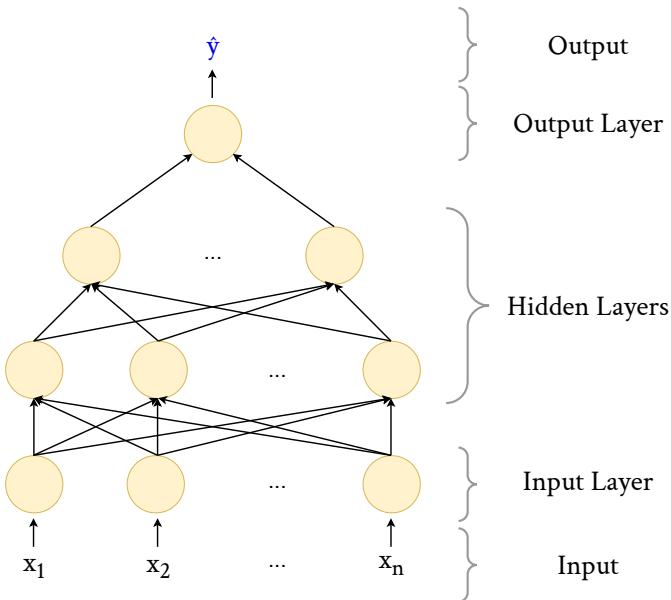


Figure 4.1. A multi-layer perceptron. This type of neural models is often used in classification tasks.

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}_o 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}_o . For example, if there are 3 neurons with 5 features each, then $\mathbf{W}_o \in \mathbb{R}^{3 \times 5}$. 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:

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

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

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

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

where f_3 produces the prediction \hat{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 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, Zarrieß, 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.

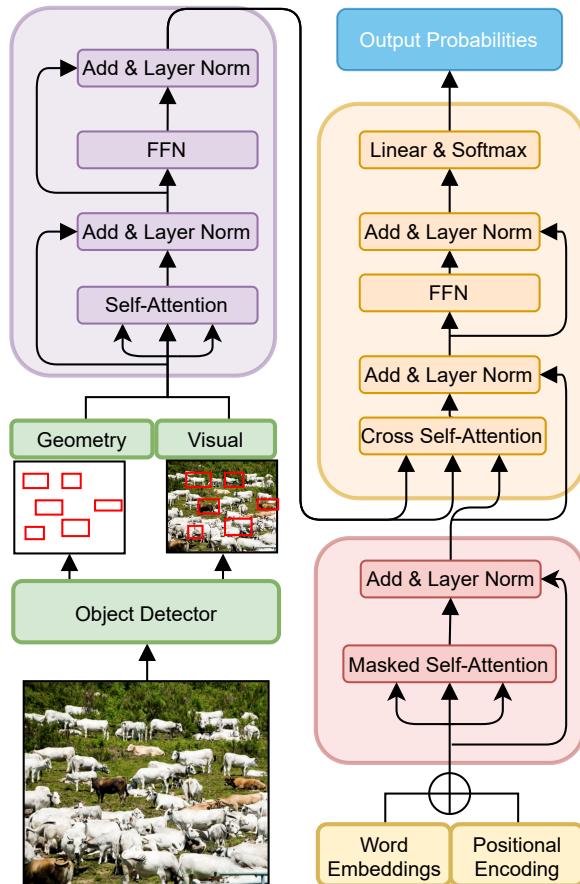


Figure 4.2. Object relation image captioning transformer. The model schema was initially described in Study I, here we duplicate it in order to explain the model's key components.

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} :

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

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

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

where $\boldsymbol{\Omega}^V$ 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 $\boldsymbol{\Omega}^V$. 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 a 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 evaluate 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 `NUM MOD` (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 `P OBJ` 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 Zarrieß, 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 unimodal (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 Zarrieß (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, Yasmeen 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 results 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 Emamipoor 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 Emamipoor. 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. Emamipoor 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, Zarrieß, 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 Fabrizio 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, Zarrieß, 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 generate 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-etal-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 (Ilinskykh, 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

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

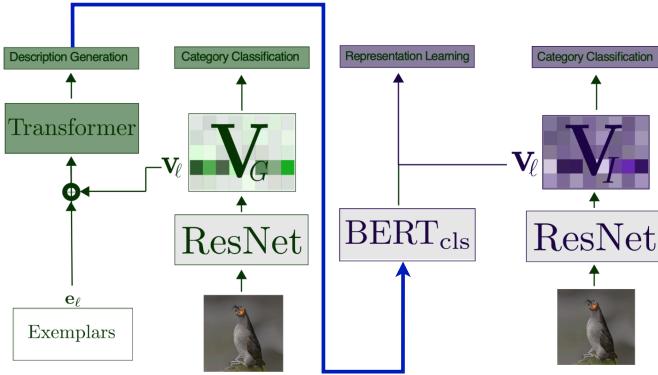


Figure 5.2. The generator model (green) and the interpreter model (purple). The blue arrow shows how descriptions produced by the generator are used by the interpreter.

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 6: Studies

6.1. How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer

6.1.1. Abstract

The problem of interpretation of knowledge learned by multi-head self-attention in transformers has been one of the central questions in NLP. However, a lot of work mainly focused on models trained for uni-modal tasks, e.g. machine translation. In this paper, we examine masked self-attention in a multi-modal transformer trained for the task of image captioning. In particular, we test whether the multi-modality of the task objective affects the learned attention patterns. Our visualisations of masked self-attention demonstrate that (i) it can learn general linguistic knowledge of the textual input, and (ii) its attention patterns incorporate artefacts from visual modality even though it has never accessed it directly. We compare our transformer’s attention patterns with masked attention in distilgpt-2 tested for uni-modal text generation of image captions. Based on the maps of extracted attention weights, we argue that masked self-attention in image captioning transformer seems to be enhanced with semantic knowledge from images, exemplifying joint language-and-vision information in its attention patterns.

6.1.2. Introduction

Recently, we have seen a surge of interest in explainability research for large-scale neural networks, e.g. transformers (Vaswani et al., 2017). A lot of the existing literature focuses on the analysis of attention (Bahdanau et al.,

2015) in terms of linguistic knowledge it encodes (Belinkov and Glass, 2019). Clark, Khandelwal, et al. (2019) show that attention heads' patterns in BERT (Devlin, Chang, et al., 2019) resemble syntactic dependencies present in the text. They also use a probing classifier to identify how knowledge of syntax is distributed between attention heads. Hoover et al. (2020) and Vig and Belinkov (2019) have shown that visualising the structure of attention in transformer models can help us see which parts of the model capture specific syntactic knowledge. Voita, Talbot, et al. (2019) demonstrate that not all attention heads are equally suitable for learning syntactic information. Thus, pruning such heads can be an option to reduce the model's complexity. While attention is not always an explanation (Jain and Wallace, 2019), some work (Ravishankar et al., 2021) has shown that extra fine-tuning on a syntax-related task can guide the model's attention to truly resemble syntactic information about the text. Other approaches to the model's interpretability include, for example, a work by Rethmeier et al. (2020), which inspects how knowledge is transferred on the neuron level rather than attention level.

While most of the existing research has placed the problem of model's explainability in the context of **uni-modal** text-based tasks, e.g. machine translation, the field of language-and-vision is somewhat lacking similar analysis for models trained to solve **multi-modal** tasks. This becomes especially important with the increasing interest in adopting transformers for learning better cross-modal representations (Tan and Bansal, 2019). In addition, using large-scale models to improve grounding between language and vision representations (Lu, Batra, et al., 2019) requires vigilance regarding how information is learned in different parts of such densely structured models. Multi-modal transformers are required to not only learn to perform *symbol grounding*, e.g. mapping natural language symbols into visual representations as defined by Harnad (1990) and a language model, but also learn *to fuse information* from two modalities, the nature of which has been an open question in the field (Caglayan, Madhyastha, et al., 2019; Ilinykh and Dobnik, 2020;

Lu, Xiong, et al., 2017). The effect that such multi-modal representations have on the attention in large-scale models has not been addressed a lot in the language-and-vision literature. More specifically, we need a better understanding of how self-attention in transformer processes the multi-modal information.

In this paper, we analyse the masked self-attention part of the image captioning transformer, which performs a standard language masking task based on the textual input, and compare its attention patterns with masked attention in distilgpt-2, a text-only transformer. Our goal is to identify what kind of knowledge is captured in representations learned by this part of the model and whether it is affected in any way by the visual modality, which is not directly accessible for this particular self-attention. We aim to answer the following questions:

- Does masked self-attention show patterns which resemble any syntactic knowledge of the input text?
- What are the differences in attention on previous words when generating the next word in either the uni-modal or multi-modal task set-up?
- What is the task's effect (uni-modal vs. multi-modal) on the semantics of words captured by masked-self attention in image captioning transformer?

In addressing these questions, we believe that we show novel insights into how the information is transferred between inner self-attentions of complex architectures such as a transformer and how representations from specific components of such models are affected by the training objective and multi-modality.

6.1.3. Model

Fig. 6.1 shows the architecture of the image captioning transformer that we use for our experiments, first introduced by Herdade et al. (2019) and built on top of the basic image captioning transformer (Luo, Price, et al., 2018).

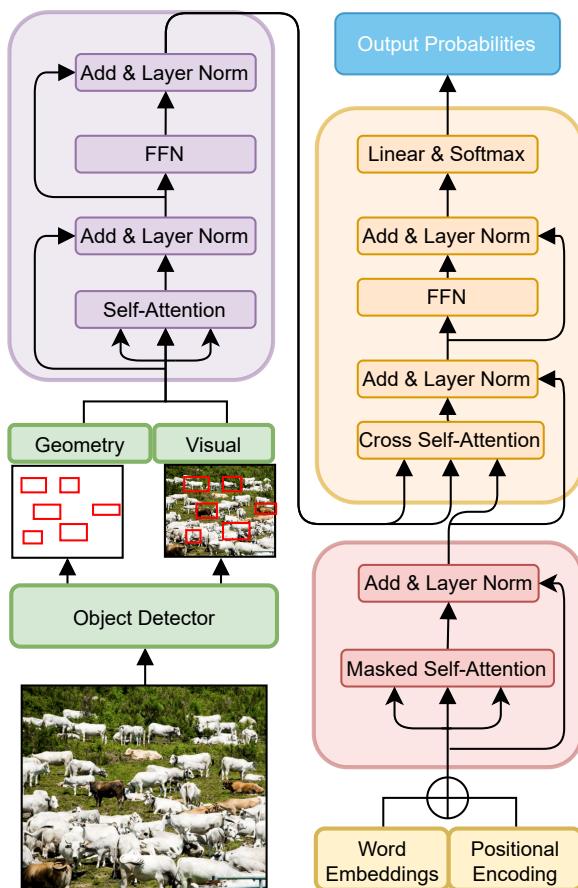


Figure 6.1. Object relation image captioning transformer. The image is first passed through a pre-trained object detector to extract visual and geometric features. The left side **self-attention (image encoder)** consists of attention heads, where each of them utilises both visual and geometry information. On the right side, the **masked self-attention (text encoder)** is given the embeddings of the caption words and their positional information. The words are fed to the text encoder in an auto-regressive manner, e.g. one word at a time plus all the preceding words. The **cross self-attention** uses keys K and values V from the visual encoder, while queries Q are coming from the textual encoder and finally predicts the output probabilities of the next word.

This architecture resembles many parts of the classic transformer (Vaswani et al., 2017), which was initially introduced for machine translation, consisting of three multi-head self-attention mechanisms. The standard transformer’s encoder learns representations of the input text by passing it through two sub-layers: multi-head self-attention and feed-forward network. Each sub-layer has a residual connection around itself, followed by layer-normalisation operation. The decoder contains masked self-attention, which is used to learn linguistic knowledge of the ground-truth target translation. In a unidirectional task, it masks the words in the future so that the model learns to attend to the previously generated words only. The third self-attention is performing a cross-modelling task, using information from both encoder and decoder. This cross self-attention identifies correlations between the source text and currently generated target text in a machine translation context.

Once we reformulate the model’s task from machine translation to image captioning (Fig. 6.1), we naturally change the encoder’s inputs. Instead of the source sentence, the encoder uses representations of objects from the image as its input. On the decoder’s side, the ground-truth captions that the model learns to generate are used as inputs during training. To prepare inputs to our encoder, we first extract visual features of the detected objects $X = \{x_1, \dots, x_N\}$, where $x_n \in \mathbb{R}^{1 \times D}$ with $N = 36$ and $D = 2048$. We use a bottom-up feature extractor (Anderson, He, et al., 2018), which is based on Faster-RCNN (Ren, He, et al., 2015) and pre-trained on Visual Genome (Krishna et al., 2017) with the ResNet-101 as its backbone (He et al., 2016). For each detected object we also extract geometry features $G = (x, y, w, h)$ (centre coordinates, width, height). In the next step, queries $Q = W^Q X$ and keys $K = W^K X$ are used to get scaled dot product Ω^V :

$$\Omega^V = \frac{QK^T}{\sqrt{d_k}} \quad (6.1)$$

Then, Ω^V and geometric features G are combined, taking into account

the displacement between the objects and producing a fused representation Ω .¹ Finally, each attention head h from each encoder layer l outputs a combination of values V and geometry-aware visual features Ω :

$$\text{head}_{l,h} = \text{self-attention}(Q, K, V) = \Omega V \quad (6.2)$$

Masked self-attention in the decoder The idea of self-attention is that each token from the input text learns to attend to the other tokens from the same sequence. However, this is not feasible for the caption generation task since attending to the future tokens is unfair and it cannot be used when generating text. Therefore, the self-attention in the decoder is using masking of future tokens to keep the auto-regressive nature of the model. In particular, the token w_t and the future tokens w_{t+1}, \dots, w_W are replaced with [MASK]. Then, w_t is predicted using the previous context in the standard left-to-right fashion: $W_{\setminus t} := (w_1, \dots, w_{t-1})$.

We have specifically focused on the analysis of the attention weights in the **decoder's masked self-attention** of the image captioning transformer. We extract the attention weights for each head h in each layer l of this self-attention and use them for our visualisations and analysis. These weights are calculated similarly to the attended visual features (Eq. 6.1). Our masked self-attention has six layers, consisting of eight heads in each of them.

For the model checkpoint, we use the best model released by the authors of the architecture². This checkpoint has been chosen on the basis of automatic evaluation scores: the model uses bottom-up representation of images, geometry features and self-critical training (Rennie et al., 2017). The captions are generated using beam search with beam width $bw = 5$ in the standard auto-regressive manner.

¹For more details on how geometric information is combined with visual features in this model, we refer the reader to Herdade et al. (2019).

²Available at: https://github.com/yahoo/object_relation_transformer

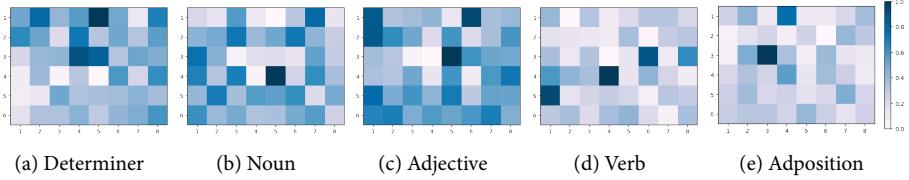


Figure 6.2. Each heat-map demonstrates the proportion of attention targeted towards a word of a specific part of speech. Vertical and horizontal axes indicate layers and heads respectively.

6.1.4. Learning syntactic knowledge

In our first experiment we investigate whether the attention weights of the masked self-attention are able to capture any general syntactic knowledge about the input text. It has been shown that the multi-head attention patterns in the transformer trained for the task of machine translation resembles syntactic properties of language at the level of part-of-speech tags and syntactic dependencies (Mareček and Rosa, 2019; Ravishankar et al., 2021). Since the self-attention that we are focused on is trained in a very similar task (masked language modelling), we first explore if particular layers and heads attend to specific part-of-speech tags the most. Then, we continue with the analysis of how information about syntactic dependencies is reflected in the learned attention patterns.

Attention on Part-of-Speech We follow Vig and Belinkov (2019) who compute the proportion of attention from each head that this head pays to tokens of a particular part-of-speech tag and accumulate the results over our test set:

$$P(\alpha | tag) = \frac{\sum_{s \in S} \sum_{i=1}^{|s|} \sum_{j=1}^i \alpha(s_i, s_j, pos(j)=tag)}{\sum_{s \in S} \sum_{i=1}^{|s|} \sum_{j=1}^i \alpha(s_i, s_j)} \quad (6.3)$$

where S is the corpus of generated captions, tag is the part-of-speech tag of the attended word, and $\alpha(s_i, s_j)$ is the attention from i^{th} word to j^{th} word for the given head. We use Spacy (Honnibal et al., 2020) to get part-of-speech tags of words and syntactic dependencies between them for all our

experiments. We also perform normalisation (linear scaling) on the values of the calculated attention proportion to place all values in a single scale from 0 to 1. The masked self-attention is always given the START token at the start of the generation. We consider attention on this token non-informative (as it is over-attended) and ignore the corresponding attention weights for better visualisations. The heads pay only ~26% of their attention to the START token on average per caption. We use BertViz tool (Vig, 2019) to produce our visualisations.

The results for the five most frequently occurring part-of-speech tags (more than 1000 individual instances) are shown in Fig. 6.2. Words of such part-of-speech tags, which can be grounded in visual signals (nouns for objects, adjectives for attributes), receive attention from a large number of attention heads. On the other hand, only specific heads focus on words describing relations (verbs, adpositions). Specifically, seventeen heads (out of forty-eight) put more than 40% of their attention to the nouns, while only three heads give more than 30% of their attention to the verbs.

We also find supporting evidence for the previous studies (Belinkov, 2018; Vig and Belinkov, 2019), showing that deeper layers focus on more complex properties, e.g. relational part-of-speech tags (verbs), which require knowledge of objects learned from earlier layers (nouns). For example, the top 3 attention heads that attend to basic parts-of-speech such as determiners are all located in the model’s first three layers. For adjectives, the top 3 heads are similarly located in the first three layers of the model, with the maximum value of the attention head being 0.25. However, attention on adjectives is more spread across many heads in different layers, with the attention value being 0.14 for more than half of the heads, which is also a mean value for attention on this part-of-speech tag. A less clustered pattern is observed for nouns: its top 3 heads are located in layers 1, 3, and 4, with thirty-three heads paying more than 30% of their attention to nouns. We argue that the reason why the attention on nouns is scattered over many heads, with most of them

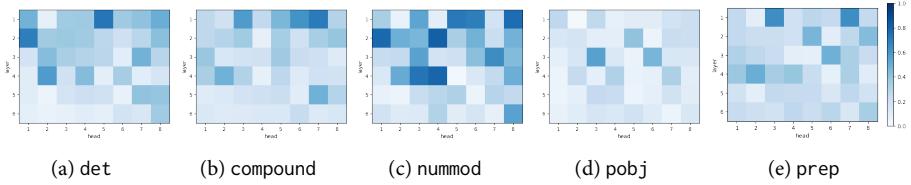


Figure 6.3. Attention distribution on different constituents of the specific syntactic dependencies. For *det*, *compound*, *nummod* we visualise which heads look the most on the non-root element of the dependency (e.g. “man” → “a” in “a man”). For *pobj* and *prep* we show attention in a different direction (e.g. “table” → “on” in “on table”, “with” → “bathroom” in “bathroom with”).

paying nearly one-third of their attention to the nouns, is because nouns are continuously required for caption generation: the model needs to take them into account when generating either a relation or an attribute.

Somewhat differently, verbs are attended mostly in the model’s deeper layers: the top 3 most attentive heads are located in layers 3, 4, and 5 with values higher than 0.3. The vast majority of the heads (forty-three) have smaller attention values (less than 0.2), indicating that the model needs verbs only for specific situations, for example when a relationship needs to be generated. Overall, our visualisations demonstrate that masked self-attention weights resemble task-specific syntactic information about part-of-speech tags. For example, nouns are similarly attended across all heads since they are required for the captioning task the most (to describe, refer to, use in phrases, etc.). In contrast, more function-dependent parts of speech (verbs, adpositions) are attended to by fewer heads in the deeper layers of the model.

Attention on Syntactic Dependencies Fig. 6.3 shows the proportion of attention from the heads in masked self-attention for the most frequently occurring syntactic dependency relations. The proportions are calculated similarly to Eq. 6.3. In particular, we used the attention weights from root to the non-root part of the dependency phrase or vice versa, extracting dependencies in advance. This choice was affected by the auto-regressive nature

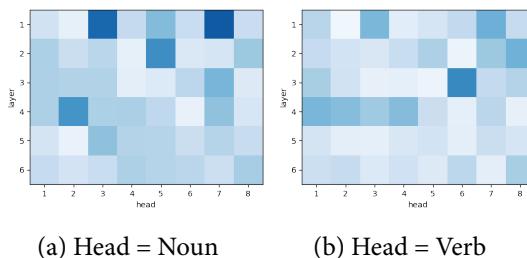


Figure 6.4. Attention distribution for the prep syntactic dependency. The left-side heat-map is computed for phrases where noun is a head in the phrase (“kitchen with”, while for the right-side heat-map it is the verb (“sitting at”).

of the generation task: for each word, we could only inspect attention focus on previous words. The attention on different dependencies seems to be distributed similarly to the attention on part-of-speech tags. More specifically, attention heads from the surface layers (1:5 and 2:1³) seem to be focused on the determiner in the det relation. Comparing heat-maps of attention distribution on part-of-speech and syntactic dependency may give us intuition about the specific heads' role. For example, the heads 3:4 or 3:5 are not intensely active for the det relation, although they are among the most active heads when attending to the determiners. This indicates that these heads 1:5 and 2:1 may be more responsible for focusing on determiners when the phrase in the det relation is generated. Interestingly, many heads strongly attend to the numeral in the nummod dependency compared to all other relations. This could be related to the importance of learning about the number of objects in the scene, while other, simpler noun-based dependencies (det, compound) do not have to be attended so strongly.

Only a few heads specialise in dependencies that capture more complex properties (e.g. relations between different objects), with heads 3:3 and 3:6 being the most attending heads for pobj. The root of the prep phrase is often attended in the first layer, with only a few more heads in the later layers being activated. *Could this pattern be mapped with the fact that roots in these*

³We use layer:head notation.

phrases are often nouns and verbs? Fig. 6.4 shows that heads 1:3 and 1:7 are the most active heads when a noun is a root in the phrase of prep dependency. Same heads in the first layer are also active the most when looking at the nouns, according to Fig. 6.2b. This indicates that the model acquires basic knowledge of language syntax (dependencies, part-of-speech information) in its first layers. Similarly, as Fig. 6.4b demonstrates, the head 3:6 is the single most active head for the prep dependency. At the same time, according to Fig. 6.2d, this particular head is one of the few most active heads when the attention focus is on verbs. This might be interpreted as if this head is better at learning information about syntactic dependencies than other activated heads. We argue that it is helpful to look at the correspondence between attention on parts-of-speech and syntactic dependency since it is informative when determining specific heads' roles and how important they are for different language tasks, e.g., part-of-speech tagging and syntactic dependency identification.

6.1.5. Multi-modality and masked self-attention

In this section, we look at how a multi-modal task of image captioning affects attention on the previous words when a masked self-attention model predicts the next word. We also compare our model's attention patterns with patterns from an auto-regressive model, distilgpt-2 (Radford, Wu, et al., 2019), which has been pre-trained on OpenWebTextCorpus. This model has 6 layers with 12 heads in each layer, which makes it more comparable to our captioning transformer than the standard GPT-2 model with 12 heads in each of the 12 layers.

Semantics of Attention Patterns Here, we compare the text-only uni-modal language model and its attention patterns with our multi-modal transformer's masked self-attention. We do this because we want to investigate to what

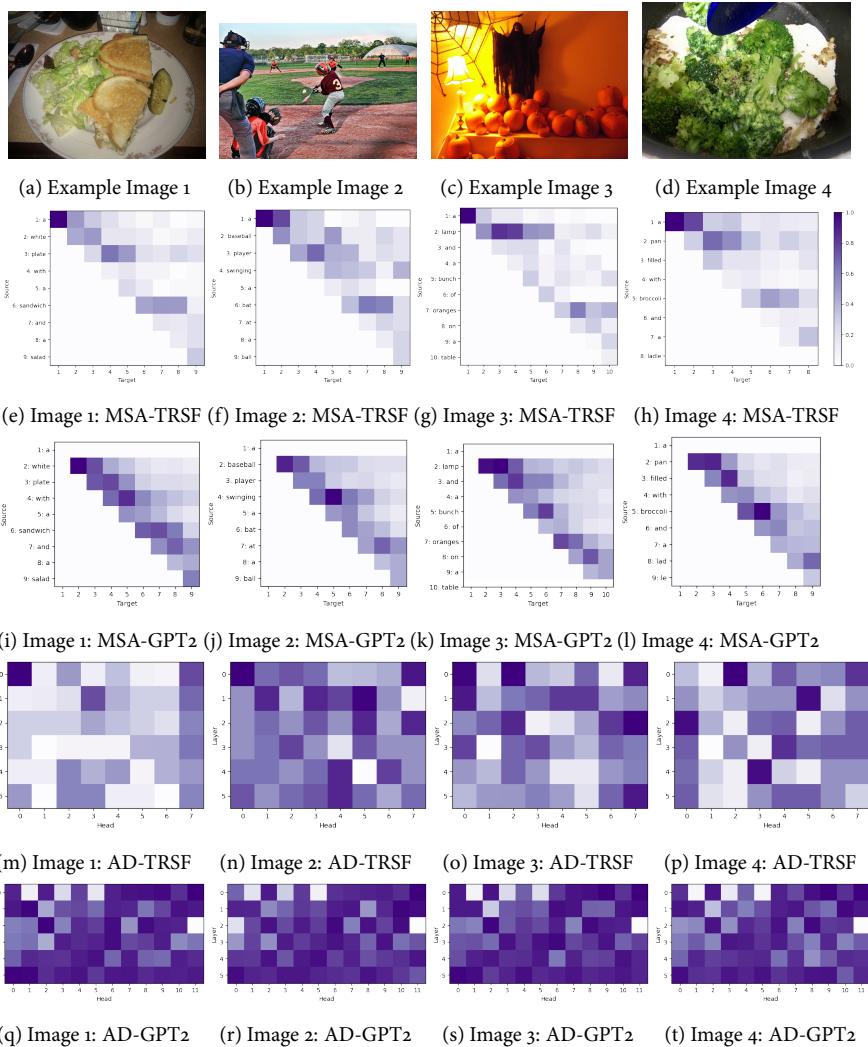


Figure 6.5. Here are several examples of different attention visualisations for masked-self attention (**MSA**) from our image captioning transformer (**TRSF**) and distlgpt-2 (**GPT2**). **The top row** shows example images for which we generate a caption. **The second and third rows** show attention on the available context (indicated by the *Source* axis) when generating the next word (the *Target* axis). Word of the generated caption are displayed on the *Source* axis. To get more fine-grained visualisations in the third row, we exclude attention on the first token of each sentence for distlgpt-2 attention patterns since, based on our experiments and literature (Vig and Belinkov, 2019), attention on the first token is always very strong and not relevant. **The fourth and the fifth rows** show attention dispersion (**AD**) for each head in each layer. The colour bar in the second row indicates the range of values in all visualisations in this figure.

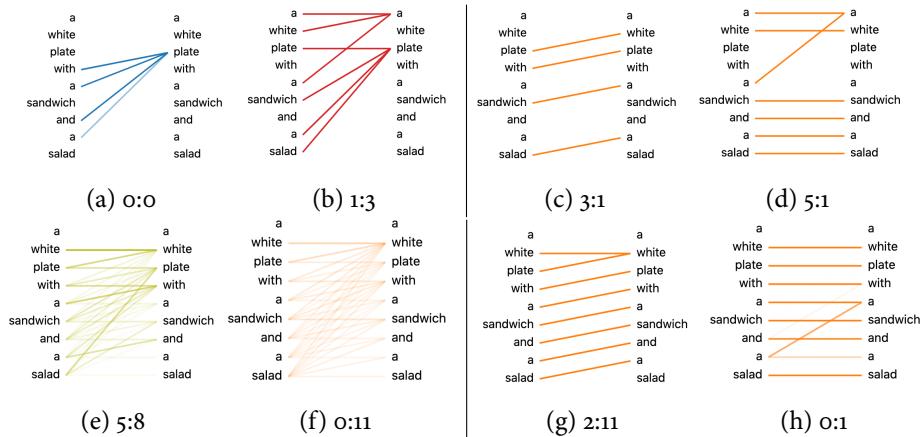


Figure 6.6. Visualisation of attention for example attention heads. The first row shows heads from the masked self-attention in our transformer; the second row depicts the head's attention from distilgpt-2. The side to the left of the vertical line in the middle includes heads with **high entropy** in either of the models, while the right side contains heads with **low entropy**. The heads are denoted by a layer:head notation; they can be traced back to the more general attention concentration in Fig. 6.5m and Fig. 6.5q. Each figure displays attention from **target (left)** to **source (right)**.

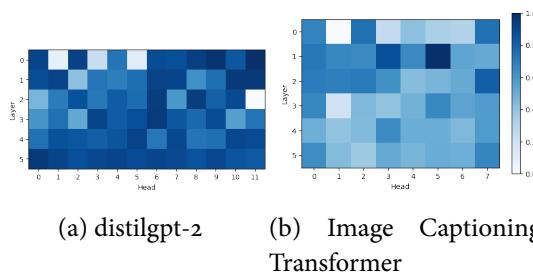


Figure 6.7. Mean normalised entropy of attention per head / layer calculated for the set of generated captions.

extent the attention patterns produced by the language model in the multi-modal setting differ from patterns where the task is uni-modal. For this, we run distilgpt-2 (Radford, Wu, et al., 2019) on the captions generated by our image captioning transformer, where both input and target are the same image descriptions. This way, we receive two sets of masked self-attention weights for the same texts from two models trained for different tasks. Both our decoder and the distilgpt-2 model are trained for the masked language modelling task; therefore, these models' attention is comparable with each other. We save the model's attention weights similar to how we did it for our captioning transformer's masked self-attention. The first three rows in Fig. 6.5 show visualisations of both models' attention for several captions and if applicable the corresponding images. Attention in our masked self-attention tends to focus on nouns much more than on other parts of the source (context). In comparison, distilgpt-2 patterns are more diagonal: every next word is focused on its surroundings the most, and the attention does not generally look at a single word for too long.

We believe that this is an artefact of the training for image captioning task: our masked self-attention learns to focus on nouns because they ground objects, and most of the time, the following words form a single phrase referring to these objects. For example, attention on “lamp” for the third image is very strong throughout the generation of the whole phrase “lamp and a bunch of”. Once a new object is introduced (“oranges”), the attention shifts to this object for a different phrase (“oranges on a table”). The visualisations show that captioning transformer's masked self-attention learns global, phrase-based semantic features of sentences. In contrast, in the text-only setting, the model learns about local relations between words in a sentence. For example, distilgpt-2 continuously shifts its maximum attention after every 2-3 words are generated, indicating that it learns to capture local relations between words (“bunch of”, “oranges on”).

Attention Focus As demonstrated by Fig. 6.5, attention can constantly focus on particular words (e.g., nouns) while the caption is generated. We seek to identify which attention heads are responsible for such observed patterns in the masked self-attention of the image captioning transformer. This is potentially important for reducing the model’s complexity by pruning non-important heads, which do not have an interpretable role defined by the measure of choice. Therefore, we calculate the entropy of attention distribution (Ghader and Monz, 2017) and use it as the measure of dispersion between attention weights:

$$Ent_\alpha(s_j) = - \sum_{i=1}^{|s|} \alpha(s_i, s_j) \log(\alpha(s_i, s_j)) \quad (6.4)$$

As Fig. 6.7a demonstrates, many heads in distilgpt-2 have high entropy scores which means that attention here is highly dispersed. The entropy increases in the deeper layers of the model. This correlates with the fact that deeper layers capture more distant syntactic relations and, therefore, lead to higher entropy scores (Vig and Belinkov, 2019). Fig. 6.7b shows the entropy scores for attention heads in captioning transformer’s masked self-attention. Here, most heads have a relatively low entropy, with only some of them with higher entropy in the model’s first layers.

Do heads have high/low entropy? Based on the examples of attention heads from Fig. 6.6, we can conclude that high entropy reflects a stronger concentration of attention from target words on *particular* source words to learn *specific* information. Such pattern can be observed, for example for captioning transformer’s masked self-attention heads in Figures 6.6a and 6.6b. Note that these heads heavily link several words with nouns (e.g. “plate”), which increases the head’s entropy - many words in the target sentence attend to a single word from the context. Another important observation is that the attention distribution from target to source is not always strong: not every word on the left side has a connecting line to the right side, indicating that

attention is used to learn only specific properties. For example, as Fig. 6.6b demonstrates, focusing on “plate” when other objects (“sandwich”, “salad”) are mentioned may indicate that the model learns the notion of scene structure reflected in the text. At the same time, Fig. 6.6a shows that focusing on “plate” can be required when generating relations between objects, e.g. “plate *with* a sandwich *and* a salad”. However, as figures 6.6e and 6.6f demonstrate, distilgpt2 learns somewhat different attention between the source and the target words. While these patterns demonstrate that many words in the target sequence tend to focus on the specific words from context, each attention connection is not as strong as for the heads of the captioning transformer’s masked self-attention. The distilgpt-2 model does not focus on the caption’s specific relations or properties. Instead, it learns weak attention between all words. The heads’ entropy is high as the attention is dispersed, but each attention connection’s is also *not as strong as* it is in the captioning transformer’s masked self-attention. The examples of heads with low entropy (the right side of the Fig. 6.6) indicate that there is a word in the context that will be attended for each generated word.

6.1.6. Attention Alignment

It may be the case that the observed differences in attention patterns discussed in the previous section are simply due to different frequencies of words (in particular nouns) in the dataset on which the models are trained. For example, the multi-modal decoder also attends on the closest syntactic relations in the same way as a uni-modal decoder, but these happen to be nouns simply because there are more nouns in image captions. To test this hypothesis we calculated the Pearson correlation coefficient between the frequency of the nouns in our captions and attention distribution on the context words attended by heads when the next word is produced. The test has not shown a statistically significant correlation between the frequency of the nouns versus attention distribution on the context words in multi-modal decoder’s self-attention

($r = 0.49, p = 0.056$). However, we observed a moderate positive correlation between the frequency of the nouns versus attention distribution in the uni-modal decoder’s attention ($r = 0.60, p = 0.014$). These differences in correlations show that the uni-modal architecture is more biased to frequencies, whereas in a multi-modal setting, the effect of noun frequency is diminished. This provides support to our hypothesis, namely that this bias towards nouns is coming from somewhere else, e.g. **the multi-modal representations that the language model is grounded in.**

Since the model’s parameters are jointly updated with an end-to-end training through back-propagation, representations learned by different self-attention mechanisms are expected to be *aligned* with each other. We present a small preliminary analysis of whether the attention weights in the cross-modal self-attention (cross self-attention from Fig. 6.1) are responsible for information fusion between image encoder and text decoder. Our hypothesis is as follows: if cross-modal self-attention pays a significant portion of attention to the objects, which are generated as nouns in the caption as content words, we can conclude that due to the learning objective and nature of the information flow within the model’s components, decoder’s self-attention *aligns* with a higher-level cross-modal self-attention. In this case, we also expect that for every non-content word (e.g., determiner, preposition), the cross-attention keeps its attention on the most recent content word similar to what we observe for decoder’s self-attention in Figures 6.5e–6.5h. We use two example images and examine the differences among the top 5 most-attended objects for every word generated in image captions. We use the predicted labels from the feature extractor (Anderson, He, et al., 2018) to refer to the detected objects.

Fig. 6.8 shows changes in cross-modal attention on objects during generation of descriptions. From Fig. 6.8a we can see that every time a new content word is generated (“plate”, “sandwich”, “salad”), the cross-modal attention tends to focus on objects with labels that are similar to the generated content

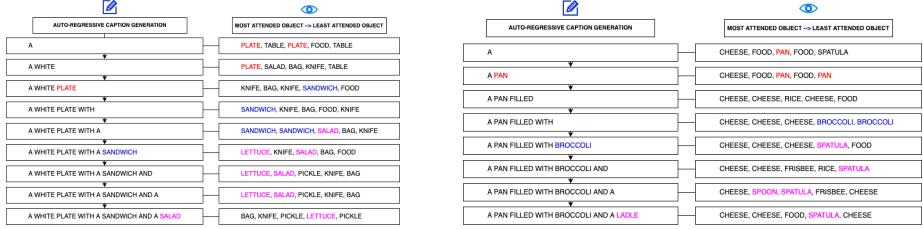


Figure 6.8. Attention shifts in cross-modal attention. The left-side column of each sub-figure shows the generated caption one word at a time. The right-side column depicts the labels of the 5 most attend objects in images when generating each word.

words. For example, “lettuce” and “salad” are among the most attended objects when the transformer is preparing to generate the content word “salad”. Also, the same objects are continued to be attended when other non-content words are generated. This example provides initial evidence how text generation of nouns as exemplified by the decoder’s attention is linked to multi-modal representations as exemplified by cross-modal attention on objects. The results suggest that in multi-modal settings models learn representations that are fused and aligned with each other. Since the self-attention in the uni-modal architecture only needs to generate the text one word at a time by taking into account only previously generated words, it learns a pattern over local syntactic dependencies. In our future work, we would like to provide a more detailed analysis of the cross-modal attention and the uni-modal visual attention and therefore further strengthen the arguments how multi-modality affects knowledge that different parts in the large scale transformer models learn.

6.1.7. Conclusion

We have shown that attention patterns learned by a sentence decoder module of a multi-modal transformer are highly affected by the task that the model is optimised for. We focused on the masked self-attention in a sentence decoder

in an image captioning transformer, demonstrating that its attention weights resemble linguistic knowledge, which is affected by the task of image captioning. This indicates that such language model acquired important aspects of grounded semantics. Simultaneously, we show that it is important to be cautious when applying large-scale pre-trained models on specific tasks to different semantic tasks as the original task does have an impact on the semantic representations learned. Our future work will focus on further examination of self-attention in the other two components of the multi-modal models which will give us an even clearer picture on what representations are learned by them.

6.2. What does a Language-and-Vision Transformer See: The Impact of Semantic Information on Visual Representations

6.2.1. Abstract

Neural networks have proven to be very successful in automatically capturing the composition of language and different structures across a range of multi-modal tasks. Thus, an important question to investigate is how neural networks learn and organise such structures. Numerous studies have examined the knowledge captured by language models (LSTMs, transformers) and vision architectures (CNNs, vision transformers) for respective uni-modal tasks. However, very few have explored what structures are acquired by multi-modal transformers where linguistic and visual features are combined. It is critical to understand the representations learned by each modality, their respective interplay, and the task's effect on these representations in large-scale architectures. In this paper, we take a multi-modal transformer trained for image captioning and examine the structure of the self-attention patterns extracted from the visual stream. Our results indicate that the information about different relations between objects in the visual stream is hierarchical and varies from local to a global object-level understanding of the image. In particular, while visual representations in the first layers encode the knowledge of relations between semantically similar object detections, often constituting neighbouring objects, deeper layers expand their attention across more distant objects and learn global relations between them. We also show that globally attended objects in deeper layers can be linked with entities described in image descriptions, indicating a critical finding - the indirect effect of language on visual representations. In addition, we highlight how object-based input representations affect the structure of learned visual knowledge and guide the model towards more accurate image descriptions. A parallel question that we

investigate is whether the insights from cognitive science echo the structure of representations that the current neural architecture learns. The proposed analysis of the inner workings of multi-modal transformers can be used to better understand and improve on such problems as pre-training of large-scale multi-modal architectures, multi-modal information fusion and probing of attention weights. In general, we contribute to the explainable multi-modal natural language processing and currently shallow understanding of how the input representations and the structure of the multi-modal transformer affect visual representations.

6.2.2. Introduction

The ability of transformers to capture contextualised representations and encode long-term relations has led to their successful application in various NLP tasks (Devlin, Chang, et al., 2019; Radford, Wu, et al., 2019; Vaswani et al., 2017). Their large size, layer depth and numerous multi-head self-attention mechanisms are the main reasons for their excellent performance. However, the structure of such ever-larger models imposes new challenges on understanding and explaining their inner workings. Since there is no clear cognitive motivation behind tremendously successful transformers (Rogers et al., 2020), a set of sophisticated methods is required to examine how information is processed and what is learned by such models. Multiple explainability methods and tools have been proposed in the ‘BERTology’ field, which investigates whether transformers can learn helpful information. In these approaches, self-attention is typically inspected for the presence of specific *linguistic knowledge* as a product of cognition. For example, some research has focused on identifying valuable information for syntactic, co-referential, and translation tasks (Belinkov and Glass, 2019; Raganato and Tiedemann, 2018). Notably, Vig and Belinkov (2019) show that more complex linguistic phenomena are captured in deeper attention heads of the model, building on top of much simpler knowledge present in earlier layers of the model. Such **hierarchical**

learning of linguistic information is further exemplified by showing that proper nouns are learned in deeper layers, and low-level constructs such as determiners are captured in lower layers. Others have inspected each attention head individually for the specific type of information (Voita, Talbot, et al., 2019), or even tried to explain attention by comparing it to human input in particular contexts (Hoover et al., 2020). However, it has been emphasized that attention is not always an explanation of the linguistic knowledge learned by the model (Jain and Wallace, 2019), and several other factors have to be taken into account when explaining such models (Kobayashi et al., 2020). Some other popular explainability methods include neuron-based analysis and transfer learning (Rethmeier et al., 2020) and promising gradient-based analysis, which directly reflects the knowledge learned by the model (Wallace et al., 2019). However, it has been recently shown that it is relatively easy to manipulate and corrupt gradient-based explainability methods (Wang, Tuyls, et al., 2020).

The transformers have also taken by storm the field of computer vision, one of the last bulwarks of CNNs. Dosovitskiy et al. (2021) have shown that vanilla transformer demonstrates impressive results on the task of image classification if supplied with simple BERT-style image representations (e.g., 2D image patches). Interestingly, the authors show that the vision transformer can gradually increase its attention on the semantically plausible parts of the image, **structuring its visual knowledge**. Specifically, attention heads in surface layers uniformly attend to many different areas in the image, with attended patches relatively close to each other. In contrast, deeper attention heads focus on specific image patches, while the distance between attended patches becomes larger. Caron et al. (2021) observed that attention heads in vision transformer capture class-specific features of objects (e.g., shapes, parts), which can be indicators of emerging *visual knowledge* of the world. Interestingly, the authors show that the model focuses more on “class-specific” features when trained with self-supervision. In contrast, using object labels in

a standard supervised setting dissolves its focus and re-distributes the model’s attention across different parts of the image. This finding raises a question of **the effect that language has on visual representations**⁴. While more focus on a single object might be beneficial for image classification, a more sophisticated multi-modal task (e.g., image captioning) requires scene-level knowledge about objects and relations between them. Thus, more global attention shaped by the conceptual knowledge from language is required for such tasks. One shortcoming of many current vision-only transformers is that the representations learned by such models lack grounding in the broader relational knowledge between different objects expressed in image descriptions. This characteristic of the vision transformers provides additional motivation for the current study and our exploration of how language-and-vision transformers can benefit from a combination of two modalities.

Somewhat surprisingly, only recently **multi-modal** transformer representations have started getting attention from scientists. Cao et al. (2020) probe the pre-trained multi-modal transformers for several language-and-vision tasks and show that these models encode a variety of useful textual, visual or cross-modal representations. However, a better understanding of **how** multi-modal representations are structured and **implicitly** learned is currently missing in the literature. Additionally, we need to know **what** is the role of **explicit** factors, such as the way the image is fed to the model. Therefore, in this paper, we address the problem of transparency of multi-modal representations and experiment with the two-stream image captioning transformer introduced by Herdade et al. (2019). In this transformer, each modality (language and vision) is first attended separately by modality-dependent self-attention, and then the two are fused by the third component, cross-modal attention. The separation of the system into three modules allows us to examine fine-grained uni-modal representations in the multi-modal architecture.

⁴Note that there is also an effect of how a target object is placed in the image. In standard image classification set-up, the target object is located in the centre of the image, intuitively simplifying the task for the model.

A dedicated module for merging of visual and linguistic information allows us to study how they are fused. Thus, a two-stream transformer can utilise the combination of conceptual knowledge of how objects can be distributed and related to each other based both on linguistic and visual information. Such models learn to perform a variety of tasks:

- ◊ Visually parse the scene: find patterns/invariances that are visually salient across different visual contexts (vision stream).
- ◊ Extract knowledge from linguistic descriptions: find salient patterns between word representations and sequences of words (language stream).
- ◊ Combine both information types to make visually and linguistically dependent representations which are grounded in how we structure and label the world as reflected in language and what we observe visually (cross-modal stream).

The analysis of representations learned by the multi-modal architectures is also relevant in the context of the call for a change in what semantic representations we use in natural language processing (Bender and Koller, 2020; Bisk et al., 2020). As the authors point out, semantic representations learned from word embeddings are insufficient and grounded representations are required. Investigating multi-modal models allows us to study how such representations differ from representations learned in completely uni-modal architectures. In Ilinykh and Dobnik (2021), we demonstrate that visual knowledge indirectly affects language representations in the multi-modal transformer. Our experiments show that the self-attention in the language stream becomes more focused on previously generated nouns, aligning with visual modality and image objects. A natural continuation of this analysis is to examine whether the structure of visual representations is influenced by conceptual knowledge of the world, reflected in language. Therefore, we examine how a multi-modal transformer proposed by Herdade et al. (2019) organises and structures learned knowledge of the visual modality. We focus on the **vision stream** and inspect (1) how visual knowledge is represented in a transformer

as exemplified by self-attention, (2) how visual knowledge is affected by the overall training task, which is image caption generation, and (3) whether the observed attentional patterns are intuitively interpretable to us.

We address the following questions:

1. Given the multi-layered nature of transformer blocks and, therefore, differences in input representations at each step, self-attention heads at each layer are expected to differ in the type of knowledge that they encode. We investigate what kind of knowledge is captured by different layers by examining visual self-attention patterns between objects.
2. Knowing that both language and perception have a hierarchical structure (Tenenbaum et al., 2011), we also expect hierarchical learning of visual information in transformers. Is there a progression of attended representations from low-level local relations to high-level global dependencies between objects corresponding to our conceptual knowledge? Moreover, is there a connection between learned dependencies and the input representations, which can be either semantically informed (e.g., object detections) or disentangled from any conceptual meaning (e.g., image patches)?
3. Does the language task have an effect on visual representations? Due to the back-propagation mechanism and the multi-modal fusion module, representations of one modality might contain artefacts of another modality in the two-stream multi-modal transformer (Ilinykh and Dobnik, 2021). Is conceptual linguistic knowledge *implicitly* reflected in visual self-attention?

The remainder of the paper is organised as follows. In Section 6.2.3.1, we review the model’s architecture and introduce the notion of the *attention link*, an important concept that we use to interpret knowledge captured by the model. We also describe our experimental setup. Then, we provide a short

analysis of how the input representations we use in our experiments might affect what and how the model learns (Section 6.2.3.2). We proceed to the main experiments in Section 6.2.4. Sections 6.2.4.1 and 6.2.4.2 describe the analysis of the knowledge in all layers in terms of thematic relatedness of the objects, visual proximity and strength of the attention links. In Section 6.2.4.3, we identify the knowledge split in what is learned by earlier and deeper layers of the model by analysing representations from different layers for similarity with each other. We also examine the spread of attention between objects as shown by attention patterns in the model’s layers. Section 6.2.4.4 describes our analysis of whether the high-level knowledge from language modality can be detected in layers of visual self-attention in some form. We inspect representations learned with a different input type (e.g., image patches) in Section 6.2.4.5 and compare them to the knowledge captured when using object detections. Finally, in Section 6.2.5, we connect our results with studies on human cognition. We conclude with a summary of how our experiments contribute to a better understanding of large-scale neural models and identify possible research questions for future work (Section 6.2.6).

6.2.3. Materials and Methods

6.2.3.1. Two-Stream Multi-Modal Transformer

Traditional multi-modal architectures (Lu, Xiong, et al., 2017; Xu, Ba, et al., 2015) learn a single set of attention weights with RNN or LSTM attending over convolutional features. In comparison, current transformer-based architectures encode information with multiple attentions, either processing each modality independently as in two-stream or multi-stream models (Lu, Batra, et al., 2019; Tan and Bansal, 2019) or simultaneously (one-stream) (Chen, Li, Yu, et al., 2020; Su et al., 2020). The separation of different modalities (*multi-stream*) allows us to (i) inspect the fine-grained attention patterns learned at each modality or level of organisation and (ii) examine the effect of

information fusion on representations of each modality. Therefore, we use the image captioning transformer by Herdade et al. (2019), which is based on the standard transformer model (Vaswani et al., 2017), consisting of three different self-attention blocks. Effectively, the modular design allows two-stream architectures to learn to encode contrasting requirements of each modality, similar to the conditions imposed by human vision and human language. In addition, the third module (cross-modal attention) fuses both types of information which is intuitively comparable with how humans combine perceptual and conceptual understanding to describe the world.

We focus on image captioning because this generation task represents a basic linguistic pragmatic case. In comparison, the VQA task (Antol et al., 2015) is a multi-label classification problem, which requires more focus on specific objects mentioned in a question. Visual dialogue (Das, Kottur, et al., 2017) imposes additional challenges, which often require knowledge beyond images and texts, e.g. memory and tracking of attention focus on objects. We believe that inspecting multi-modal representations for visual dialogue is out of the scope of the current paper and hope to address this problem in future work.

The primary difference between the model introduced by Herdade et al. (2019) and other two-stream architectures (notably, LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu, Batra, et al., 2019)) is the method that is used to encode spatial information about objects. Both LXMERT and ViLBERT do not incorporate any object relative geometry. Instead, they simply utilise coordinates of bounding boxes or their spatial location. In addition, we find BERT-inspired LXMERT and VilBERT to be more suited for learning general multi-modal representations, while the multi-modal transformer by Herdade et al. (2019) is particularly tailored for better use of modalities for a specific task of caption generation.

Figure 6.9 describes the main components of our model. The model consists of three modules, where each of them operates with different input

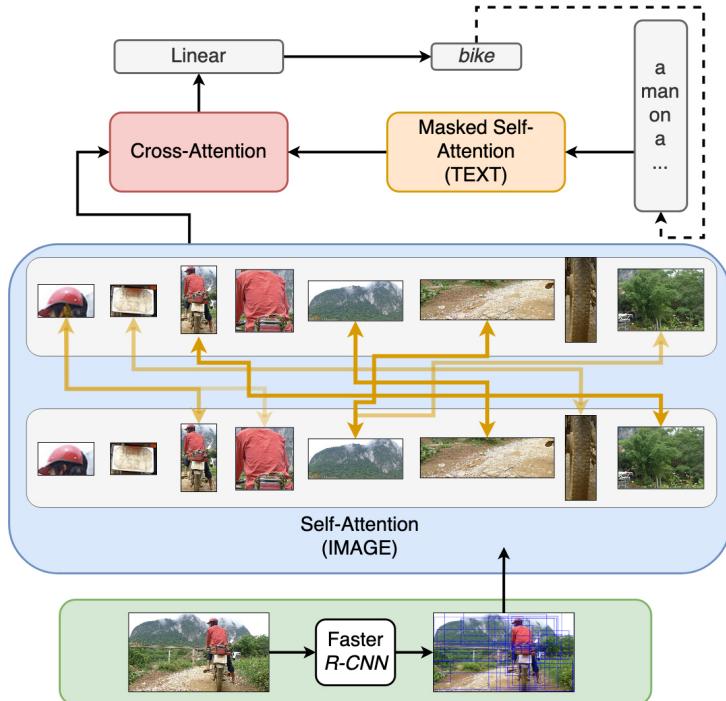


Figure 6.9. The architecture of the multi-modal transformer and a detailed visualisation of its image encoder (visual self-attention). The model consists of three parts: **self-attention on visual information**, **masked self-attention on textual input** and **cross-attention**, which learns multi-modal fusion. The predicted word is concatenated with the previously generated words and passed as the new input to the masked self-attention. The image encoder takes (i) visual representations of objects produced by a pre-trained faster R-CNN and (ii) geometric representations between detected objects. The self-attention operates at the level of detected image objects, building attention links of various strength between them. The intensity of the orange lines indicates the intensity of the attention links between objects. The attention links are created by every attention head in all layers of the image encoder and, if present, can vary in terms of attention strength.

representations and consists of $L = 6$ layers and $H = 8$ attention heads in each of these layers. For the **masked self-attention** (the orange box), we mask all words (w_{t+1}, \dots, w_T), which follow the word at the current timestep w_t , with the [MASK] token. We set T to 16. Such design is necessary for the uni-directional task of image caption generation in which sequences of words are formed gradually as the description unfolds. The masked self-attention produces representations for the next word w_{t+1} given the previous context and the most recently generated word (w_1, \dots, w_t). In our experiments, we refer to the masked self-attention module as the *text encoder* since this part of the model effectively learns the representation of the next word from the text. The cross-attention module later uses this representation with the output from visual self-attention to predict the next word along the lines of the standard decoding process.

Self-attention on visual information (the blue box) is another important part of the model. In a typical text-based transformer such as text-to-text transfer transformer (T5) (Raffel et al., 2020), this part of the model learns textual representations. However, in a multi-modal task, the self-attention is performed on image objects, which are detected and labelled prior to the model’s training. We refer to this part of the model as *image encoder* (the motivation behind the naming is similar to text encoder). To prepare input for the image encoder, we use the released feature extractor (Anderson, He, et al., 2018)⁵ that has been pre-trained on object annotations from Visual Genome (Krishna et al., 2017). This model is based on Faster-RCNN (Ren, He, et al., 2015) with the ResNet-101 (He et al., 2016) as its visual backbone. Its output consists of visual and geometric features of objects within bounding boxes, labels (“cup”) and attributes (“red”). Each attribute-label pair is accompanied by a specific score, indicating the model’s confidence about the correctness of the attribute. In our experiments, we keep only such attributes that have a confidence score of 0.1 or more. Each detected object is represented by a

⁵<https://github.com/peteanderson80/bottom-up-attention>

single feature vector $\mathbf{f}_n \in \mathbb{R}^{1 \times D}$, where $D = 2048$. All object features form a feature set $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_N\}$ with $N = 36$.

In addition, Herdade et al. (2019) also extract geometric features $\mathbf{G} = \langle x, y, w, h \rangle$ which represent the centre coordinates, width and height of every object in the image. The image encoder is provided with geometric representations, used as the positional encoding of object representations. The idea of using positional information in visual self-attention is motivated by the fact that objects in images do not have a natural order in terms of their arrangement, unlike words, and supplying models with such geometric knowledge might provide information about the topology of objects. Here we briefly describe how visual and geometric features are combined, while referring the reader to Herdade et al. (2019) for more information. First, inspired by Hu, Gu, et al. (2018), a 4-dimensional displacement vector between every two objects is computed. Similar to Vaswani et al. (2017), authors learn positional (geometric) embeddings by applying the sinusoidal function to the displacement vector and get a high-dimensional intermediate geometric representation $\mathbf{E} \in \mathbb{R}^{1 \times N \times N \times d}$. This representation is then flattened and multiplied with a learned linear matrix $\mathbf{W}_g \in \mathbb{R}^{1 \times H \times N \times N}$ and passed through ReLU non-linearity to obtain *geometric attention weights* as follows:

$$\Omega^G = \mathbf{E}\mathbf{W}_g. \quad (6.5)$$

At the same time, visual representations are used to learn queries \mathbf{Q} and keys \mathbf{K} , two standard parameters of the self-attention:

$$\mathbf{U} = \text{Dropout}(\text{ReLU}(\mathbf{W}_p \mathbf{F})), \quad \mathbf{Q} = \mathbf{U}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{U}\mathbf{W}_k, \quad (6.6)$$

where \mathbf{F} is the set of visual features or output of the previous layer, $\mathbf{W}_p \in \mathbb{R}^{D \times M}$, $\mathbf{W}_q \in \mathbb{R}^{1 \times H \times d \times N}$ and $\mathbf{W}_k \in \mathbb{R}^{1 \times H \times d \times N}$ are learned during training. \mathbf{W}_p is used to reduce the dimension of visual features resulting in \mathbf{U} . The *visual attention weights* are calculated as follows:

$$\Omega^V = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}. \quad (6.7)$$

$1/\sqrt{d_k}$ is a scaling factor, which provides efficient learning for larger inputs since their size can affect the gradient of the softmax function and make it too small. Next, geometric Ω^G and visual Ω^V attention weights are combined as follows:

$$\Omega = \log(\Omega^G) + \Omega^V, \quad (6.8)$$

where \log is used to normalise the distribution of geometric weights. Finally, the third standard parameter of transformer's self-attention, value \mathbf{V} , is multiplied with the combined features:

$$\mathbf{V} = \mathbf{U}\mathbf{W}_v, \quad \text{head}_{h,\ell}(\mathbf{F}) = \text{softmax}(\Omega)\mathbf{V}, \quad (6.9)$$

where $\mathbf{W}_v \in \mathbb{R}^{1 \times H \times d \times N}$ is the learned matrix, Ω is the $N \times N$ matrix with combined attention weights, $\text{head}_{h,\ell}$ corresponds to the specific attention head h in the layer ℓ . The authors set M to 512, d to 64 and d_k to 64. Note that the geometric features are provided to every visual self-attention layer. Every attention head at each layer learns its own set of transformer parameters (\mathbf{Q} , \mathbf{K} , \mathbf{V}) and merges *the fixed set of geometric features with the output from the previous layer*, using it to produce the final result. In the end, the output of the last layer is passed to the **cross-attention module** (the red box), which attends to both visual and textual representations to generate the next word w_{t+1} .

In our experiments, we are using attention weights between objects predicted by individual attention heads h within each layer ℓ of the image encoder. We refer to them as **attention links** to emphasise their role in connecting different objects. The attention weights are extracted from the image encoder

according to Equation 6.7. Then, we apply softmax over these representations to obtain the set of attention links. The notion of attention link represents observable attention between two identical or different objects. The attention link is *strong* if the weight that it establishes between the objects is close to 1. Otherwise, the attention link is *weak* if it is close to 0 which indicates no attention. Note that the attention weights of a specific head will typically focus on particular objects rather than all objects.

In terms of the dataset, we use the Karpathy test split (Karpathy and Fei-Fei, 2015) of the MSCOCO image captioning dataset (Lin, Maire, et al., 2014) to test our model and extract attention weights. The test split consists of 5,000 images with five captions per image, while train and validation splits contain 113,000 and 5,000 images, respectively. We take the pre-trained checkpoint of the multi-modal transformer released by Herdade et al. (2019)⁶. The choice of the checkpoint was based on the CIDEr-D score (Vedantam, Lawrence Zitnick, et al., 2015). The released checkpoint is not perfect and does not achieve human performance in generating image descriptions. In this evaluation exercise, we are interested in the attention that a model would predict for natural, therefore, human-generated descriptions. Hence, we perform the model testing in a teacher forcing setting: every next generated word is replaced with the corresponding word from the ground-truth captions and used as part of the following input to the model. We collect attention links from the image encoder for every caption generated this way.

6.2.3.2. Learning from Object Detections, not Pixels

The first layer of visual self-attention is provided with visual features of detected objects and geometric information about them. These objects become connected by attention links of various strengths at every layer during learning, resulting in many pairwise relations. We ask whether learned attention links are *influenced* in any way by the type of input that the model uses. In

⁶https://github.com/yahoo/object_relation_transformer

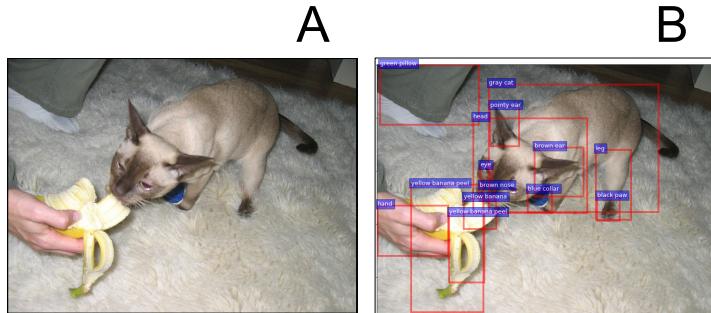


Figure 6.10. Example of object detection output from the MSCOCO dataset. The original image (A) and a subset of the detected objects represented by bounding boxes, labels and attributes (B). The caption for the original image: “a cat that is eating some kind of banana”.

particular, we want to know how the input data guides the model’s learning and what knowledge the model builds between attended objects.

Figure 6.10 shows an example of the output from the object detector. The sub-figure B demonstrates that many bounding boxes capture parts of what we would consider being the same entity: e.g. “black paw” of the “leg” which belongs to the “cat” object. The extractor can detect the same object multiple times (“yellow banana peel”), and identifying the wrong object is also possible. Therefore, the input to the first layer of the transformer consists of features of objects of different granularity, ranging from entire entities (“cat”) to their parts (‘head’). In comparison, vision transformers (Dosovitskiy et al., 2021) start with the analysis of images on the pixel level when only perceptual information is available. The image encoder in our model is expected to benefit from the *conceptual knowledge of the objects*, as provided by the object extractor, pre-trained on *human* annotations of visual scenes. The model thus might be primed to group various objects into larger concepts (“ear”, “leg”, “paw” belong to “grey cat”), acquiring lower-level cognitive information about part-whole relations and a better understanding of **local relations** between objects.

In addition, the input is represented by visual features of bounding boxes

representing objects *over the entire image* and their geometric information, and so the transformer is expected to capture **a global understanding** of an image. The self-attention in transformers can be interpreted as a parallel to larger receptive fields (Parmar et al., 2018) since it operates on a group of objects across the image. In contrast, convolution operators fall back on determining dependencies and relations between objects, while efficiently detecting objects and extracting their features (Kelleher and Dobnik, n.d.). For instance, relations such as “the leg is next to the ear” or “the banana is in front of the nose” will not be learned by CNNs because of the small size of the convolutional kernel while objects and their features (“leg”, “ear”, “banana”, “nose”) will be detected (Ghanimifard and Dobnik, 2018; Kelleher and Dobnik, n.d.). Thus, one type of knowledge that is encoded by attention weights is the knowledge of long-distance visual dependencies between objects (see, for example, the study by Ghanimifard and Dobnik (2019)).

Self-attention can capture both local and global knowledge as it is not limited by the size of the receptive field nor by the kind of information that we present to it in the vector, i.e. we can mix features of a different kind. The knowledge that we give to the transformer is a higher-level knowledge that is the output of the feature extractor, which detects objects based on the convolution algorithm. In the following experiments, we examine the attention links established by different layers of visual self-attention and the type of relations they resemble by learning from features of detected objects.

6.2.4. Experiments

6.2.4.1. Thematic Analysis of Attended Objects

We inspect to what degree objects that are linked by attention are thematically related to each other. The output of the feature extractor provides us with both object features and object descriptions. Thus, we can measure the association

between objects through semantic similarity of their descriptions. We represent the descriptions of detected objects as word embeddings by using the Word2Vec model (Mikolov, Sutskever, et al., 2013; Mikolov, Chen, et al., 2013) pre-trained on Google News dataset and available in Gensim (Řehůřek and Sojka, 2010). Next, we cluster the label embeddings into C clusters, using the k-means algorithm for clustering (Lloyd, 1982). Prior to clustering, we remove all attributes from the object descriptions since they provide an irrelevant dimension of comparison and might affect the clusters, making them less object-specific. For example, excluding “grey” from “grey table” and “grey cat” prevents situations when two descriptions of thematically unrelated objects (table and cat) are placed in the same semantic cluster of grey objects because of the shared colour dimension. We examine the cluster membership of every pair of detected objects. If they are in the same cluster, then the two objects are thematically associated; otherwise, the two objects are not related thematically. We set the number of thematic clusters $C = 3$ as this is the average number of objects described by humans in noun phrases from captions. This indicates that there are on average 3 relevant objects present in images.

We calculate the average proportion of attention links between objects that are within the same semantic cluster according to Equation 6.10:

$$\text{Prop}(\alpha \mid \ell, h) = \frac{1}{|\text{IMG}|} \times \frac{\sum_{img \in \text{IMG}} \sum_{i=1}^N \sum_{j=1}^N \alpha(n_i, n_j, clust(n_i) = clust(n_j))}{\sum_{img \in \text{IMG}} \sum_{i=1}^N \sum_{j=1}^N \alpha(n_i, n_j)}, \quad (6.10)$$

where ℓ, h stands for a specific layer and attention head, **IMG** is the test set of images, N is the number of objects in an image, *clust* denotes the cluster of a specific object, α is the attention link between the objects.

The results in Figure 6.11 demonstrate that surface layers encode visual properties within thematic categories, whereas deeper layers focus on visual properties that go beyond the automatically identified thematic categories. For example, the attention links in the first layer are created between objects

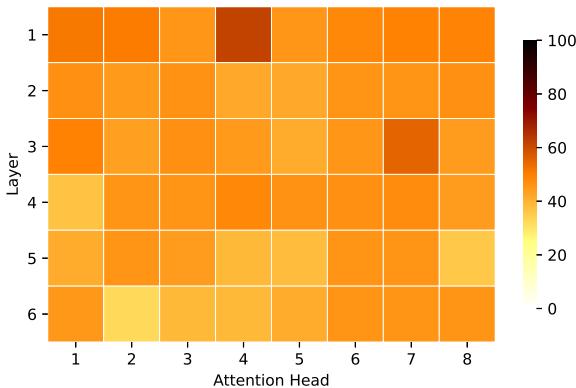


Figure 6.11. For every attention head, we show the proportion (in %) of attention links between objects in the **same** semantic cluster vs **all** attention links disregarding the cluster. The results are averaged across all images in the test set. Attention heads and layers are shown in horizontal and vertical axes accordingly. The darker the colour, the larger the ratio. The colour scale on the right indicates the range of attention proportions (minimum is 0, maximum is 100).

within the same thematic cluster on average in 50% of cases, compared to 41% in the last layer. More specifically, the top-5 (in the descending order) attention heads that link thematically related objects are all located in the first three layers of the visual encoder (1-4, 3-7, 1-1, 1-2, 3-1)⁷ with four of them located in the first layer. The best attention head (1-4) builds links between thematically related objects in 62% of cases, while the head that builds such connections the least does so in 33% of cases (6-2). The results indicate that the knowledge of thematically related objects, which possibly includes local dependencies (e.g., part-whole relations), is primarily captured in the first layers.

We support the results with the visualisations of attention links between objects in different layers in two images in Figure 6.12 and Figure 6.13. We use the tools provided by Vig (2019) for visualisations. Looking at Figure 6.12

⁷We use layer-head notation.

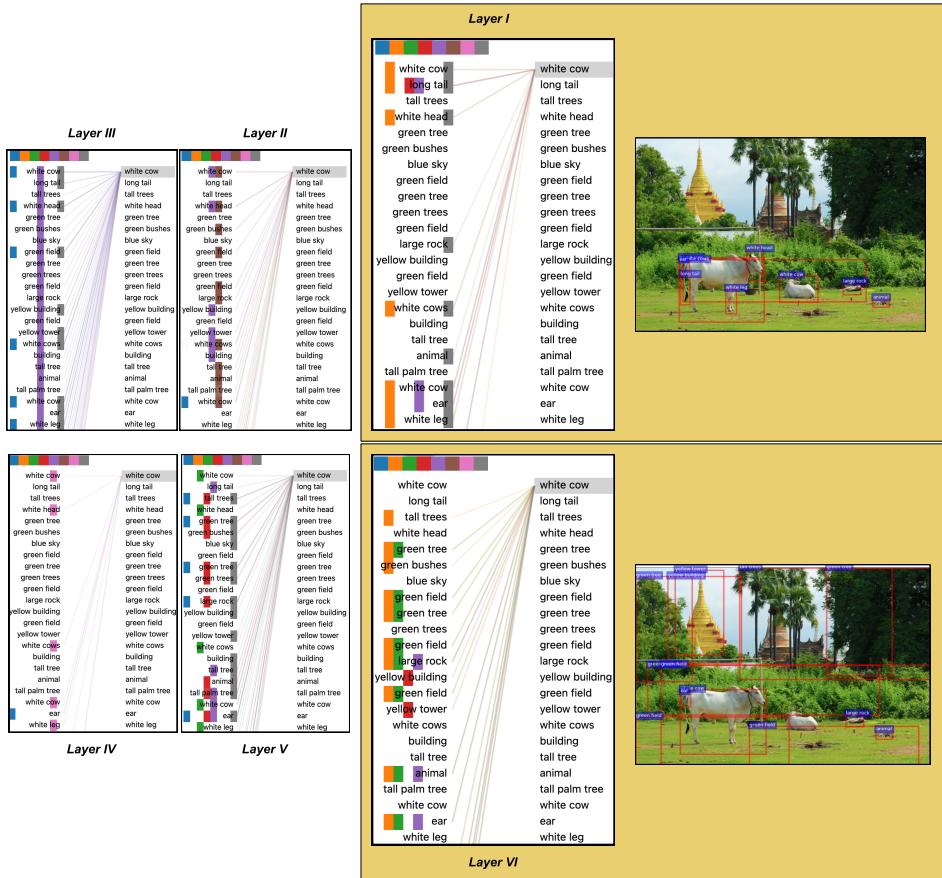


Figure 6.12. An image with (i) bounding boxes for a subset of detected objects (in red), and with (ii) visual self-attention connections for all six layers between this subset of 36 objects. Caption for the image: “two cows outside, one laying down and the other standing near a building”. Differently coloured squares on top of each layer visualisation indicate different attention heads. Each layer is displayed with two identical lists of objects: the left column shows the source objects, while the right column depicts the target objects which receive attention from source objects.

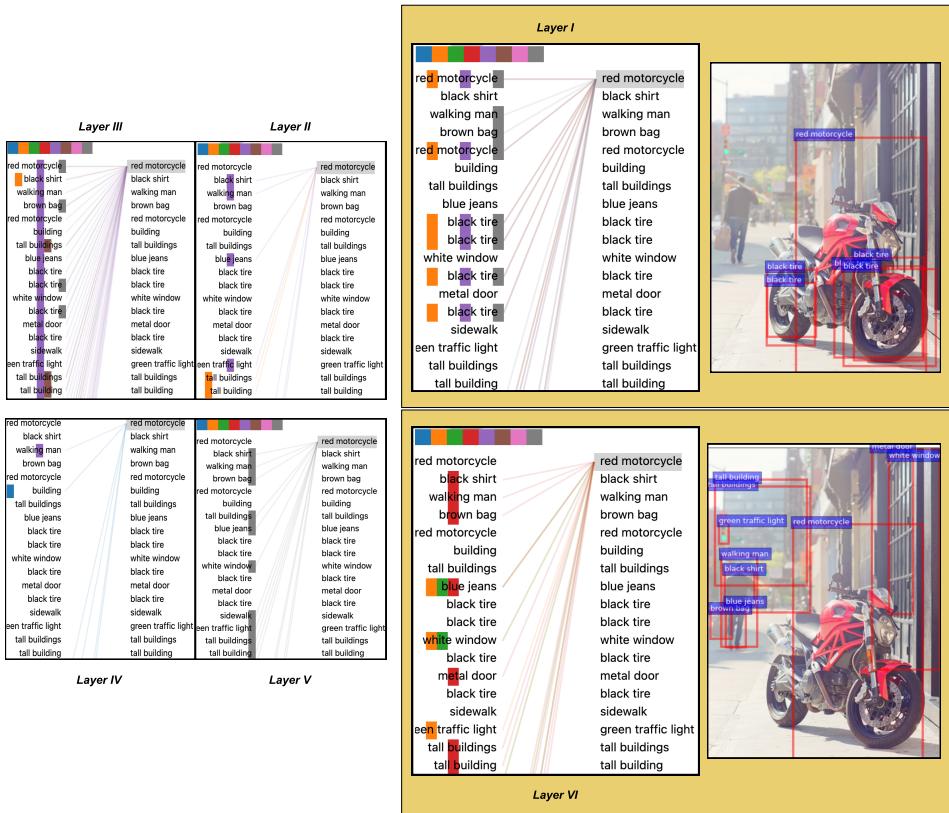


Figure 6.13. An image with attention links between objects. Caption for the image: “red motorcycle parked outside of large building in the city”. All other details about this figure are as for Figure 6.12.

(layer 1), we see that “white cow” is strongly attended by objects thematically related to cows as entities, indicating the learning of local dependencies. For example, “long tail”, “white head” and “white leg” are all parts of the entity “white cow”. Similarly, in Figure 6.13 (layer 1), attention heads relate parts of an object (“tires”) to the object itself (“motorcycle”). However, the attention links in deeper layers capture a different kind of knowledge. In particular, in Figure 6.12 (layer six), the attention heads link “white cow” with objects describing animal’s surroundings: “tall trees”, “green tree”, “yellow building”.

At the same time, in Figure 6.13, attention heads in layer six also connect “red motorcycle” with other objects in the scene: “walking man”, “metal door”, “green traffic light” etc. These differences in attention between objects in earlier and deeper layers indicate that earlier layers might focus on learning relations, which are more local such as part-whole relations. In comparison, deeper layers capture a different type of thematic relatedness, e.g. relations between *different* objects.

In earlier layers, the model operates *only* with low-level visual features of bounding boxes. Nevertheless, it can capture *semantic* similarities between objects in terms of their labels. This semantic information must come from elsewhere, possibly as a side effect of image segmentation into objects performed by the object detector. At the same time, the immediate input to the model is the set of visual object features, and learning similarities and differences between these features has made models perform so well on the number of visual and multi-modal tasks, e.g. image classification (Krizhevsky et al., 2012). Thus, we examine if semantically similar objects also share similar visual features produced by the object detector. We calculate the cosine distance \cos_{vis} between feature vectors of objects within the same thematic cluster and objects from two different clusters, following the Equation 6.11:

$$\cos_{vis} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j^\top}{\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|}, \quad (6.11)$$

where \mathbf{v}_i and \mathbf{v}_j are two feature vectors from the set of visual features \mathbf{V} , belonging to either the same thematic cluster or two different clusters for a specific image. We first averaged the scores across all combinations of object features, then across all combinations of clusters, and, finally, across images. We ignore clusters that consist of a single object. We found that the visual features of objects placed in the same thematic cluster are more similar to each other (0.50) than the objects in two different clusters (0.31). This finding indicates that semantic similarity entails visual similarity, as

has been observed in the learning behaviour of both humans (Rosch, 1975b; Rosch, 1978) and machines (Deselaers and Ferrari, 2011). This result can be attributed to bounding boxes in the input, which are in either part-whole relations and tend to largely overlap (e.g., “leg” and “gray cat”) or visually similar (e.g., bounding boxes of two cows in Figure 6.12). Thus, in earlier layers, the model links thematically related (semantic bias) and visually similar objects (visual bias). Deselaers and Ferrari (2011) also illustrate that distant objects are semantically less similar, while visually close elements are more similar to each other. In the following experiment, we examine if a similar bias is observed in our model based on the attention links of the image encoder.

6.2.4.2. The Effect of Geometric and Thematic Biases

In this experiment we inspect if there is an association between the number of attention links relating object pairs and the distance between the centres of these two objects. We make an analysis for objects in the same vs different thematic clusters. The center coordinates of each object (ObjCent_x and ObjCent_y) are calculated according to Equation 6.12, where x_{min} and y_{min} stand for the bottom-left coordinate point of the 2D bounding box covering the object, w is width, h is height. We calculate the Euclidean distance between two points in terms of image pixels. Attention weights are taken as they are without any modification.

$$\begin{aligned}\text{ObjCent}_x &= x_{min} + w/2 \\ \text{ObjCent}_y &= y_{min} + h/2\end{aligned}\tag{6.12}$$

Figure 6.14 shows the distribution of attention links for different configurations of layers and thematic clusters in terms of attention strength (e.g., high or low attention weight) and distance between two objects. The left-skewed pattern in the marginal histogram for the horizontal axis of Figure 6.14 A shows that more than 500,000 of the attended and thematically related pairs

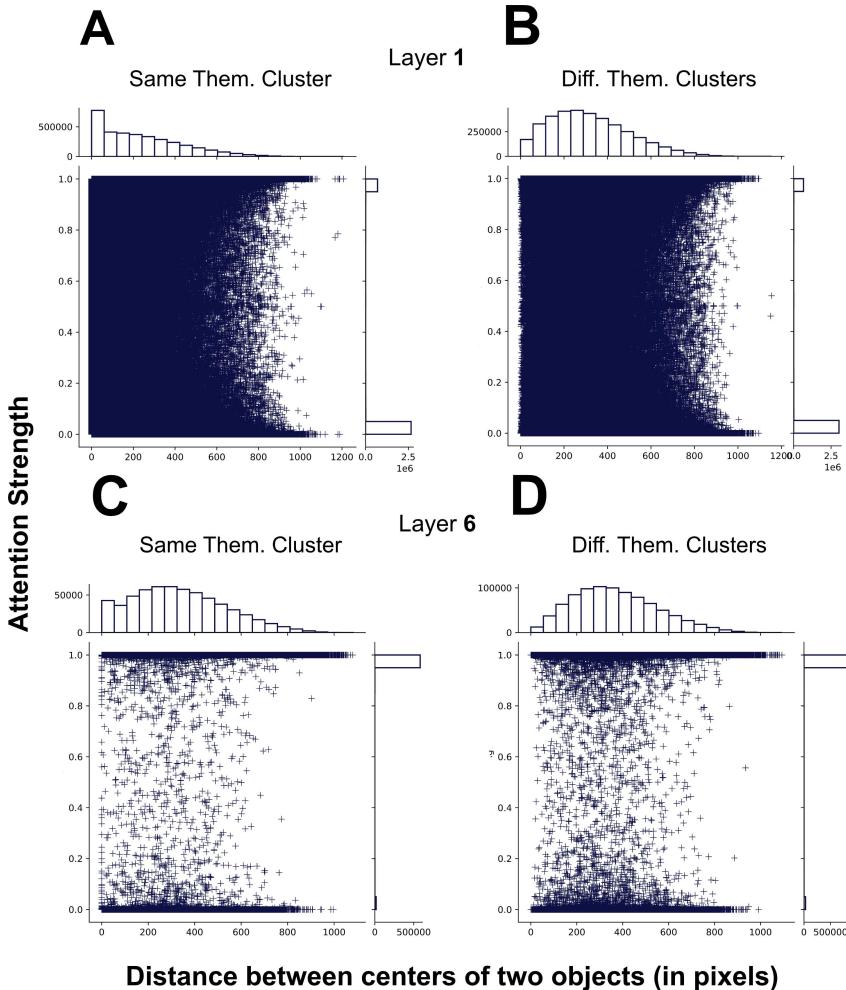


Figure 6.14. Four visualisations portraying the distribution of attention links (depicted as + marks) in terms of their strength (the vertical axis) and distance between centres of two connected objects (the horizontal axis). We split attention links into two groups: those between the objects that are either in the same or a different thematic cluster. A and B demonstrate the patterns observed in the first layer, while C and D demonstrate patterns in the last layer of the image encoder. A marginal histogram accompanies each visualisation for both the vertical (top) and the horizontal (right) axes. The histograms show a distribution of + marks for each dimension (either horizontal or vertical) in the scatter plot defined by a maximum of 20 bins. The scales for the bar sizes (e.g., frequencies of marks in each bar) are shown on the left side of the horizontal histogram and below the vertical one.

of objects are close to each other (~ 50 pixels). While there are still many thematically related objects that are not next to each other (e.g., numerous + in the range between 200 and 600 pixels), the majority of the related objects are immediate neighbours of one another. For example, in Figure 6.10 the centre of the bounding box of “grey cat” is 105 pixels away from the centre of “brown ear” and 219 pixels from “leg”. A gradual and stable left-to-right decrease of the bars’ sizes in the horizontal histogram is followed by an increasing distance between the objects on the horizontal axis, indicating that this layer learns a direct association between visual proximity and semantic similarity of objects. In comparison, we see fewer links between adjacent objects in Figure 6.14 B since according to the horizontal histogram on top of the sub-figure, there are less than 250, 000 links between objects in close visual proximity (~ 50 pixels), which is much smaller than the number of the closest thematically related objects observed in Figure 6.14 A. Also, most of the thematically non-related objects are at least $\sim 200 - 300$ apart. Note that the decreasing frequency of marks is accompanied by the increase in distances between paired objects. However, this decrease starts with the objects approximately ~ 300 pixels away from each other. The pattern for thematically unrelated objects is very similar in the last layer (Figure 6.14 D). While the overall number of links between such objects is much smaller (888, 103 vs 3, 773, 134 in the first layer), the vast majority of the objects ($\sim 100, 000$) are nearly 300 pixels apart.

On the other hand, in the last layer (Figure 6.14 C), the pattern observed for thematically related objects is different from what we have seen in the first layer. There are dramatically fewer links between neighbouring objects (< 50, 000). Most objects are also more distant from each other, between 200 and 400 pixels. The last layer builds only 600,233 links between thematically related objects, while the first layer learns 3,439,620 such links. These differences indicate that the first layer might contain links between numerous objects in part-whole relations, where parts are both thematically related and visually close to each other, e.g., “leg” and “grey cat” in Figure 6.10, while the last layer

learns thematic relations of a different kind with fewer objects, which are in a larger distance from one another, e.g., bounding boxes of two cows in Figure 6.12.

We used the non-parametric independent Mann–Whitney U test (Mann and Whitney, 1947) to examine if the differences in distances between attended objects are statistically significant between four different sets of attention links in Figure 6.14 (A, B, C, D). In this experiment, we refer to distances between objects according to the label of the corresponding sub-figure in Figure 6.14. For example, A refers to distances between thematically related objects as captured by the first layer. We found that the differences between combinations of all four sets of distances are statistically significant with extremely large U^8 and p consistently being 0.0. Such behaviour can be attributed to the size of our sets since it has been shown that statistical tests suffer from diminishing p-values when the size of the samples gets bigger, and even slight differences between large groups are considered significant (Lin, Lucas, et al., 2013). For example, A and B have more than 3,000,000 elements, while C and D have fewer items (600,000), but these numbers are still very large. Thus, we also compute Cohen's d (Cohen, 1988) to measure the effect size between two populations to estimate the degree of differences, which might give us the indication of significance. The test has shown that the effect size of significance is medium between A-D⁹ ($d = 0.701$), small for A-C ($d = 0.494$), A-B ($d = 0.423$) and B-D ($d = 0.292$), very small for C-D ($d = 0.182$) and B-C ($d = 0.096$). This result demonstrates that, in general, distributions seen in earlier layers are different from those in the deeper layers. Intra-layer, the differences are insignificant for the last layer (C-D) and somewhat significant for the first layer (A-B). In addition, the difference between set A and other sets is more significant than all other set combinations. Therefore, we argue that local relations between semantically

⁸The U values were ranging from 237983526041.0 to 1391402498035.0. Medians of the sets in pixels: A = 203, B = 295, C = 318, D = 352.

⁹The notation X-X means that we compare two different sets, e.g., A and D.

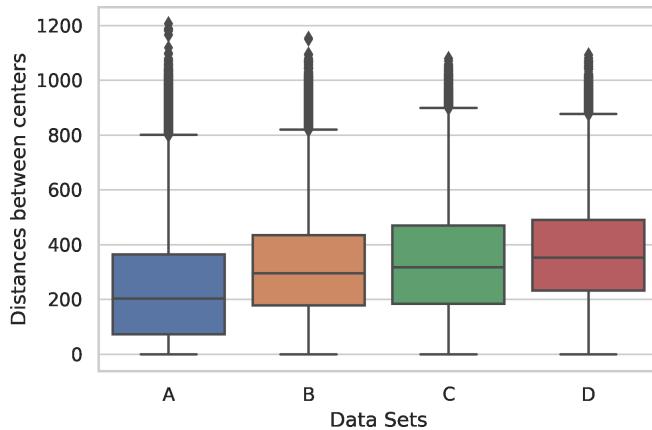


Figure 6.15. Box plots of distances between attended objects depending on thematic relatedness and depth of the layer. The vertical axis has measurements in pixels. Names of the data sets correspond to the conditions described in Figure 6.14.

similar objects are captured in the first layer, less so in the last layer. Overall, the results show that the layers of the image encoder capture different kinds of knowledge, supporting the hypothesis of separation between learning of local and global dependencies in early and deeper layers of the model, respectively.

In Figure 6.15, we provide additional analysis of the data and show the box plot of distances between attended objects for different distributions. The distance between 50% of all objects across all data sets (as witnessed by the quartiles) is lower than 500 pixels, while the outliers are nearly 1000 pixels away from each other. The data in A is skewed to lower distances more than in other sets, supporting the idea of learning of local dependencies of thematically related objects in the first layer. This result suggests that the model links two objects that are both approximately in the image's central area, not on the periphery. Intuitively, such learning mirrors the perspective of how the pictures are typically taken, e.g. most of the salient objects are usually distributed in the centre of the image, not around its corners. However, there are numerous outliers in all four sets, indicating, possibly, links between

false object detections or simply non-informative links. Also, outliers might indicate such objects in images that are far away from each other but still belong to the same thematic cluster, as is the case for A. Additionally, the outliers in terms of distances between attended objects might generally occur due to the ability of self-attention to attend across all objects in the scene simultaneously. The outliers in the first layer are more spread and distant from each other, reaching differences as much as 1200 pixels, compared to the last layer. At least half of the objects in C and D (two quartiles representing each box) are also more distant than in A and B. This finding indicates that deeper layers focus on more distant objects but not on objects that are extremely far from each other. Overall, the box plot shows that the first layer learns dependencies between neighbouring objects and a general understanding of the scene, attending between several very distant objects. In contrast, the last layer expands its focus across more distant objects and shrinks its scene-level attention, linking fewer objects on different ends of the image.

We also examine whether the strength of attention is significantly different across all four plots shown in Figure 6.14. We used the Mann-Whitney U test and found that the differences between all sets of attention strength are significant¹⁰. As we previously argued, significance testing for large data sets has to be accompanied by measuring the effect size to decide if the difference can be neglected. We used Cohen's d and found that the effect size of significance is large between sets from the first and the last layer (A-C: 2.447, A-D: 2.381, B-C: 2.661, B-D: 2.586) and very small between sets within the same layer (A-B: 0.073, C-D: 0.042). This result shows striking differences between attention strength observed in earlier and deeper layers. For example, in Figure 6.14 the strength of the attention links ranges from 0 to 1 in the first layer. Most of the links are weak, as indicated by the marginal histogram on the right side of both sub-figures. In contrast, the deeper layers are generally

¹⁰ U values were ranging from 264300345444.0 to 6185043831303.0; $p = 0.0$ for all combinations, except for C-D ($p = 2.176e - 154$). Medians of the sets in attention strength: A = $8.228e - 22$, B = $1.154e - 23$, C = 1.0, D = 1.0.

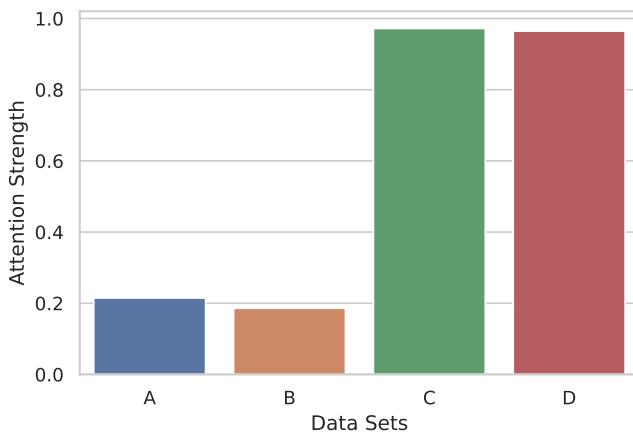


Figure 6.16. Bar plot that shows means of the attention strength in each of four conditions. The y-axis describes the strength of the attention, ranging from 0.0 to 1.0. Names of the data sets correspond to the conditions described in Figure 6.14. Standard deviation for different sets: SD(A) = 0.406, SD(B) = 0.384, SD(C) = 0.164, SD(D) = 0.184.

more confident in connecting pairs of objects as most connections are close to the maximum strength (1.0). One possible explanation for such difference is that the model operates with original features of objects in lower layers and, thus, builds a large number of attention links of various strengths. In comparison, at higher layers, the representations are built on top of the information from lower layers and might closely correspond to the conceptual knowledge of objects. Therefore, we observe more confidence in the attention patterns in deeper layers rather than in earlier layers.

In Figure 6.16, we provide a bar plot that visualises differences in the attention strength of links built in the first and the last layer of the model. The figure caption also describes standard deviation (SD) of values in each set. In earlier layers, the model has very diverse attention based on the SD values, its mean is relatively small (~ 0.2). Attention in deeper layers has much smaller SD values, and its mean is nearly 1.0. These results additionally supports the idea that deeper heads are more focused on specific relations, while in earlier

parts of the model the attention is distributed across many objects.

Overall, we have found a noticeable and statistically significant difference in the knowledge captured by earlier and deeper layers of the image encoder. The first layer, in particular, captures a general understanding of the scene by linking neighbouring objects with low certainty (small attention strength). In contrast, the last layer is highly confident in linking objects, additionally spreading its attention across more distant objects. There is also a difference in the thematic knowledge captured between the layers: the first layer might acquire information about the thematic relatedness of objects in local dependencies (e.g., part-whole relations). In contrast, the last layer broadens the notion of thematic relatedness and capture similarities between whole entities in larger distances.

6.2.4.3. Knowledge Split between Self-Attention Layers

We also examine the differences between layers in terms of the similarity of the corresponding weights with each other. Specifically, given the previous experiments on differences between the first and the last layer only, we want to inspect where (between which layers) the shift in the learned knowledge happens in the image encoder. Following Zhou, Kang, et al. (2021), we compute the cosine similarity between attention patterns observed in two neighbouring layers for all images in the test set. In particular, for every object n_i in the image, we compute cosine score between two vectors representing attention originating from this object to every other object in the image at layer k and layer m :

$$\cos_{h,n_i}^{k,m} = \frac{\mathbf{A}_{h,n_i}^k \cdot \mathbf{A}_{h,n_i}^{m\top}}{\|\mathbf{A}_{h,n_i}^k\| \cdot \|\mathbf{A}_{h,n_i}^m\|}, \quad (6.13)$$

where \mathbf{A}_{h,n_i}^* is the self-attention vector for a specific object, k and m are two neighbouring self-attention layers (e.g., layer three and layer four), and h is

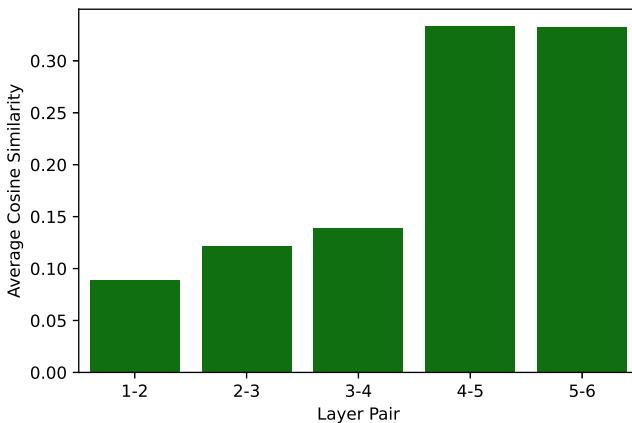


Figure 6.17. The average cosine similarity between attention weights in neighbouring layers. The horizontal axis shows the specific pair of layers that the scores were calculated for, while the vertical axis shows the cosine scores.

the attention head. The final cosine similarity scores for each attention head are averaged over all objects (divided by N). We also average the scores over heads and images to obtain a single score per pair of layers. Here, we examine the similarity between attention patterns of *neighbouring* layers only since we want to inspect how visual knowledge is sequentially processed from earlier to deeper layers.

Figure 6.17 shows the results. The attention patterns are highly dissimilar in the first three layers. In contrast, deeper layers (4, 5, 6) encode more similar knowledge, showing an increase in similarity by almost 0.25 points. This indicates that the results of the analysis in Sections 6.2.4.1 and 6.2.4.2 can be valid for the layer 4 and 5 because of their strong similarity with the last layer. At the same time, the dissimilarity between the first three layers can be explained based on what we know about the first layer, which builds many different links of various strengths between the objects. Thus, it is possible that layers 2 and 3 also build a large number of varied links of different strengths. In general, the results suggest that the shift in similarity occurs somewhere in

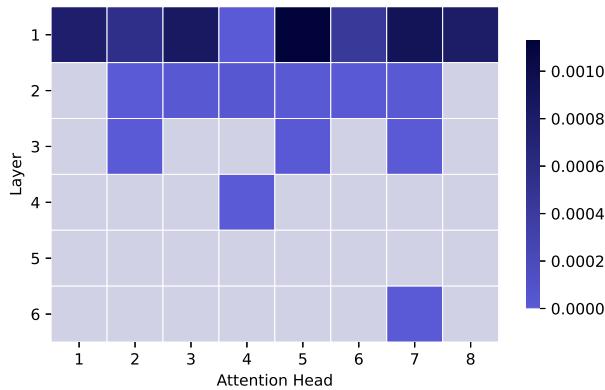


Figure 6.18. We show the normalised entropy of visual self-attention heads. The result has been normalised by the maximum attainable entropy $\log(N)$. The darker colour indicates higher entropy.

the middle of the image encoder. Such change can be attributed to the type of information each layer operates with and how.

We also examine the level of dispersion of attention links in different layers of the model and compute the attention entropy E of the attention distribution α for each attention head according to the Equation 6.14, where s_i and t_j are specific source and target objects, α is the attention value between them.

$$E_{\alpha_{\ell,h}}(t_j) = - \sum_{i=1}^{|S|} \alpha(s_i, t_j) \log(\alpha(s_i, t_j)) \quad (6.14)$$

The results in Figure 6.18 show that all attention heads do not contribute equally to the formation of attention links. In general, the entropy is highest in the first layer and slowly diminishes with the depth of the model. This pattern demonstrates that attention converges to relate fewer objects in deeper layers compared to the earlier ones. Based on the results in Section 6.2.4.2, this knowledge is also established between more distant objects. This shows

that the attention in deeper parts of the model becomes saturated and focused on specific relations, possibly reflecting the knowledge of core entities in the scene, perceived on the scene level, not local level. Overall, the image encoder captures at least two types of knowledge, which can be associated with either the first layers of the model or the deeper ones. Disperse and dissimilar attention in the initial layers indicates that the model starts with learning general and varied aspects of the scene. In contrast, highly focused and similar links in deeper layers show that the model converges to a more concrete understanding of the image, potentially establishing task-dependent relations between objects. At the same time, the representations in deeper layers are not extremely similar (e.g., they do not reach a 0.9 score or higher). It indicates that the attention patterns in these layers still show active learning of diverse connections between image objects and are less likely to encode any layer-redundant information. Note that it is much harder to establish strong attention links in earlier layers, given only a few steps of visual information processing in previous layers. The confidence of deeper layers, thus, might be not only built on top of the previous processing but also shaped by the task.

Eventually, the image encoder is required to produce valuable representations for caption generation, and the last layer outputs such representations. In turn, captions describe only specific relations between objects and not the whole set of all possible relations. Thus, the reasons for the presence of focused relations in deeper layers can be two-fold. First, the emergence of more concentrated knowledge can be due to the learning of complex information from low-level image understanding in the initial layers, as observed in the experiments. Second, indirect interference of the language task, leading to high-level conceptual knowledge of the scene in deeper layers, can also be an important factor. Representations from these layers are used as input to the cross-attention module, which fuses them with language representations to generate the image description. In the following experiment, we examine whether there is an association between the caption and representations from

the last layer of the image encoder.

6.2.4.4. Tracing the Knowledge of Language in Visual Representations

Here, we examine if deeper layers of visual self-attention achieve a better pairing of **noun phrases in captions**, typically describing objects in relations on the global scale, with **objects in the scene** through the attention links. By inspecting whether visual representations include signs of conceptual (language) knowledge, we enrich our understanding of how different modalities affect each other in the multi-modal transformer (see our work in Ilinykh and Dobnik (2021) for the discussion on how language representations are affected by visual information).

Our goal is to compute the proportion of attention links, which mirror the pairing of objects and noun phrases in captions. In our analysis we name objects that receive attention as **target** objects, normally shown on the right side of attention visualisations (e.g., Figure 6.12 and Figure 6.13), while objects that are source of attention are named as **source** objects. Note that each target object might be connected with any other object in the scene, receiving attention from multiple source objects. This can lead to such attention links, which are hard to interpret since the target is attended by many different source objects. An example of this can be observed in Figure 6.13, Layer 3: the attention head represented in purple predicts the “motorcylce” as the target of attention for all the source objects on the left. Similar to *null attention* observed in the analysis of textual models (Vig and Belinkov, 2019), we treat such attention links as non-informative. Specifically, if a target object is linked with more than 30 source objects, we ignore all such links for the current attention head.

Next, we prepare the set of noun phrases and object labels for linking. We use spaCy¹¹ to identify and extract noun phrases from image captions. Since the labels of the detected objects describe specific *objects* (e.g., “a cat”) rather

¹¹<https://spacy.io/>

than other parts of the scene (e.g., “the right corner”), we perform additional filtering of noun phrases. We remove any phrase which contains at least one word from the special word list¹². We keep adjectives in noun phrases since object labels are typically provided with attributes. Note that numerals and determiners are removed from the noun phrases in order to reach structural similarity with object descriptions for the linking process. After obtaining the set of noun phrases from the caption, we process the detected objects to collect their labels. For every object, we retrieve the predicted label (noun) and its attribute (adjective). The attribute is removed if the extractor is not confident about the attribute’s correctness, when the confidence is lower than 0.1.

We want to inspect whether the linking between the target object and source objects corresponds to *two different* noun phrases, which are also related, but in the context of image description. According to the attention links between objects (e.g., Figure 6.12), a target object (the right side of visualisations) is often linked with multiple source objects (the left side). While linking one object label depicting the target with the noun phrase is relatively simple, connecting multiple source objects with a single noun phrase is not straightforward. Objects might be associated with more than one noun in the caption, and the other way around, a caption might be related to more than one visual entity in the image. We take a simpler approach and inspect if at least a subset of source objects for a specific target can be grouped into a single thematic cluster, thus, describing a single entity. By identifying from all the detected objects those representing the core scene entities, we ensure that the large part of different links captures identifiable relation for a single target. For example, if many source objects can be associated with a single entity that is linked with a different entity depicted by the target (e.g., the target “cow” is attended by a large subset of source objects which form a thematic

¹²The list of words used to filter noun phrases: [‘right’, ‘left’, ‘top’, ‘bottom’, ‘the left’, ‘the right’, ‘the top’, ‘the bottom’, ‘the back’, ‘the front’, ‘back’, ‘front’, ‘far’, ‘the far’, ‘close’, ‘the close’].

cluster of “building”), the attention head can be interpreted to be confident in linking these two entities. In contrast, highly diverse links between target and source objects (e.g., the target “cow” is linked by several kinds of objects representing “building”, “trees”, and “street”) leads to a weaker knowledge of specific relations in the scene since no semantic category is dominating the weights. Such a spread of focus might lead to less interpretable and diluted knowledge. To distinguish between confident and non-confident patterns of attention, we cluster source objects based on the same strategy we used in the first experiment (Section 6.2.4.1). In particular, we impose a relatively soft requirement and examine if at least 25% of the source objects can be grouped into a single thematic cluster. We ensure that this cluster is different from the cluster of the target object. If we can not group at least a quarter of source objects’ labels into a single cluster, we ignore the whole set of links for this particular set of source objects and a target. This simple mechanism ensures that we select only those attention links between objects that are strongly focused and connect entities from two different thematic clusters.

Once the objects are clustered, we match their labels with noun phrases by computing their semantic similarity score. We use BERTScore¹³ (Zhang, Kishore, et al., 2020) to get the most semantically similar object label for every noun phrase. This model allows us to use the power of pre-trained contextual BERT embeddings (Devlin, Chang, et al., 2019) to match each noun phrase with every object description and the other way around by computing cosine similarity. The model also correlates well with the human judgements and outputs results of multiple performance metrics: precision, recall, F1 score. A recall metric is computed as the averaged sum of maximum cosine similarities between each token in the noun phrase and a token in the object description. Precision is calculated similarly, but between each token in the object description and a token in the noun phrase. F1 score is a classic combination of precision and recall. In the end, we receive a single F1 score for

¹³We take the best performing layer (40) of the microsoft/deberta-xlarge-mnli model.

every combination of a noun phrase and the whole set of object descriptions, which can be either targets or sources of attention.

We match the source objects and noun phrases as follows. First, as described previously, we take the subset of object descriptions if one-fourth (the size of the subset) of the whole set can be grouped into a single cluster. F1 scores of the nine objects in the subset (25% from $N = 36$) are ranked from highest to lowest for each noun phrase. We examine if there is an F1 score higher than 0.6 in at least *one* of these subsets of source objects to ensure that the word similarity scores are sufficiently high. If that is the case, a set of nine objects is linked with the noun phrase that has the highest similarity score with this set compared to other noun phrases. For example, “a cat” will be chosen as the linked noun phrase for nine source objects if their highest F1 score is 0.8 for the object description “white cat” from this set, compared to a different noun “the table”, for which the highest F1 score (0.6) for an object description from the set is smaller. The target object’s label is inspected for the biggest F1 score across all noun phrases and matched with the noun phrase with an F1 score higher than 0.6.

Given two noun phrases describing the source and target objects, we check if these two phrases are different. If these phrases are different, we proceed to calculate the proportion of attention links between mapped objects vs all objects for all images and target-source connections. Note that if we were able to successfully map nouns with object descriptions, the computed proportion for the specific target is $\text{Prop}_{\text{target}} = 9/36 = 0.25$. We get the final results by averaging these proportions across all targets and images.

The results are shown in Figure 6.19 in percentages. The knowledge of relations between two different entities mentioned in the caption and expressed as linked objects is more pronounced in deeper layers of the model. In fact, Figure 6.11 shows that deeper attention heads are more active when connecting objects in different thematic clusters, which possibly correspond to two different noun phrases. These objects describe entities that consti-

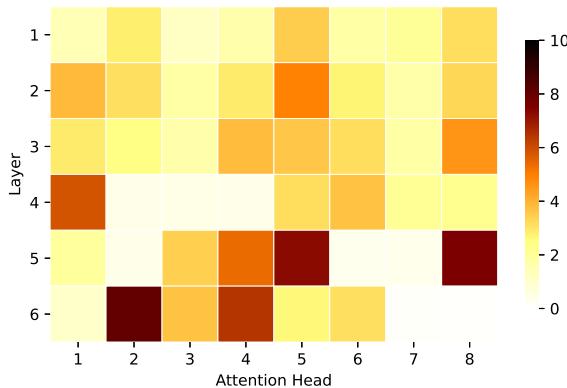


Figure 6.19. The proportion of attention links between target and source objects which can be associated with at least one noun in the caption. The nouns cannot be identical (e.g., “cows” and “building” in “two cows outside, one laying down and the other standing near a building”), and both the target and the source objects linked with them must correspond to two different thematic clusters. For better visualisation, the highest value is set to 10%.

tute the visual scene as a whole and are also more distant from each other (Figure 6.14B). For example, in Figure 6.10 the “grey cat” is 305 pixels away from the “yellow banana”, which is much further than the distance between the vast majority of semantically similar objects (~ 50 pixels, Figure 6.14A). In addition, the caption for this image (“a cat that is eating some kind of banana”) describes it with global relations between objects, not with local, thematic relations (e.g., “a cat with paws and a tail is eating some kind of banana with a yellow peel”). The structure of image descriptions depends on the pragmatic context of the task and instructions given to the describers. In the case of the MSCOCO dataset, the instructions forced the participants to describe the image as a whole with a single sentence. These regulations have primed humans to mention only a subset of the most important objects in the scene, leading to 11.30 words per sentence on average, according to the statistics of MSCOCO dataset. We have identified that such knowledge of described objects is concentrated in deeper attention heads, indicating the

traces of language information in visual representations. This finding also corresponds to the idea that deeper layers provide cross-attention with the necessary task-dependent representations, encoding more global knowledge of the scene in terms of whole entities (not numerous detections of objects of smaller granularity)¹⁴.

Note that the attention heads in the first layer have bright uniform patterns, indicating lesser mapping of nouns in captions and object descriptions. We attribute such knowledge to the design of self-attention, which might somewhat understand relations between objects on the scene level in the initial layers. However, this knowledge is less expressed in the first layers and more articulated in deeper layers. Besides, the patches for multiple heads in deeper layers are nearly entirely white. We believe that these heads might also capture mapping between nouns and objects but in terms of a different and hypothetical image description. For example, one can describe images in various ways (e.g., each image in MSCOCO comes with five descriptions). Figure 6.10A can also be described as “a cat is on the white floor” or “someone is feeding a cat with a banana in their hand”. In these texts, the focus is on such image entities that differ from those mentioned in the original caption (“a cat that is eating some kind of banana”). It is possible that due to the nature of self-attention, deeper layers were able to capture these potential relations between different objects as well, given that the model was provided with all five captions for each image during training. In addition, similar entities co-occur across images in different conditions and configurations. This type of knowledge might give self-attention even more power to juggle the observed objects in myriad ways. However, an additional set of experiments is required to unveil what else is learned at the output of the visual self-attention, and we leave this for future investigation.

¹⁴Note that some descriptions might be not necessarily global, e.g. “a cat that has whiskers and ears”. Thus, there is a noise in the model because there is no guarantee that descriptions only describe details at a scene level.

6.2.4.5. Representing Image with Patches, not Objects

In our experiments, the model attends across the visual features of bounding boxes, which correspond to objects in images, providing the model with semantic information about images. Such representations have shown to be more suitable for image captioning and visual question answering (Anderson, He, et al., 2018; Li, Tang, et al., 2017). However, more recently, dividing images into a uniform grid of patches of the same size and feeding them to a BERT-inspired visual transformer has demonstrated improvements on the task of image classification Dosovitskiy et al. (2021). Note that prior to their application in vision transformers, grid-like representations have been widely used in encoder-decoder networks for image captioning (Lu, Xiong, et al., 2017; Xu, Ba, et al., 2015), but were shown to be less informative than object-level representations.

Here we perform an ablation study, examining the distances between attended objects in two settings: (1) when the model is given grid-level features (image patches), (2) or when the model uses object-level features. Note that we used object-level representations in all other experiments in this manuscript. We divide each image into the set of 6×6 patches, resulting in 36 patches per image, comparable with all our experiments in which we detected 36 object regions. Next, we pass image patches through ResNet-101 (He et al., 2016) and extract the set of patch-based visual features $\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_N\}$. These representations are used as an input to the self-attention layers similar to linearly projected patch embeddings in Dosovitskiy et al. (2021). Our feature extractor is fixed and not updated during training. We train the model from scratch on the set of image patches and the corresponding caption. We follow the instructions in the original paper by Herdade et al. (2019) and use the official code implementation¹⁵ to train the model in two stages: first, with a standard cross-entropy loss, then we use self-critical reinforcement learning (Rennie et al., 2017). The model is also provided with the relative positions

¹⁵https://github.com/yahoo/object_relation_transformer

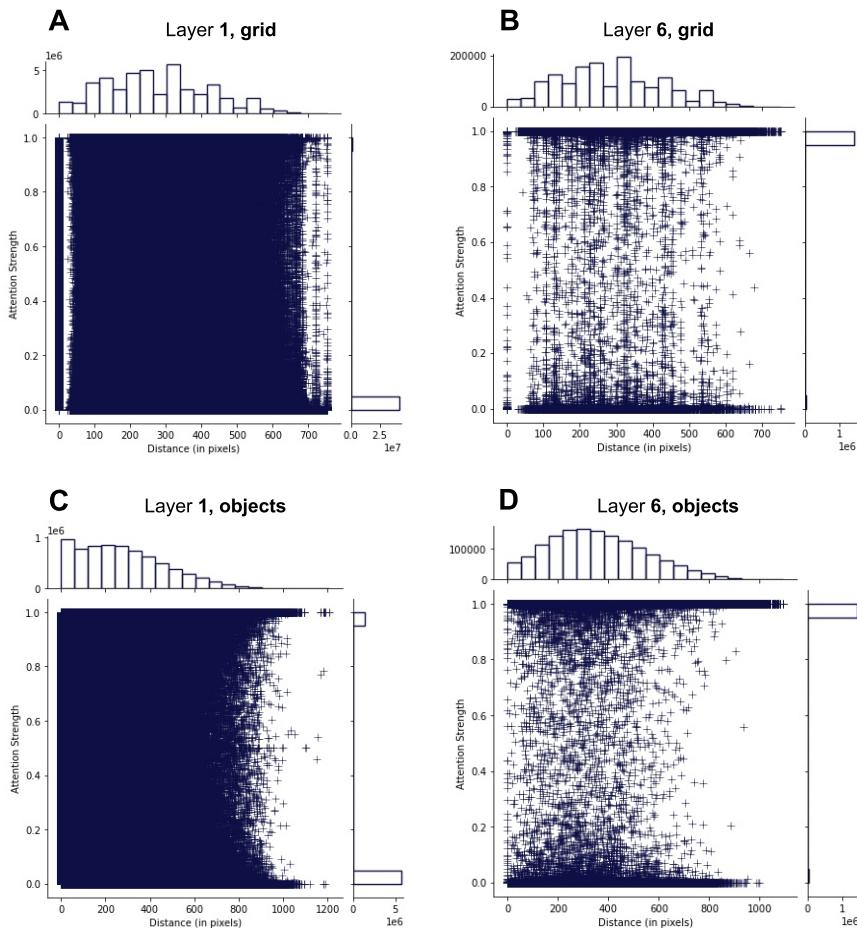


Figure 6.20. The distribution of the attention links (depicted as + marks) in terms of their strength (the vertical axis) and distance between centres of two connected objects (the horizontal axis). A and B correspond to the patterns from the first and the last layer of the model for grid-based features (e.g., image patches), while C and D represent links when we use object representations as input features. We disregard objects' thematic clusters in these visualisations for a fair comparison with the grid-based approach. Information about other parts of the figures (e.g. histograms) is identical to the description in Figure 6.14.

of image patches. We test the model on the Karpathy test split (Karpathy and Fei-Fei, 2015) in a teacher-forcing setting and extract its attention links with the corresponding weights. The grid-based representations are not informed semantically; hence, they might not correspond to a semantically plausible object in the region. Therefore, we perform the analysis of distances between attended parts of the image identical to the geometric analysis in section 6.2.4.2.

Figure 6.20 shows the distribution of attention links when the model is provided with either image patches (**A**, **B**) or object representations (**C**, **D**). We focus on the analysis of distances between attended objects, similar to our experiment in section 6.2.4.2. First, note that the two distributions of attention links between layer one and layer six for the grid-based approach are visually very similar (**A** and **B**, histograms on top). We compute Cohen's d score to measure the effect size between the observed populations to get the estimation of the degree of differences since a large size of the data makes results of significance testing non-informative, which has been observed in section 6.2.4.2. Our results have shown that the effect size is small for A-D ($d = 0.489$), B-D ($d = 0.422$), and C-D ($d = 0.402$), very small for A-B ($d = 0.082$), A-C ($d = 0.030$), B-C ($d = 0.039$). Layer-wise, we observe that object representations affect the structure of the learned information in earlier vs deeper layers, which is reflected in a more noticeable difference between C-D, while this difference is practically absent between patterns in A and B. This leads to the conclusion that semantic information indeed provides *more structure* to what and how the model learns. Also, the distances patterns that are learned with grid-based features have not shown a large effect size of significance when compared with the patterns captured in the *first* layer of the object-based approach (**A** and **B** vs **C**). In contrast, the effect size becomes bigger when comparing patterns in both **A** and **B** with distances in the *last* layer of the model that uses semantically informed representations (**D**). Combined with the results in section 6.2.4.4, we conclude that there is no clear

distinction between local and global knowledge in layers of the model when it is provided with image patches. Overall, our ablation experiment has shown how semantic information allows the model to organise the information that it learns hierarchically. We note that our results echo what has been observed by Raghu et al. (2021), who have shown that visual transformers that operate with image patches (aka grid features) do not structure their knowledge in the context of the image classification tasks.

6.2.5. Discussion and Implications

Our primary goal in this paper was to identify what kind of representations are learned in the vision stream of the multi-modal transformer. Specifically, we have examined how different parts of the visual self-attention learn to attend between objects of varying granularity. In addition, we have also inspected if visual representations contain artefacts of world knowledge (language). Our model benefits from higher-level object detections and builds a variety of relations between *objects*, in contrast with the vast majority of vision transformers that receive pixel-level image representations as their input and lack grounding in richer language representations. In addition, the CNN-based object detector produces numerous bounding boxes, frequently corresponding to what humans consider parts of a single object. This, in turn, affects the type of knowledge our model captures or “sees” when attending across objects. The effect of using semantically informed visual representations has been validated by the ablation experiment in Section 6.2.4.5. Figure 6.21 illustrates how all of the results reported in this paper can be placed in a single visualisation. We conclude with the following results:

- Thematic analysis of objects based on their labels has shown that lower layers more frequently form attention links between objects, which belong to the same thematic cluster.
- Analysis of distances between attended objects demonstrates that objects in the same thematic class are also closer together. Thus, lower

layers encode knowledge of local dependencies between objects, which might correspond to part-whole relations.

- The attention strength increases in deeper layers and stays at much lower levels in the earlier layers. This finding reflects gradually increasing confidence of the model about the objects that it has to relate for a better understanding of the whole image.
- One important finding is the effect of the input representations on how and what visual self-attention learns. We have demonstrated that

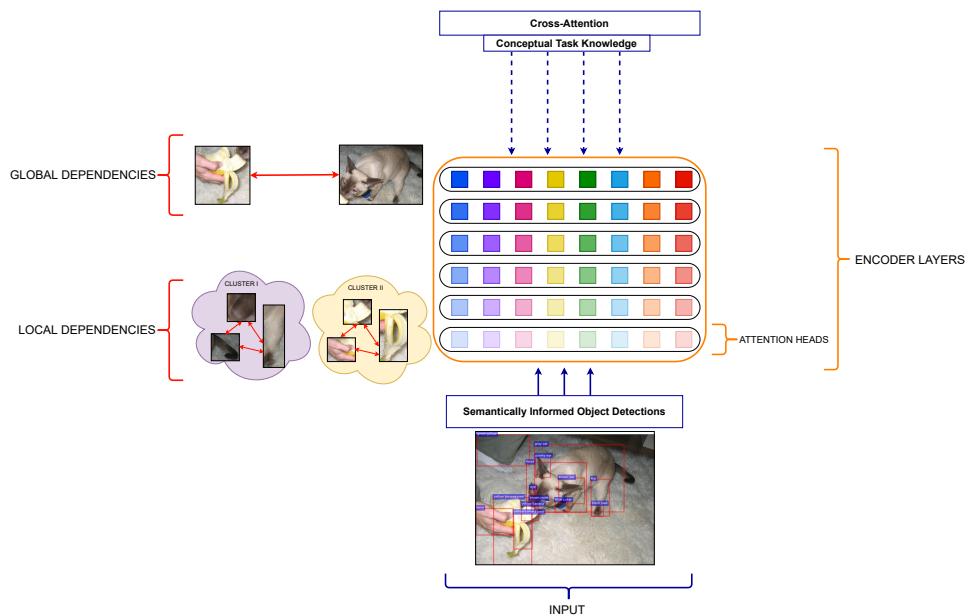


Figure 6.21. A visual conceptualisation of the most important findings of the current study. The colour intensity of each attention head increases with layer depth, resembling the increase in attention strength. **Left:** as shown, earlier layers capture more local dependencies between thematically related objects, while deeper layers connect distant objects. **Below:** the input to the model is the set of *semantically informed* features of detected objects, which is in contrast to the grid-based approach, where image patches carry no semantics. **Above:** we also denote an indirect influence (dotted lines) from the cross-attention module, which possibly occurs due to backpropagation and is reflected in the heavier grounding of noun phrases in deeper heads that we examine in Section 6.2.4.4. Overall, this figure demonstrates the whole variety of findings that we present in this manuscript.

higher-level object detections structure the internal representations of the model, while patch-based representations do not inform the model about object semantics; hence, there is no gradual learning from local to global knowledge.

- Lower layers compared with each another lead to very dissimilar attention patterns; higher layers exhibit more similarities. This result shows a shift in knowledge between the layers, with deeper ones being more specialised.
- Larger entropy in the earlier layers indicates their dispersion and attention between many different combinations of objects. In contrast, deeper layers are much more confident and focused on individual objects, indicating learning of saturated information about relations.
- Lastly, matching attended objects with noun phrases in image descriptions indicates that objects referred to in a description are more often attended in the deeper layers. This result shows the indirect influence of language on the model's visual representations in the context of the image captioning task.

The first conclusion that we make here is that visual information is built-up **hierarchically**, starting with learning of local dependencies between objects and finishing with more global understanding of the scene, also in terms of the described objects. Similarly, structured representation learning has been observed for linguistic information, starting with more local dependencies in earlier layers and expanding to the global dependencies (e.g., subject-verb agreement) in deeper layers of BERT (Jawahar et al., 2019). In the beginning, the multi-modal transformer connects objects which are visually, semantically and locally close to each other and expands onto learning relations between more distant objects, collecting information about the image as a whole. In earlier layers, the model builds many links of lower strength between objects. In deeper layers, the model is re-distributing its attention towards fewer objects, but with much more confidence, reflected in larger

attention strength. The deeper layers are affected not only by local visual dependencies from previous layers but also by the pragmatic nature of the image captioning task. In this task, there is a preference for depicting global scene-level relations in the caption, as participants were instructed to describe a scene with a single sentence. At the same time, all modalities are a part of a system as they are trained together, and therefore, one does expect their co-influence. Both language and vision information might optimise each other because the model needs to find a sensible mapping of one to another. Thus, due to the back-propagation of the information in the model (from caption to cross-attention to visual self-attention), deeper layers of visual self-attention might be indirectly guided by conceptual knowledge. At the same time, as our experiment in section 6.2.4.5 has shown, the structure of the learned knowledge is directly affected by the type of input representations (e.g., image patches or semantically informed object detections) given to the model.

The hierarchical processing of visual information has also been observed in humans. Our results recall the results of Hubel and Wiesel (1959): visual information is processed hierarchically with simpler biological cells responding to such phenomena as light orientation and more complex cells capturing movements. However, we can build a stronger parallel between our results and **the theory of visual routines** (Ullman, 1984). According to this theory, humans process visual information in sequential order, starting from more straightforward representations and later applying task-dependent rules to them. In particular, *base representations* are such features, which depend solely on visual input and typically are 2D image patches or sketches. Such features encode basic information about the scene: depth, colour and orientation. They are uniformed and bottom-up driven. As the next step, *visual routines* are applied to base representations to produce more complex features. Visual routines, in turn, are divided into a two-step process: in the first step, *universal routines* are used to achieve some general understanding of the scene. Univer-

sal routines constitute the set of rules that allow us to perform some initial analysis of the scene and capture general aspects of the scene to isolate objects and describe their colour, shape, and other characteristics. Such routines are not task-dependent, and they are required to define such information, which more complex visual routines can later use. For example, universal routines might provide us with sufficient information to classify an image as a whole with objects in it. Finally, given the holistic understanding of the image from applying universal routines, we use more specific and top-down driven visual routines to complete the task at hand. When explaining visual self-attention and its processes through the prism of visual routines, it is possible to say that its input are base representations (object features), with different layers learning both universal and visual routines. In other words, the first layers of the model would mirror the behaviour of universal routines, capturing general information about the image. In contrast, deeper layers are specialised in building relations regarding the global arrangement of objects and their importance for the captioning task.

Moreover, humans learn about elements ‘below’ object level (e.g., the set of semantic and visual features corresponding to part-whole relations) and ‘after’ (e.g., identification of object relations) (Ben-Yosef and Ullman, 2018). This structure mirrors the process of learning of local and global information observed in visual self-attention. Local knowledge, in particular, can also be related to the innate ability of humans to represent objects through the hierarchy of their parts. The principle of compositionality can be observed not only in the language in the sense of Fregean tradition but also in human vision, e.g. computational vision (Geman et al., 2002). Parts often cannot be represented in isolation, without any contextual constraint on them, that would allow them to form a coherent whole. Consolidating parts into a whole is not task-dependent and can be seen as a bottom-up guided feature of human cognition. Yosinski et al. (2014) show that such basic abilities may be considered as general knowledge that is acquired in earlier layers.

We have also observed that the objects attended in deeper layers are more likely to be matched with noun phrases in image descriptions. This result suggests that conceptual (language) knowledge is indirectly present in visual representations of deeper layers. Here, we show that this result corresponds to the insights from a well-developed **load theory of selective attention and cognitive control** (Lavie, Hirst, et al., 2004). As this theory suggests, in terms of human cognition, attention acts through two selective mechanisms: *perceptual selection* and *cognitive control* (Dobnik and Kelleher, 2016). In the generation of image descriptions, both of these processes are required. In particular, describing an image is a complex cognitive task as the describer must select what information to include in the description. Using all perceptual and background knowledge, humans decide what objects they should mention. Thus, the perceptual selection is identical to filtering perceptual information by our sensors, while cognitive control is a selection of elements from this information given the task. In other words, humans cognitively control *when and how* to describe specific objects and relations between them, which are perceived and filtered visually. Note that the pragmatic nature of the image captioning task places strong restrictions on what is to be described. For example, in the task of image paragraph generation as Ilinykh and Dobnik (2020) note, there is a progression in image description from the general (“an image with two chairs.”) to more specific knowledge (“the chair on the left is black”). In this case, the restrictions on filtering visual scenes and controlling what to describe are not substantial since the model generates descriptions of a larger set of objects and relations throughout multiple sentences.

In our experiments, the model back-propagates representations from cross-attention to each uni-modal stream. The cross-modal representations include both vision and language information combined. Hence, we expect that there is an effect that task-dependent information has on visual representations. We attribute this effect to the mechanism of cognitive control, which influences visual representations in deeper layers to be more beneficial for

cross-attention and, eventually, the task at hand (image captioning). At the same time, deeper representations are created from the lower-level knowledge of local object dependencies coming from earlier layers. In our experiments, we have observed that the number of attention links decreases with the increased depth of the model, and they also become stronger (more focused). We see parallels with perceptual selection in this effect: initially, the model constructs many different attention links and filters them layer after layer. Overall, we have provided a preliminary evidence that the representations in deeper layers are affected not only by visual information coming from visual input, but also by conceptual knowledge, that indirectly makes deeper representations to be more language-aware (Section 6.2.4.4). However, we leave a more detailed analysis for future work.

One important implication of our work is the effect that the structure of large-scale models has on the representations learned by different modules responsible for processing different modalities and their fusion. As all modules are trained end-to-end and optimised jointly, it becomes impossible to avoid information leaks from one modality to another. However, in multi-stream architecture, these effects can be seen and analysed in isolation for each modality. One benefit of using such representation is their indirect grounding in a different modality. For example, as we have revealed in this paper, visual representations alone contain perceptual knowledge about the scene, which is structured and partially organised by task-related language knowledge. Combined with the results in Ilinykh and Dobnik (2021), we argue that uni-modal representations resemble at least partial grounding in a different modality, which is just as good as a result of the cross-modal fusion of two modalities, that is often too complex to explain and utilise. We believe that an extensive set of experiments is required to examine if the training task and structure of the multi-stream transformers is the exact reason for such exciting blends of different modalities in a single modality's representation.

6.2.6. Conclusion

A large number of papers has focused on the analysis of representations captured by uni-modal architectures, e.g. BERT (Devlin, Chang, et al., 2019). This manuscript shifts the attention from uni-modal to multi-modal architectures and presents the analysis of visual representations learned by the two-stream image captioning transformer. We show that the visual knowledge is hierarchically structured as resembled by the self-attention weights of the visual stream. In particular, while earlier layers are better at learning the information about thematically related and visually close objects in the scene, deeper layers focus on objects that depict core entities on the image scale, capturing relations between them. We also demonstrate that the task affects the high-level knowledge in deeper layers, resulting in the artefacts of language found in visual information. We support our findings with several insights from the experiments in cognitive science. Overall, our extensive analysis touches upon fundamental questions on the effects of the model’s architecture and multi-modality on the model’s representations. We argue that representations of each modality can be enriched with important information from a different modality, which helps build more efficient and robust architectures. In future work, we are planning to test each of the three modules in the multi-stream transformer for several multi-modal tasks such as visual co-reference resolution and multi-modal human-object interaction.

6.3. Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer

6.3.1. Abstract

We explore how a *multi-modal transformer* trained for generation of longer image descriptions learns syntactic and semantic representations about entities and relations grounded in objects at the level of masked self-attention (text generation) and cross-modal attention (information fusion). We observe that cross-attention learns the visual grounding of noun phrases into objects and high-level semantic information about spatial relations, while text-to-text attention captures low-level syntactic knowledge between words. This concludes that language models in a multi-modal task learn different semantic information about objects and relations cross-modally and uni-modally (text-only). Our code is available here: <https://github.com/GU-CLASP/attention-as-grounding>.

6.3.2. Introduction

In this paper, we examine what kind of knowledge is encoded in the multi-modal transformer. Existing work has mostly looked at the knowledge captured in models that operate with **a single modality** (text). For instance, previous research has shown that the attention weights in large-scale models, e.g. BERT (Devlin, Chang, et al., 2019), implicitly encode knowledge of sentence structure (Raganato and Tiedemann, 2018; Ravishankar et al., 2021), part-of-speech tags, syntactic dependencies (Clark, Khandelwal, et al., 2019; Vig and Belinkov, 2019), subject-verb agreement between words (Goldberg, 2019), and even information about textual co-reference (Tenney, Das, et al., 2019). Only a few papers have inspected what is captured by **multi-modal architectures**.

Cao et al. (2020) demonstrate that the attention heads in image-and-text transformers effectively encode linguistic and cross-modal knowledge. Ilinykh and Dobnik (2021) provide the analysis of how language representations are indirectly affected by visual information in language-and-vision model.

Here we inspect what the model learns about two types of words in the multi-modal setting: (i) words denoting objects in the scene (e.g. “a red chair”), (ii) words depicting spatial relations between objects (e.g. “a chair *next to* the table”). While it is relatively simple to associate nouns with specific image regions, words describing relations are much harder to ground (Lu, Xiong, et al., 2017), possibly because visual representations are typically designed to capture objects without any explicit knowledge of relations. Secondly, grounded relations depend on knowledge from *both* vision and language modalities which contains information about the objects and their mode of interaction (*what*) as well as their physical location (*where*) (Ghanimifard and Dobnik, 2019). Ideally, each relation (and also other types of words) should be grounded in both modalities, but to a different degree.¹⁶ However, studies of language-and-vision models indicate that they are frequently biased towards one modality, most often to language (Goyal et al., 2017). Therefore, *the main research challenge* is to develop architectures that learn to utilise an appropriate ratio of visual and language knowledge for generation (or understanding) of each word in its context. Towards this goal we investigate grounding of different semantic types and answer the following questions:

- Q1:** Does attention across two modalities learn visually grounded semantics of nouns?
- Q2:** What syntactic knowledge is encoded in attention on text in the multi-modal set-up?

¹⁶Of course, in uni-modal word-embeddings the semantics of words are grounded in word-contexts only but such representations give us only common sense knowledge not linked to particular situations.

Model Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	WMD
CNN+LSTM+LSTM (Ilinykh and Dobnik, 2020)	25.10	13.88	8.11	4.61	11.30	26.38	7.61
Multi-Modal Transformer (this paper)	39.68	24.12	14.71	8.33	14.97	17.54	8.66

Table 6.1. Automatic evaluation of image paragraphs generated by two different model architectures.

Q3: What does cross-modal attention learn about grounded semantics of spatial relations?

We use a two-stream multi-modal transformer (Herdade et al., 2019), which first attends to each modality independently and then learns to attend cross-modally. This architecture uses rich relative geometry between objects, while many other two-stream models (Lu, Batra, et al., 2019; Tan and Bansal, 2019) simply use either coordinates of bounding boxes or their spatial location. We train the model for *image paragraph generation* (Ilinykh, Zarrieß, et al., 2019b; Krause et al., 2017), allowing examination of the knowledge of semantic types in extensive contexts. Our experiments show how language and vision are bridged in the multi-modal transformer. In addition, our work provides insights into how multi-modal representations are learned for different word types.

6.3.3. Experimental Set-Up

Model We train a multi-modal transformer for image paragraph generation. The model is based on the image captioning transformer proposed by Herdade et al. (2019)¹⁷. We use the object detector provided by Anderson, He, et al. (2018)¹⁸. This model comes pre-trained on object annotations from Visual Genome (Krishna et al., 2017). We extract features of N objects per image, resulting in the set $\mathbf{V} = \{v_1, \dots, v_N\}$ with $v_n \in \mathbb{R}^{1 \times D}$. We set $N = 36$ and $D = 2048$. The object extractor also provides us with labels (“table”) and

¹⁷https://github.com/yahoo/object_relation_transformer

¹⁸<https://github.com/peteanderson80/bottom-up-attention>

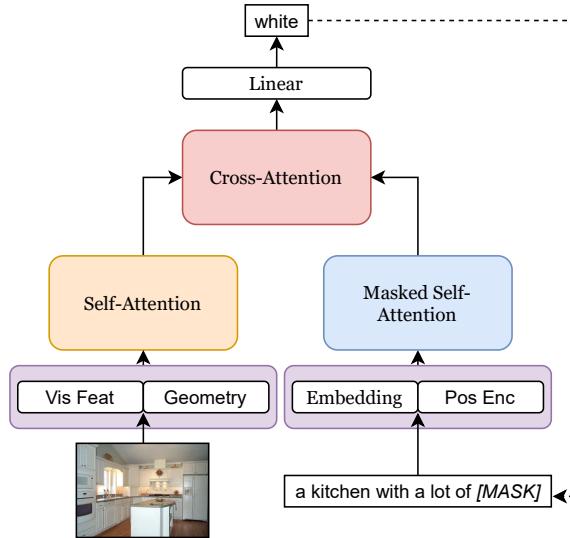


Figure 6.22. Multi-modal image description transformer. Every next generated word is concatenated with the previously generated words and passed to the model to output the next word prediction.

attributes (“round”) for the objects, which will be used in our experiments. Following Herdade et al. (2019), we also extract geometry information about each object $\mathbf{G} = \langle x, y, w, h \rangle$ (centre coordinates, width, height) and use it as an additional input along with visual features. Figure 6.22 describes the architecture of the model. In this model, each attention mechanism consists of six layers with eight attention heads in them. The *image encoder* (orange box) learns to combine visual and geometric features¹⁹ and passes them through the standard self-attention block, consisting of multi-head self-attention, feed-forward network, residual connections and layer-normalisation. Due to uni-directional nature of description generation, the *text decoder* (blue box) produces representation of the current token w_i , based on previously generated tokens (w_1, \dots, w_{i-1}) , while (w_{i+1}, \dots, w_W) are replaced with $[MASK]$. Finally, the *cross-attention* (red box) uses information from both textual and visual streams

¹⁹For more information on how image encoder employs both visual and geometric information, we refer the reader to the original implementation by Herdade et al. (2019).



Figure 6.23. **Ground truth description of the image:** It's a room with a bar on the side. There is a pink couch in the center. There's a coffee table in front of the couch. It has a light purple rug. There are three chairs at the bar.

Generated description of the image: This appears to be a bonus room that is red and white. There is a wooden table in the center of the room. There is a red couch. There is a large plant in the corner.

to output a probability of the next word in the sequence.

Dataset We train our model on Tell-Me-More (Ilinykh, Zarrieß, et al., 2019b), a dataset of natural multi-sentence descriptions of real-world images of rooms in the house setting (Zhou, Zhao, et al., 2017). The descriptions in this dataset are paragraphs produced by human describers in an image captioning task which are different from annotated relationships between object pairs in the Visual Genome (Krishna et al., 2017) which were examined in earlier work (Ghanimifard and Dobnik, 2019). Figure 6.23 shows an example of the ground truth text and generated paragraph. For training, we use train and extra splits, providing us with 4820 image-sequence pairs, while for validation and testing we use 441 and 441 pairs respectively. We use beam search to generate sequences with beam width $bw = 2$. The model is trained with standard cross-entropy loss. The best model's checkpoint is chosen based on the highest CIDEr score (Vedantam, Lawrence Zitnick, et al., 2015) for the

test set after training for 100 epochs. As Table 6.1 shows, our model achieves higher scores across most of the standard automatic metrics compared to the baseline architecture (CNN + LSTM + LSTM). Although our transformer performs slightly worse in terms of CIDEr score, note that different from previous work on multi-sentence image description generation (Chatterjee and Schwing, 2018; Ilinykh and Dobnik, 2020; Krause et al., 2017), we do not restrict the model to generate a specific number of sentences, instead stopping the generation when either the *END* token is encountered or the maximum number of words has been generated ($W = 150$). In addition, our dataset is much smaller than the Stanford image paragraph dataset (Krause et al., 2017), that the first model has been trained on.

6.3.4. Methods and Metrics

We extract the attention weights from both cross-modal attention and masked self-attention. Here, we could examine attention of the model while it is generating a new description or attention of the model receiving a ground truth description using teacher-forcing. Since our task is a validation task where we want to examine the behaviour of the model under fixed conditions we opted for ground truth descriptions. Using generated descriptions could produce identifiable attention patterns but the descriptions are not guaranteed to contain entities and relations that are in the image and we are interested in. If the model has approximated the training data well, then the unseen ground truth descriptions will not be far off from its predictions. Using ground truth descriptions that are not the model’s predictions imposes more uncertainty for the model and therefore harsher conditions for evaluation of attention patterns. Identifying interpretable attention patterns under these conditions therefore makes the conclusions stronger.

For every generated word w_i , the attention weight α per head h in each layer ℓ is extracted. In transformers the attention weights are computed as the scaled dot-product of the query matrix Q with all the keys in K followed by a

softmax operation. These weights are focusing on either previously generated words (masked self-attention MSA, Equation 6.15) or image objects (cross-attention CA, Equation 6.16).

$$\alpha_{\ell,h}(w_i \mid w_1, \dots, w_{i-1}) = \text{softmax}\left(\frac{Q_{MSA}K_{MSA}^T}{\sqrt{d_k}}\right) \quad (6.15)$$

$$\alpha_{\ell,h}(w_i \mid v_1, \dots, v_N) = \text{softmax}\left(\frac{Q_{CA}K_{CA}^T}{\sqrt{d_k}}\right) \quad (6.16)$$

We inspect how much attention is focused on specific parts of the input sequence when particular parts of the target sequence are generated. We refer to this measure as the **attention focus** or **attention proportion**. In our experiments, we calculate the proportion of total attention from a specific head that is focused on specific parts of the source sequence, e.g. previously generated words or image objects. Attention proportions are generally calculated as follows:

$$P_{\ell,h}(\alpha \mid S, T) = \frac{\sum_{u \in U} \sum_{i=1}^{|S|} \sum_{t=1}^{|T|} \alpha(s_i, S | t_j, T)}{\sum_{u \in U} \sum_{i=1}^{|S|} \sum_{t=1}^{|T|} \alpha(s_i, t_j, T)}, \quad (6.17)$$

where $P_{\ell,h}$ is the attention proportion for a specific head, S and T are the specific conditions imposed on the source and target sequences unique for every experiment (described below), U is the set of image descriptions sequences, t_j is the text span for either a noun phrase or relation from the target (generated) sequence T , s_i is the particular object or a text span from the source sequence S .

Conditions on P for Q1 For our experiments on visual grounding in cross-modal attention, T limits the target sequence to the text span of a noun phrase, while S defines the ground truth object that this noun phrase depicts. The attention proportion is calculated by computing the accumulated attention

weight from the words in the noun phrase towards the corresponding object and then divided by the overall attention on all objects attended when this noun phrase is generated. We use spaCy (Honnibal et al., 2020) to extract noun phrases from image paragraphs which might introduce some errors, see Appendix 6.3.7.1 for examples. We skip any phrases which contain at least one word from the list specified in Appendix 6.3.7.2. We keep determiners and adjectives in the noun phrases and any numerals if they occur. Some of the paragraphs might contain noun phrases that cannot be grounded in the bounding boxes in the image; either because the bounding boxes are not identified or because the noun phrases refer to abstract concepts. These phrases typically contain words such as “room”, “image” or “photo” and are generally placed at the beginning of the description (e.g., “the image is of a kitchen with ...”). In future experiments, we plan to investigate how the model grounds general descriptions of the scene (“the nursery room”).

Conditions on P for Q2 For the experiments on word-to-word attention, T is set to the generated word at the specific time-step t_j , while S accumulates attention on words of specified part-of-speech tags when the target word t_j is generated. Ilinykh and Dobnik (2021) show that masked self-attention on text is indirectly affected by vision in the multi-modal set-up. Nouns that often describe objects are attended to a greater extent than some other words of specific part-of-speech tag (e.g., verbs) even though this model has never seen the image directly. Interestingly, the same phenomenon is not observed in text-only models such as distilgpt-2: its attention is much more local, focusing on the words that surround the target word instead of attending to more distant nouns. This finding suggests that a multi-modal transformer can learn *semantic* differences between words of various part-of-speech tags not just their structural arrangement which would be their syntax. Therefore, we construct two sets of part-of-speech tags, which reflect semantic differences between words in terms of the possibility of their grounding. The first set

contains determiners, adjectives and nouns used in descriptions of objects, while the second set includes verbs and adpositions used in descriptions of relations between objects.

Conditions on P for Q3 To examine grounding of spatial relations, both S and T are determined based on the set of static spatial relations extracted from the texts. We extracted *target-relation-landmark* triplets from each description (there are likely to be multiple relations mentioned in a single image description sequence), based on the annotation schema described in Kolomiyets et al. (2013) and publicly available tool²⁰. We obtained 1015 relations of *region* type (“clothes on hangers”), 239 relations of *direction* type (“a gold chandelier above the table”), and 6 relations of *distance* type (“a large vase in the middle of the table”). Each of these relations consists of three spatial elements: a target (*a cup*), a landmark (*a table*) and a relation (*on*) in “a cup on the table”. Given that the word order describing relations is typically a *target-relation-landmark* sequence, the attention proportion for masked self-attention can be extracted only in following directions: *relation* → *target*, *landmark* → *relation*, *landmark* → *target*, and *landmark* → *target + relation*. For example, a possible T could restrict currently generated word to *relation* (typically expressed with adpositions), while S could limit the calculation of the attention focus to *target* (expressed as a noun phrase) in case of *relation* → *target* experiment.

6.3.5. Linking Nouns and Objects

To inspect attention heads for visual grounding, we require ground truth annotations of correct linking between image objects and noun phrases. We construct such links automatically using semantic similarity between noun phrases and object labels provided by the object feature extractor. First, we use spaCy (Honnibal et al., 2020) and extract noun phrases on different levels

²⁰<https://github.com/mmixgn/spr1-spacy>

of nesting. For example, a noun chunk “a window with white lace curtains” and the nested chunk “white lace curtains” are identified as two different noun phrases. Potentially, this design choice allows for more accurate linking between noun phrases focusing on different objects (“window” and “curtains”) and corresponding fine-grained object detections. In addition, noun phrases with specific details potentially disambiguate linking when multiple objects of the same type are in the image, e.g., several windows. As for object labels, for every detected object in every image, we take the predicted label and its attribute if the extractor’s confidence for this attribute is higher than 0.1. We determined this threshold manually allowing a lower degree of confidence to generate a sufficient number of adjectival attributes in order to disambiguate objects, e.g. “a brown chair” vs “a black chair”.

Noun phrases and object descriptions typically include multiple words. Therefore, we compute semantic similarity between phrases. We examine several methods for linking noun phrases and object descriptions and compare them against the small subset of image paragraphs with manually annotated linking. Specifically, we randomly sample ten image-text pairs, consisting of 196 detected noun phrases. Then, 158 noun phrases were manually linked with image objects by the first author. The subset of the remaining 38 noun phrases included pronouns and abstract descriptions, too ambiguous to be linked with the specific object in the scene. In addition, we found that some noun phrases describe either a non-detected object or were extracted by mistake. A fraction of noun phrases that were not linked with any object is shown in Appendix 6.3.7.1.

Table 6.2 shows the results of our search for the best linking method. We use GloVe embeddings (Pennington et al., 2014) to represent each word in a phrase and combine them by either element-wise multiplication (*GloVe Multiply*) or addition (*GloVe Add*), inspired by methods for phrase meaning representation (Mitchell and Lapata, 2008). The resulting vectors for a noun phrase and object description were compared based on cosine similarity. For

<i>Combination Method</i>	<i>Measure</i>	<i>mAP@K</i>	<i>Acc</i>
GloVe Multiply	cos	0.095	13.78
GloVe Add	cos	0.276	41.84
BERTScore	F_1	0.232	41.84
Sentence Transformer	cos	0.313	44.39

Table 6.2. Results of the search for the optimal method of linking noun phrases and object descriptions.

BERTScore we follow Zhang, Kishore, et al. (2020) and use contextual word embeddings (Devlin, Chang, et al., 2019) to represent every word. Words in a noun phrase and object description are then matched against each other by cosine similarity, and the F_1 score can be used to examine the similarity. Finally, for *Sentence Transformer* we represent each word with the embedding from Sentence Transformer (Reimers and Gurevych, 2019). This model fine-tunes BERT embeddings for numerous NLI tasks and applies a mean pooling operation to get the fixed-size vector representing embedding of a whole phrase. We report accuracy *Acc* against manual annotations of ten image-text pairs. We also compute mean average precision *mAP@K*, a metric that allows us to see whether a particular combination method generally rates relevant object descriptions more similar to a noun phrase:

$$AP@K = \sum_{k=1}^m P_k(R_k - R_{k-1}), \quad (6.18)$$

where P_k and R_k are the precision and recall at cut-off k , m is the number of noun phrases detected in an image paragraph. K is set to the number of objects (36) since we inspect the linking of noun phrases with the whole set of objects. The final *mAP@K* score is the mean of average precisions for noun phrases in descriptions of images. Our search results for the linking method demonstrate that using embeddings from Sentence Transformer and comparing them for cosine similarity performs the best in terms of both metrics. Interestingly, simply using BERT embeddings and match them for similarity (*BERTScore*) is not enough to achieve a high *mAP@K* score, and this method also performs

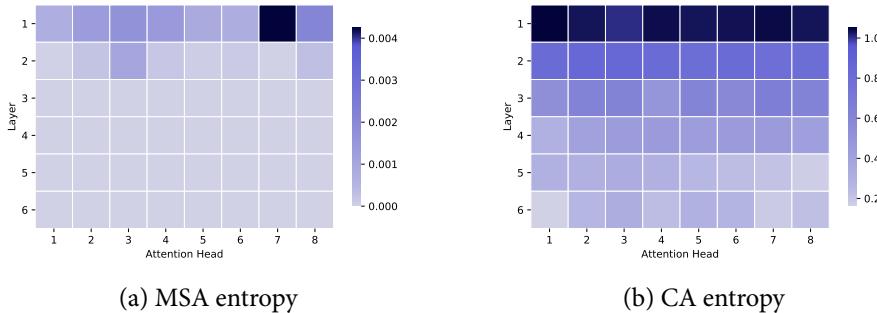


Figure 6.24. Normalised entropy of attention heads in different layers for masked-self attention (MSA) and cross-attention (CA). The darker the colour, the higher the entropy. All values were normalised by the maximum achievable entropy $-\log_2(O)$. Note that the range of values is different between the graphs.

worse than a simple addition of non-contextualised embeddings (*GloVe Add*). A more complex fusion of information from different words is required to represent a phrase. When examining attention heads for visual grounding of nouns and relations, we thus use the best performing linking method (*Sentence Transformer*). Noun phrases might describe a group of objects in the scene (“six chairs”), corresponding to multiple object detections (several chairs). Labels of such objects are often identical, which makes their cosine similarity scores also identical. Therefore, we link a noun phrase with multiple objects on top of the similarity ranking if they have the same cosine score. Otherwise, a noun phrase is linked with the object that is ranked the highest.

6.3.6. Experiments and Results

Attention Entropy We compute entropy of the attention weights in both modules for each attention head. Specifically, the entropy E of an attention head h in layer ℓ is defined as follows:

$$E_{\ell,h}(t_j) = - \sum_{i=1}^{|S|} \alpha(s_i, t_j) \log(\alpha(s_i, t_j)) \quad (6.19)$$

where s_i and t_j are specific source and target sequence items and α is the attention weight between them. As Figure 6.24 shows, the entropy pattern is similar across both attention modules. Attention heads have lower entropy in deeper layers, focusing more on specific parts of the source sequence. In contrast, surface layers scatter attention across many items (either objects or previously generated words). Intuitively, such progressive increase of attention focus from surface to deeper levels indicate that both modules first learn to generalise over low-level features, gradually moving to capture more specialised, high-level conceptual knowledge (Ullman, 1984). Here, a fair question to ask is *what kind of low-level and high-level knowledge do masked and cross-modal attention learn in different layers with different entropy?*

As Ghader and Monz (2017) show for the task of machine translation, lower attention entropy is mainly observed when looking at nouns and adjectives, while higher entropy is witnessed when attending to adpositions and verbs. This finding demonstrates that attending to nouns in purely textual syntactic dependencies is less complex than focusing on verbs. In the context of our task, adpositions and verbs would be used when generating spatial relations, while objects are described with nouns and adjectives. Learning nouns in a multi-modal setting implies their visual grounding, a more complex task that requires knowledge of the scene. Similarly, in general, understanding spatial relations is a much more sophisticated task for the multi-modal transformer. It requires higher-level semantic knowledge and identification of objects and relations, compared to simple attention on verbs and adpositions as part-of-speech tags in a uni-modal setting. It has also been shown that attention on highly complex phenomena (named entities) would happen in deeper layers of the model, while low-level constructs (determiners) are attended much earlier in the layers of both uni-modal (Vig and Belinkov, 2019), and multi-modal (Ilinykh and Dobnik, 2021) architectures. Therefore, in our experiments, we examine how attention heads in different layers of masked and cross-modal attention capture either **syntactic knowledge** (nouns and

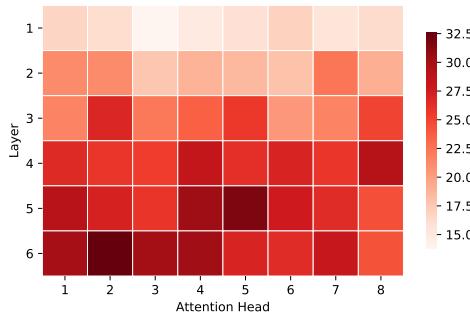


Figure 6.25. Attention proportions P on correct noun-object pairs (as determined by linking) for each attention head in the cross-modal attention. The darker the colour, the **bigger** the proportion. The proportions are averaged over the noun phrases in descriptions.

relation phrases as words) or **semantic information** (visually grounded nouns and spatial relations).

Visual Grounding in Cross-Attention Here we investigate whether the high focus of cross-attention heads in deeper layers can be attributed to their specialisation in visual grounding of nouns. Specifically, *based on the linking method*, we compute the proportion of attention that radiates from words in a noun phrase towards **corresponding objects** described by this noun phrase. Figure 6.25 shows the results.

We can see that attention heads in deeper layers concentrate on linking bounding boxes of detected objects with noun phrases that describe them when these phrases are generated. Specifically, while in the first layer, attention heads pay on average 16% of their attention to the linked objects, in the deeper layers, the average attention focus reaches 29%. The most activated head is the second head in the sixth layer, which places 33% of its attention on connecting noun phrases with the bounding boxes of objects linked with this phrase. These findings show that the model captures complex visually grounded semantics of nouns in deeper layers of cross-attention. In addition,

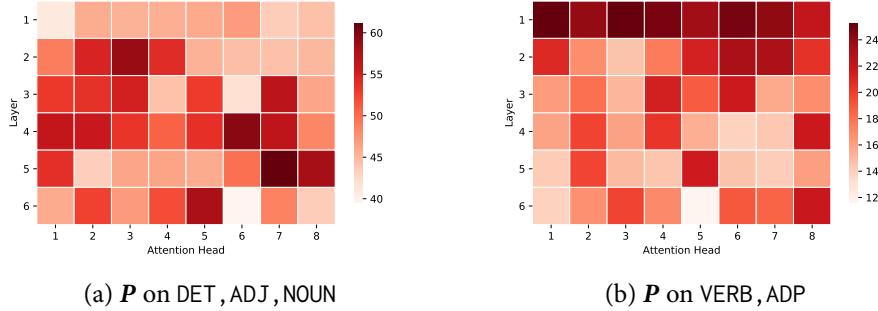


Figure 6.26. Attention proportions on words of specific part-of-speech tags for every head in the masked self-attention module. The proportions are averaged over the samples in the test set.

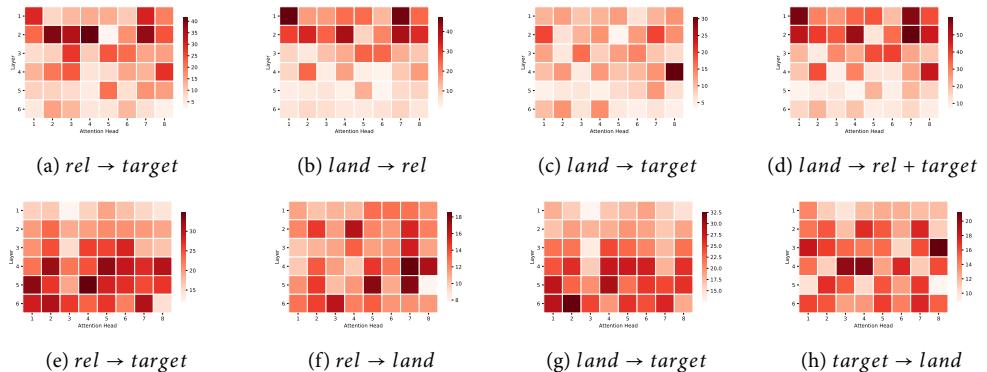


Figure 6.27. Heat-map visualisations of P for masked self-attention (**the top row**) and cross-attention (**the bottom row**) for different possible configurations of attention between words constituting spatial relations. All attention proportions are normalised by the number of spatial relations in the test set.

lower entropy observed in these layers (Figure 6.24b) also indicates that deeper heads are strongly focused and specialised in grounding of nouns.

Masked Self-Attention on Specific Part-of-Speech Tags Figure 6.26 demonstrates the attention focus on previously generated words of specific POS tags. We separate between tags which either describe objects (DET, ADJ, NOUN) or relations (VERB, ADP). Based on the heat-maps, we can see that previously generated determiners, adjectives and nouns are more attended in all layers except the first one, in which the focus is on relation part-of-speech tags. At the same time, according to Figure 6.24a, the attention in the first layer is more dispersed, which means that when attending to verbs and adpositions, attention is also looking at other words to a lesser degree, possibly such words which are involved in the action described by the verb. We calculated the Pearson correlation coefficient between both heat-maps in Figure 6.26. The test has shown a significant negative correlation ($r = -0.71, p = 1.7e - 08$), indicating that there is a clear separation in attention focus on two types of words in masked self-attention. Overall, text-to-text attention is able to capture local and non-grounded syntactic knowledge of objects and relations between them.

Masked Self-Attention on Spatial Relations Figures 6.27a–6.27d show the attention focus in masked self-attention for several possible directions between parts of the phrase describing spatial relation. For example, *rel* → *target* shows the attention on the noun phrase describing the target object when a phrase describing relation is generated. Note that in masked self-attention, we are not able to look into the future; thus, we cannot inspect attention on *rel* → *landmark* or *target* → *landmark*. The first important observation is a clear difference between attention on the word depicting the target object depending on where this attention is coming from. Numerous attention heads in the first layers focus on the target when relation is generated (Figure 6.27a),

while only a few heads are looking at the target when landmark is generated. According to Figure 6.27b, relation is more important for landmark since it is widely attended by many heads, compared to only a few heads in Figure 6.27c and only a single head (head 8, layer 4) being highly active. In addition, there are three attention heads in the second layer (2, 3, 4) in Figure 6.27a, which are also highly activated in Figure 6.26a. This might indicate that these heads do not simply look at the words depicting objects but specialise in such words, which are playing the part of the “target” object in spatial relations. Therefore, we can identify particular heads that learn knowledge of syntactic dependencies between words describing spatial relations in the textual encoder. Also, based on Figure 6.27b, we can see that the focus on relation phrases is mostly captured in surface layers, which supports our statement that the model first needs to learn general knowledge about existing relations in the scene, later starting to exploit it for better focus on correct target and landmark nouns.

Cross-Attention on Spatial Relations Figures 6.27e–6.27h show how much each head looks at the specific object that corresponds to a target or a landmark in spatial relations. Similar to our experiment on visual grounding, we linked every noun phrase describing either a target or a landmark with a bounding box of the detected object by computing semantic similarity between the noun phrase and the label of every object. Note that here we look at how words of *semantic categories* describing relations between objects are grounded in *visual representations* (objects) rather than other words, as in the case of the masked self-attention. One noticeable difference between the top and bottom rows in the Figure 6.27 is that the attention focus in the cross-modal part of the architecture is much more distributed across heads.

Given that, according to Figure 6.25, while multi-modal grounding of nouns into objects is clearly observed in the deeper parts of the model, grounding of relations in objects is much less interpretable. First, relations cannot

be straightforwardly linked to the visual features of objects in a scene. When grounding relations the system needs to rely on several sources of knowledge, both linguistic and visual, and here systems tend to rely on linguistic knowledge more than on visual information (Ghanimifard and Dobnik, 2019). Learning is further complicated by the fusion of information in cross-attention. For example, the model needs to simultaneously rely on the semantic information from the language representations and identify objects that are targets and landmarks in spatial relations. Therefore, cross-modal attention activates several attention heads when trying to learn about spatial relations, which require attention on multiple sources of knowledge.

Interestingly, as Figure 6.27f and Figure 6.27h show that attention on landmark in cross-attention is distributed across multiple layers. However, the first layer of $rel \rightarrow land$, which generally has the highest entropy (cf. Figure 6.24b), is more activated compared to the first layer of the $target \rightarrow land$ attention map. This shows that certain attention heads in the first layer specialise to identify landmarks from relations (Figure 6.27f), whereas there are less such heads that identify landmarks from targets (Figure 6.27h). This can be attributed to the fact that the model learns to confidently attend targets only in the deeper layers of the network because targets require much more complex inference. Landmarks are intuitively semantically closer to relations as in descriptions they are used together to identify targets. For example, Dobnik, Ghanimifard, et al. (2018) show that there is a strong asymmetry between knowledge about targets and landmarks. Landmarks are generally much easier to predict, and they contribute less to the perplexity of the model than targets. Intuitively, a speaker would like to describe the target, and they need to find a suitable contextually salient landmark, which then selects an appropriate relation and finally produce a full description including the target. Therefore, it might happen that the model first distributes its attention between heads in surface and deeper layers to identify landmarks in the context of particular relation, and then learns to strongly map this relation-landmark

context with the specific target in deeper layers. This idea is also supported by strongly localised and focused attention on the target object in deeper layers when either a relation or a landmark are generated (Figure 6.27g and Figure 6.27e).

Note the differences between attention patterns in Figure 6.27a and Figure 6.27e for the *relation* → *target* direction. Surface layers in masked self-attention, as we have shown, seem to learn local syntactic dependencies between words in the source input (text). This is different from the multi-modal scenario, where deeper layers are much more activated for visual and language inputs. This indicates that spatial relations are much more sophisticated in the language-and-vision context: they need to capture semantic dependencies between words and objects in the scene. Also, the complexity of information might be the reason why *rel* → *target* attention is much more scattered across many heads in deeper layers in cross-modal attention, compared to more focused attention in specific heads in surface layers for masked self-attention.

6.3.7. Conclusion

We have shown that the language model in a multi-modal task captures linguistic phenomena of different kind depending on the source knowledge (text or objects) and semantic type of the output words (noun phrases or spatial relations). Cross-modal attention visually grounds objects and, therefore, semantic dependencies in its deeper layers (addresses **Q1**). Text-only attention learns low-level linguistic phenomena, e.g. local syntactic dependencies (addresses **Q2**). This is also exemplified for target-relation-landmark descriptions which are attended in a sequential order that they appear in the text. We have also shown that there is a difference in a way objects and relations are grounded cross-modally and such grounding is particularly challenging for relations (addresses **Q3**). The grounding of landmarks depends on relations to a greater degree than on targets in both masked and cross-modal self-attentions.

This could be attributed to the auto-regressive nature of the image paragraph generation task. However, there are important differences in terms of activations across attention layers for different semantic pairs. Deeper heads in the cross-modal attention tend to be activated more than the surface heads which is the opposite tendency compared to masked self-attention. Overall, our work demonstrates that attention on vision and language captures considerably more diverse linguistic knowledge, *both syntactic and semantic* which is not *linearly aligned*, compared to uni-modal (language only) architectures.

One possible follow-up experiment is to use attention as input to the probing classifier and identify a specific knowledge encoded by the weights. However, the performance of the probing model does not tell us whether the original model utilises acquired knowledge since it is *detached* from the original architecture (Belinkov, 2022). Although attention is not necessarily an explanation (Jain and Wallace, 2019), inferring linguistic properties from attention weights does not require learning a new set of parameters. Other methods include fine-grained analysis of features preferred by specific neurons in the model architecture by examining their maximum activation values (Rethmeier et al., 2020). This method would identify the neurons that are active at each step of generation, but would not straightforwardly tell us how words and objects are linked together, which is clearly expressed in attention. Our results indicate that the way relations are grounded in a transformer model is not completely transparent. Future research should focus on examining the effect of different feature representations that are relevant for spatial relations (e.g., RGB-D and different models of geometry, common sense knowledge about objects' affordances) as well as the models that can be built around them. In another follow-up study we could examine grounding of relations in a different task, for example in vision-and-language navigation (Anderson, Wu, et al., 2018) which is rich with descriptions of relations between objects and compare whether the same observations also hold for those models.

6.3.7.1. Appendix A

Pronouns such as `it` and `his` were not linked with any object in the scene. Noun phrases depicting spatial descriptions or locations were also ignored, e.g. `the right`, `the background`, `the corner`. Some noun phrases are describing properties of objects in the scene (e.g., `color`, `the overall color of the room`) or positional arrangement (`a straight line in three paintings hang in a straight line`). Other noun phrases describe a general understanding of the image, and not a single bounding box could cover it (`a beachside hotel in a room that looks like inside a beachside hotel`). Some noun phrases were incorrect either due to an error made by spaCy or human producing the original description, e.g. `the walls floor sofa`.

6.3.7.2. Appendix B

When extracting noun phrases for the experiment on visual grounding we ignore all pronouns and spatial phrases found on this list: `right`, `a right`, `the right`, `left`, `a left`, `the left`, `top`, `the top`, `bottom`, `the bottom`, `back`, `the back`, `front`, `the front`, `far`, `the far`, `close`, `the close`, `side`, `each side`, `background`, `the background`, `foreground`, `the foreground`, `middle`, `the middle`, `corner`, `a corner`, `the corner`.

6.4. When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions

6.4.1. Abstract

Generating multi-sentence image descriptions is a challenging task, which requires a good model to produce coherent and accurate paragraphs, describing salient objects in the image. We argue that multiple sources of information are beneficial when describing visual scenes with long sequences. These include (i) perceptual information and (ii) semantic (language) information about how to describe what is in the image. We also compare the effects of using two different pooling mechanisms on either a single modality or their combination. We demonstrate that the model which utilises both visual and language inputs can be used to generate *accurate* and *diverse* paragraphs when combined with a particular pooling mechanism. The results of our automatic and human evaluation show that learning to embed semantic information along with visual stimuli into the paragraph generation model is not trivial, raising a variety of proposals for future experiments.

6.4.2. Introduction

The quality of automatically generated image captions (Bernardi et al., 2016) has been continuously improving as evaluated by a variety of metrics. These improvements include use of neural networks (Kiros et al., 2014; Vinyals et al., 2015), attention mechanisms (Lu, Xiong, et al., 2017; Xu, Ba, et al., 2015) and more fine-grained image features (Anderson, He, et al., 2018). More recently, a novel open-ended task of image paragraph generation has been proposed by Krause et al. (2017). This task requires the generation of multi-sentence image descriptions, which are highly informative, thus, include descriptions of a large variety of image objects, and attributes, which makes

them different from standard single sentence captions. In particular, a good paragraph generation model has to produce descriptive, detailed and coherent text passages, depicting salient parts in an image.

When humans describe images, especially over longer discourses, they take into account (at least) two sources of information that interact with each other: (i) perceptual information as expressed by visual features and (ii) cognitive reasoning that determines the communicative intent of the text and the use of language (Kelleher and Dobnik, 2019). Perceptual information mainly determines *what* to refer to while the reasoning mechanisms tell us *how* and *when* to refer to it. Both mechanisms interact: that a particular object is described at a particular point of discourse and with particular words depends not only on its perceptual salience but also whether that object should be referred to at that point of the story that the text is narrating which is its discourse salience. Compare for example: “two cows are standing in the field”, “there are trees in the field” and “a few of them are close to the trees”. The selection and the order of the relevant features are described by a cognitive mechanism of attention and memory (Dobnik and Kelleher, 2016; Lavie, Hirst, et al., 2004).

In this paper, we investigate the interplay between visual and textual information (reflecting background knowledge about the world and communicative intent) and their ability to generate natural linguistic discourses spanning over several sentences. Our primary research question is as follows: does using both visual and linguistic information improve *accuracy* and *diversity* of generated paragraphs? We experiment with several types of inputs to the paragraph generator: visual, language or both. We also investigate the effects of different kinds of information fusion between visual and textual information using either attention or max-pooling. We demonstrate that multimodal input paired with attention on these modalities benefits model’s ability to generate more diverse and accurate paragraphs.

We evaluate the accuracy and diversity of our paragraphs with both

automatic metrics and human judgements. We also argue that, as some previous work shows (van der Lee et al., 2019), n -gram-based metrics might be unreliable for quality evaluation of generated texts. The generated paragraph can be accurate as of the image, but because it does not match the ground truth, this would score low based on the automatic evaluation. To provide a different view on paragraph evaluation, we asked humans to judge the subset of generated paragraphs across several criteria, more specifically described in Section 6.4.4.4 and Appendix 6.4.6.1.

In language and vision literature, “diversity” of image descriptions has been mostly defined in terms of lexical diversity, word choice and n -gram based metrics (Devlin, Gupta, et al., 2015; Lindh et al., 2018; van Miltenburg, Elliott, et al., 2018; Vijayakumar et al., 2016). In these papers, the focus is on generating *a diverse set of independent, one-sentence captions*, with each describing image as a whole. Each of these captions might refer to identical objects due to the nature of the task (“describe an image with a single sentence”). Then, diversity is measured in terms of how different object descriptions are from one caption to another (e.g. a man can be described as a “person” or “human” in two different captions). However, as argued above, a good image paragraph model must also introduce diversity at the sentence level, describing *different scene objects* throughout the paragraph. Here, we define *paragraph diversity* with two essential conditions. First, a generative model must demonstrate the ability to use relevant words to describe objects without unnecessary repetitions (*word-level diversity*). Secondly, it must produce a set of sentences with relevant mentions of a variety of image objects in an appropriate order (*sentence-level diversity*).

Producing structured and ordered sets of sentences (e.g. *coherent paragraphs*) has been a topic of research in NLG community for a long time with both formal theories of coherence (Barzilay and Lapata, 2008; Grosz, Joshi, et al., 1995) and traditional rule-based model implementations (Deemter, 2016; Reiter and Dale, 2000). The coherence of generated text depends on several

NLG sub-tasks: *content determination (selection)*, the task of deciding which parts of the source information should be included in the output description, and *text structuring (micro-planning)*, the task of ordering selected information (Gatt and Krahmer, 2017). We believe that the hierarchical structure of our models reflects the nature of these tasks. First, the model attends to the image objects and defines both their salience and order of mention and then it starts to realise them linguistically, first as paragraph visual-textual topics and then as individual sentences within paragraphs.

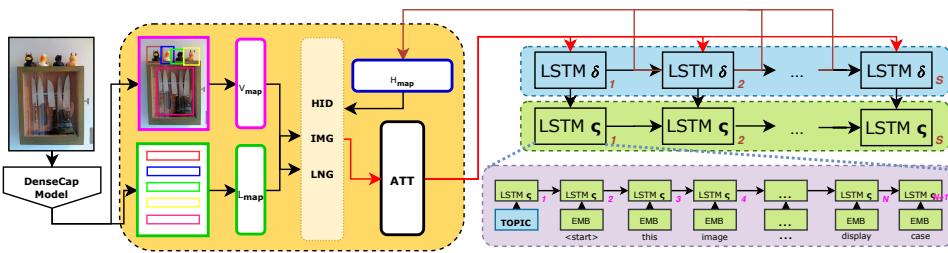


Figure 6.28. Multimodal paragraph generator architecture. The orange area on the left is the learned space where two modalities are attended to (vision in purple, language in green). The mapped features are concatenated together and passed to the attention mechanism, that outputs a vector which is used as an input to the discourse LSTM (in blue, marked with δ). The attention module also uses the last hidden state of the discourse LSTM at each timestamp. The sentence LSTM (in green, marked with c) is given the sentence topic and word embeddings. Due to limited space, we omit the linear layer and the softmax layer which are used to predict the next word from the output of the sentence LSTM.

6.4.3. Approach

Overview For our experiments we implement and adapt the hierarchical image paragraph model by Krause et al. (2017).²¹ We deliberately chose to reimplement an existing model to study the effects of using different modalities (visual or language). However, through our implementation and extensions, we propose several new models based on the original model in (Krause et al.,

²¹The authors have not publicly released the code of their model and hence the model implementation is based on our interpretation of their paper.

2017). To prepare input features, we utilise the pre-trained model for dense captioning (Johnson et al., 2016) in two ways. First, we use it to extract convolutional features of identified image regions. We also use its hidden states from the RNN layer as language features. In the original model, these states are used to generate region descriptions; therefore, these vectors represent semantic information about objects. We construct a *multi-modal space*, in which we learn mappings from both text and vision features. Lastly, we concatenate both modalities and attend to them to form a multi-modal vector, which is used as an input to the paragraph generator. Our paragraph generator consists of two components: discourse-level and sentence-level LSTMs (Hochreiter and Schmidhuber, 1997). First, the discourse-level LSTM learns the topic of each sentence from the multi-modal representation, capturing information flow between sentences. Second, each of the topics is used by sentence-level LSTM to generate an actual sentence. Finally, all generated sentences per image are concatenated to form a final paragraph. An overview of our model and a more detailed description is shown in Fig. 6.28. Our model is different from the model by Krause et al. (2017) in the following ways: (i) we use either max-pooling or attention in our models, (ii) we do not learn to predict the end of the paragraph, but generate the same number of sentences as we find in ground-truth paragraph per each image, (iii) we use semantic information about objects in the visual scene. The focus of our work is not to improve on the results of Krause et al. (2017) but to investigate the effects of different multi-modal fusion on the accuracy or the diversity of paragraph descriptions.

6.4.3.1. Input Features

Visual Features We use DenseCap region detector (Johnson et al., 2016)²² to identify salient image regions and extract their convolutional features. First, a resized image is passed through the VGG-16 network (Simonyan and Zisserman, 2015) to output a feature map of the image. A region proposal network

²² Available at: <https://github.com/jcjohnson/densecap>

is conditioned on the feature map to identify the set of salient image regions which are then mapped back onto the feature map to produce corresponding map regions. Each of these map regions is then fed to the two-layer perceptron which outputs a set of the final region features $\{v_1, \dots, v_M\}$, where $v_m \in \mathbb{R}^{1 \times D}$ with $M = 50$ and $D = 4096$. This matrix $V \in \mathbb{R}^{M \times D}$ provides us with fine-grained image representation at the object level. We use this representation as features of visual modality.

Language Features In the dense captioning task, a single layer LSTM is conditioned on region features to produce descriptions of these regions in natural language. We propose to utilise its outputs as language features, using them as additional semantic background information about detected objects. Specifically, we condition a pre-trained LSTM on region features to output a set $Y = \{y_1, \dots, y_M\}$ with $y_m \in \mathbb{R}^{1 \times T \times H}$, where $T = 15$ and $H = 512$. We condense each vector over the second dimension T , which determines the maximum number of words in each description. We achieve this by summing all elements across this dimension and dividing the result by the actual length of the corresponding region description, which we generate from Y . The final matrix $L \in \mathbb{R}^{M \times H}$, contains language representations of M detected regions.

Multimodal Features First, we learn two different mappings, using V_{map} for vision and L_{map} for language. These linear projections learn to embed modality-specific information into the attention space. Then, we concatenate these mappings to form the multimodal vector f , which is then combined with the mapping from the hidden state. We have experimented with fusing two attended modalities into a single vector via an additional linear layer but observed no improvement. We also tried to use modality-dependent attention (*early attention*) as such setting has shown to produce good joint representation for the task of multimodal machine translation (Caglayan, Barrault, et al., 2016; Caglayan, Madhyastha, et al., 2019), which is very similar

to image captioning in its nature. However, this set-up provided us with worse scores of automatic metrics. Therefore, here we use *late attention*: attending to the visual and textual features when they are already concatenated.

As shown in Eq. 6.20, at each timestamp t we concatenate mapped features from both modalities to output the multimodal vector $mult_t$, where $t \in \{1, \dots, S\}$ and S is the maximum number of sentences to generate. We use δ to refer to the discourse LSTM and ς when referring to the sentence LSTM. Concatenation, the logistic sigmoid function and element-wise multiplication are indicated with \oplus , σ and \odot respectively. We set S depending on the number of sentences in the ground-truth paragraph with the maximum $S = 6$. Then, as Eq. 6.21 indicates, we generate attention weights for our multimodal vector $mult_t$. We use additive (concat) attention mechanism and concatenate multimodal representation with the previous hidden state of the discourse LSTM. Finally, as in Eq. 6.22, we obtain a weighted multimodal vector $f \in \mathbb{R}^{1 \times H}$, which encapsulates and merges salient information from attended visual and textual modalities.

$$mult_t = [W_m^V V_t \oplus W_m^L L_t] \quad (6.20)$$

$$\alpha_t^{mult} = softmax(W_a^L \tanh(mult_t \oplus W_h h_{t-1}^\delta)) \quad (6.21)$$

$$f_t = [\alpha_t^{mult} \odot mult_t] \quad (6.22)$$

6.4.3.2. Discourse LSTM

Our discourse-level LSTM is responsible for modelling multi-modal topics of each of the individual sentences in the paragraph. At each timestamp, it is conditioned on the weighted multimodal vector f_t , and its output is a set of hidden states $\{h_1, \dots, h_S\}$, where each state is used as an input to the sentence-level LSTM. In its nature, the discourse LSTM has to simultaneously

complete at least two tasks: produce a topic with a relevant combination of visual and linguistic information for each sentence, while preserving some type of *ordering* between the topics. Such topic ordering is essential for keeping a natural transition between sentences (discourse items) in the paragraph (discourse). We expect attention on the combination of two modalities to assist the discourse LSTM in its multiple objectives since attention weights specific parts of the input as more relevant for a particular sentence. We expect that this allows discourse LSTM to learn better sentence representations and sentence order.

Similar to Xu, Ba, et al. (2015), we also learn a gating scalar β and apply it to f_t :

$$\beta = \sigma(W_b h_{t-1}^\delta), \quad (6.23)$$

where W_b is a learnable model parameter. Thus, the input to discourse LSTM is computed as follows:

$$f_t^\delta = \beta \odot f_t \quad (6.24)$$

6.4.3.3. Sentence LSTM

Our sentence-level LSTM is a single-layer LSTM that generates individual sentences in the paragraph. We run the sentence LSTM S times. Each time we use a concatenation of the corresponding hidden state of the discourse LSTM with the learned embeddings of the words in the target sentence y_s as its input:

$$x_s^\varsigma = [h_s^\delta \oplus E y_s] \quad (6.25)$$

Our word embedding matrix $E \in \mathbb{R}^{K \times H}$ is learned from scratch, K is the vocabulary size. This is different from (Krause et al., 2017), who use word embeddings and LSTM weights from the pre-trained DenseCap model. We

have also experimented with transferring DenseCap weights and embeddings into our model but observed no significant improvement.

At each timestamp t , our sentence LSTM is unrolled $N + 1$ times, where N is the number of words to generate. At each step, its hidden state is used to predict a probability distribution over the words in the vocabulary. We set $N = 50$. The final set of sentences is concatenated together to form a paragraph.

6.4.3.4. Learning Objective

We train our model end-to-end with image-paragraph pairs (x, y) from the training data. Our training loss is a simple cross-entropy loss on the sentence level:

$$\text{loss}^c(x, y) = - \sum_{i=1}^S \sum_{j=1}^{M_i} \log(p_{j,s}) \quad (6.26)$$

where $p_{j,s}$ is the softmax probability of the j^{th} word in the i^{th} sentence given all previously generated words for the current sentence $y_{1:j-1,i}$. For the first sentence, the hidden states of both LSTMs are initialised with zeros. For every subsequent sentence, both LSTMs use the last hidden states generated for the previous sentence for each respective layer. During training, we use teacher forcing and feed ground-truth words as target words at each timestamp. We use Adam (Kingma and Ba, 2015) as an optimiser and choose the best model based on the validation loss (early stopping). For decoding we use beam search (Freitag and Al-Onaizan, 2017) with beam width $B = 2$ (we tested several values for the beam width $B \in \{2, 4, 6, 8, 10\}$). We leave the investigation of the effects of using different decoding strategies such as nucleus sampling (Holtzman et al., 2020) or various techniques for controlling decoding (length penalty, n-gram repetition penalty (Klein et al., 2017; Paulus et al., 2017)) for future work.

Model Input	Type	WMD	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG	+MAX	7.48	25.66	11.20	24.51	13.67	7.96	4.51
LNG	+MAX	7.19	22.27	10.81	23.20	12.69	7.34	4.19
IMG+LNG	+MAX	7.61	26.38	11.30	25.10	13.88	8.11	4.61
IMG	+ATT	7.47	26.01	11.26	24.88	13.99	8.13	4.67
LNG	+ATT	7.20	22.11	10.82	23.20	12.55	7.16	3.97
IMG+LNG	+ATT	7.54	26.04	11.28	24.96	13.82	8.04	4.60

Table 6.3. Automatic evaluation results. Models are separated based on the input features (one modality / multi-modal) and type of the mechanism used to compactly describe content of the image (max-pooling / attention). Best scores for both +MAX and +ATT modes are shown in bold. The colour intensity indicates how good the score is compared to the other models' scores.

Model Input	Type	mBLEU	self-CIDEr
IMG	+MAX	50.63	76.43
LNG	+MAX	52.24	75.59
IMG+LNG	+MAX	52.09	76.46
IMG	+ATT	51.82	75.51
LNG	+ATT	50.93	76.41
IMG+LNG	+ATT	47.42	78.39
GT	-	18.84	96.51

Table 6.4. Automatic paragraph diversity evaluation. mBLEU stands for the average score between all self-BLEU scores for n -grams (1, 2, 3, 4). Self-CIDEr stands for the average score of the LSA-based diversity metric. We also include ground-truth scores calculated from the test set (GT, coloured in blue). Best models are shown in bold. All scores are multiplied by 100 for better interpretability.

6.4.4. Experiments and Evaluation

6.4.4.1. Models

We describe six configurations of our model, which we train, validate and test on the released Stanford paragraph dataset splits (14,575, 2,487, 2,489 for training, validation and testing respectively) (Krause et al., 2017). Our models are described as follows: the **IMG** model is conditioned only on the mapped visual features, while the **LNG** model only uses the mapped semantic information to generate paragraphs. The **IMG+NLG** is conditioned on both

mapped visual and semantic information. All models with **+ATT** use late attention on either uni-modal or multi-modal features. We also test another configuration of the models with max-pooling of input features across M regions, represented by mapping from either language features $x = W_m^L L_t$ or visual features $x = W_m^V V_t$:

$$x_s^\zeta = \max_{i=1}^M(x) \quad (6.27)$$

In the **IMG+LNG** model we apply max-pooling on both modalities and concatenate them into a single vector:

$$x_s^\zeta = [\max_{i=1}^M(W_m^L L_t) \oplus \max_{i=1}^M(W_m^V V_t)] \quad (6.28)$$

6.4.4.2. Metrics

Typically, a variety of n-gram based automatic metrics is used to measure the correctness/accuracy of image captions. We evaluate our models with the following metrics: CIDEr (Vedantam, Lawrence Zitnick, et al., 2015), METEOR (Denkowski and Lavie, 2014), BLEU-{1, 2, 3, 4} (Papineni et al., 2002), and Word Mover’s Distance (Kilickaya et al., 2017; Kusner et al., 2015). We also measure lexical diversity of sentences within the generated paragraphs. For this we report self-BLEU (Zhu et al., 2018) which is sometimes referred to as mBLEU (Shetty et al., 2017). Estimating lexical diversity is important for paragraph generation as their sentences should be neither too similar nor too different from each other. We calculate self-BLEU as follows: we split each generated paragraph into sentences and use one sentence as a hypothesis and the other sentences as references. A lower score indicates more diversity, e.g. fewer n -gram matches between compared sentences. We also calculate the diversity metric introduced by Wang and Chan (2019). This metric applies Latent Semantic Analysis (Deerwester et al., 1990) to the weighted n-gram feature representations (CIDEr values between unique pairs of sentences)

and identifies the number of topics among sentences. Compared to self-BLEU, which measures n-gram overlap, LSA combined with CIDEr-based kernel metric measures semantic differences between sentences as well. More identified topics in paragraph sentences indicate a higher level of diversity. However, this intrinsic metric does not evaluate if the paragraph demonstrates discourse coherence in terms of how these topics are introduced and the quality of the generated sentences and their sequences (Section 6.4.2).

6.4.4.3. Results

As the results in Table 6.3 demonstrate, models which utilise both semantic and visual information (any **IMG+LNG** configuration) outperform their single modality variants in both attention and max-pooling settings. When using max-pooling, **IMG+LNG** model improves on CIDEr by 0.72 and METEOR by 0.10. Also, two-modal architecture is slightly lexically more diverse from the ground truth paragraphs, according to the WMD scores. This result comes at no decrease in other metrics, concerned with lexical accuracy.

When replacing max-pooling with late attention, we observe that the **IMG** model reaches the highest scores in BLEU-{2, 3, 4}, while finishing second in all other metrics. However, **IMG+LNG** model does not seem to benefit from the attention that much, reaching lower scores in comparison to its version with max-pooling. Interestingly, semantic information is beneficial to WMD, CIDEr and METEOR, which also take into account the syntactic structure of the sentences.

Table 6.4 contains the scores of the lexical diversity metrics. The best (i.e. the lowest) mBLEU scores are achieved by models which use either a visual modality with max-pooling (**IMG+MAX**) or both modalities with attention (**IMG+LNG+ATT**). The best self-CIDEr scores are achieved by both bi-modal architectures. In addition, **IMG+LNG+ATT** strongly outperforms all other models in both lexical diversity metrics: mBLEU is reduced by 3.21%

indicating a smaller n -gram overlap between paragraph sentences, while self-CIDEr increases by 1.93% demonstrating that attention in the model which uses multimodal features helps to generate a more diverse set of sentences in terms of topicality.

We include two examples of generated texts by humans and our models. As Figure 6.29a demonstrates, the **IMG+LNG+ATT** model can generate less redundant/repetitive descriptions compared to the **IMG+LNG+MAX** model. Figure 6.29b demonstrates a case where **IMG+LNG+ATT** generated a paragraph which seems correct but different from the human-generated text (e.g. ‘human’ instead of a ‘dog’). However, to a human eye it is not entirely clear whether the surfer is a human or a dog. Neither of our models was able to see a ‘dog’ in this image. Similar challenging examples include cases where different human describers have a different take on what the image is about or what is the focus of the image.

Therefore, the intrinsic metrics might not be the best indicator for identifying clear differences in diversity and accuracy of the generated texts. In addition, such diversity metrics as mBLEU under-represent the diversity, being unable to take into account semantic differences between sentences. Therefore, we conduct a human evaluation experiment to achieve a better understanding of which input features and which pooling mechanism assists in the generation of both accurate and diverse paragraphs.

6.4.4.4. Human Evaluation

In the human evaluation task we are interested in the following properties of generated paragraphs covering both accuracy and diversity aspects: word choice, object salience, sentence structure and paragraph coherence. We randomly chose 10% of the images from our test set, resulting in 250 images. For each of these images, we gathered seven paragraphs (six from the models and one from the test set). We presented workers with the instructions shown in Appendix 6.4.6.1. To ensure quality and variety of workers’ judgements, we



(a) **HUMAN**: *There are several cars parked along a street. There are many trees in a field in front of the street. There are small blue parking meters on the sidewalk next to the street.*

(b) **HUMAN**: *A large splash is in front of a wave in the water. There is a large white and black surf board in the water. There is a black dog that is riding on top of the surf board.*

IMG+MAX : *There are several cars parked on the road. There are cars parked on the street. There are trees behind the street.*

IMG+MAX : *A man is riding a wave. He is holding a surfboard. The man is wearing a black wet suit.*

LNG+MAX : *There are several cars on the street. There are trees on the street. There are trees on the street.*

LNG+MAX : *A person is surfing in the water. The surfboard is black and white. The surfboard is black and white.*

IMG+LNG+MAX : *There are several cars on the street. There are two cars on the street. There are cars parked on the sidewalk.*

IMG+LNG+MAX : *A man is standing on a surfboard. The surfboard is black. The man is wearing black shorts.*

IMG+ATT : *There are several cars parked on the road. There are two cars parked on the road. There are two cars parked on the road.*

IMG+ATT : *A man is standing on a surfboard. The surfboard is black and white. The man has black hair.*

LNG+ATT : *There are several signs on the street. There are signs on the street. The pole is white.*

LNG+ATT : *A person is standing in the water. The person is wearing a black suit. The person is holding a black surfboard.*

IMG+LNG+ATT : *There is a parking meter holding a black surfboard on a sidewalk. There are cars next to the street. There is a parking lot next to the street.*

IMG+LNG+ATT : *A person is surfing in the ocean. She is wearing a black wet suit. She is holding a white surfboard.*

Figure 6.29. Two example images with generated paragraphs from our models (incl. ground truth descriptions).

presented our tasks only to the Master workers (those with the high reputation and task acceptance rate) and controlled for the number of tasks a single worker is able to submit (we set it to 30). We paid 0.15\$ per task to a single worker. Finally, we obtained judgements from 154 unique Master workers for 1,750 image paragraphs overall. For each judgement criteria, we took the average score across all models; the results are shown in Table 6.5.

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	59.57	37.87	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	34.20	38.72	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	37.06	46.78	72.95	40.88	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

Table 6.5. Human evaluation results. WC, OS, SS, PC stand for word choice, object salience, sentence structure and paragraph coherence. Each value in the table is the average of all scores for the corresponding criterion. The mean values per each model and type of pooling mechanism are coloured in light cyan.

As shown by human evaluation, looking at the overall mean, the multi-modal information does help the generation of better paragraphs when using max-pooling. The **IMG+LNG** model with max-pooling might be a beneficial choice (scores first in two criteria out of four) in terms of word choice and identification of salient objects. The performance of the **IMG+LNG** model with max-pooling is close to the performance of the **IMG** model while the performance of the **LNG** model is slightly lower. Overall, attention is judged as more advantageous in general than max pooling, having higher mean scores across all criteria compared to the mean scores of max-pooling models. However, here the **IMG+LNG** model is outperformed by both uni-modal models. The **LNG** model which utilises semantic information and uses attention is judged as the best configuration by humans, which is in line with some previous work that reports strong bias on the semantic information (Agrawal, Batra,

Parikh, and Kembhavi, 2018). Note that while its performance is close to the **IMG** model in terms of word choice and object salience, the improvement of the **LNG** model is much more expressed in terms of sentence structure and paragraph coherence, categories where one would expect that semantic information matters most. Interestingly, max-pooling does not seem to have the same effect on utilisation of semantic information: the **LNG+MAX** model achieves the lowest scores. A possible explanation for this is that when using max-pooling, the same semantic information is chosen for every sentence topic. At the same time, attention learns to select different semantic information for a sequence of topics. This appears to affect semantic features more than visual features. Note that humans mostly judge models that incorporate linguistic information as the best ones for the word choice criterion. This supports the idea that utilising semantic information reduces redundancy in terms of the number of repeated words in the generated paragraph.

Overall, the results indicate that both visual and semantic information are beneficial for the generated paragraphs as they affect different evaluation categories differently. The main challenge lies in information fusion of visual and semantic information in the model with attention. We believe that these results suggest the following future experiments: (i) detailed investigation of early vs. late attention (when to fuse two modalities and how), (ii) as van Miltenburg, Elliott, et al. (2017) argue, more control over human evaluation can provide us with better, more precise human judgements, (iii) training with other decoding strategies such as top- k sampling or nucleus sampling (Holtzman et al., 2020).

6.4.5. Related Work

Neural image paragraph captioning The task of generating image paragraphs has been introduced in (Krause et al., 2017) along with the dataset of image-paragraph pairs. The authors hierarchically construct their model: sentence RNN is conditioned on visual features to output sentence topics.

Then, each of these topics is used by another RNN to generate actual sentences. Our models are based on this hierarchical model. However, we substantially change its structure and also remove the end of paragraph prediction.

Liang et al. (2017) also use the hierarchical network, but with an adversarial discriminator, that forces model to generate realistic paragraphs with smooth transitions between sentences. Chatterjee and Schwing (2018) also address cross-sentence topic consistency by modelling the global coherence vector, conditioned on all sentence topics. Different from these approaches, Melas-Kyriazi et al. (2018) employ self-critical training technique (Rennie et al., 2017) to directly optimise a target evaluation metric for image paragraph generation. Lastly, Wang, Pan, et al. (2019) use convolutional auto-encoder for topic modelling based on region-level image features. They demonstrate that extracted topics are more representative and contain information relevant to sentence generation. We also model topic representations, but we use additional semantic representations of image objects as part of the input to our topic generator. Lin, Kong, et al. (2015) has proposed a non-neural approach to generate texts describing images. However, this approach depends on multiple components: visual scene parsing, generative grammar for learning from training descriptions, and an algorithm, which analyses scene graphs and extracts semantic trees to learn about dependencies across sentences.

Language representation for image captioning Several existing models for image captioning are conditioned on both visual and background information. You et al. (2016) detect visual concepts found in the scene (objects, attributes) and extract top-down visual features. Both of these modalities are then fed to the RNN-based caption generator. Attention is applied on detected concepts to inform the generator about how relevant a particular concept is at each timestamp. Our approach does not use any attribute detectors to identify objects in the scene. Instead, we use the output of another pre-trained model for the task of dense captioning. Lu, Xiong, et al. (2017) emphasise that image

is not always useful in generating some function words (“of”, “the”). They introduce adaptive attention, which determines when to look at the image and when it is more important to use the language model to generate the next word. In their work, the attention vector is a mixture of visual features and visual sentinel, a vector obtained through the additional gate function on decoder memory state. Our model is guided by their approach: we are interested in deciding which type of information is more relevant at a particular timestamp, but we also look at how *merging* two modalities into a single representation performs and how it affects attention of the model. Closest to our work is the work by Liang et al. (2017), who apply attention to region description representation and use it to assist recurrent word generation in producing sentences in a paragraph. Similar to our approach, they also supply their model with embeddings of local phrases used to describe image objects. However, they use textual phrases directly, while we are using hidden representations from the model trained to generate such phrases (Johnson et al., 2016). Also, our approach explores a different application of semantic information encoded in language: we use phrase representations to define sentence topics to choose from (topic selection) rather than directly guide the generation of words (micro-planning).

6.4.6. Conclusion

In this paper, we addressed the problem of generating both accurate and diverse image paragraphs. We demonstrated that utilising both visual and linguistic information might benefit the quality of generated texts depending on the pooling mechanism that is used. We showed that intrinsic evaluation metrics are insufficient for evaluation of paragraphs as they focus on lexical choice and do not capture human level of judgement: **LNG+ATT** is judged as the best model in human evaluation, while it is not among the leaders according to the automatic evaluation. We believe that our work is a good starting point for further investigation of the ways multiple sources of information

about the world can be merged for learning generation of high-quality multi-sentence stories, describing real-world visual scenes. In our future work we also intend to test how our models can generate task-dependent paragraphs. For this task we will use the dataset of image description sequences (Ilinykh, Zarrieß, et al., 2019b) which consists of paragraphs collected in a task-based setting to train our models. In contrast, in the Stanford dataset humans were not given a specific task when describing images. We believe that generation from more context-dependent and structured descriptions can open up new perspectives for the research on image paragraphs.

6.4.6.1. Human Evaluation: AMT Instructions

Short Summary: You are going to be shown an image and several sentences describing the image. Below you will see statements that relate to the image descriptions. Please rate each of these statements by moving the slider along the scale where 0% stands for ‘I do not agree’, 100% stands for ‘I fully agree’.

Detailed Instructions:

In general, you are required to judge image descriptions based on the following:

- choice of words: does the text correctly describe objects and events in the scene and with the right detail?
- relevance: does the text describe relevant objects and events in the scene?
- sentence structure: do the sentences have a good and grammatical structure?
- coherence: does the text progresses in a natural way forming a narrative?

You can enter any feedback you have for us, for example if some questions were not easy to answer, in the corresponding feedback field (right after the survey).



DESCRIPTION: there are two cows standing in the field. there are trees behind them.

How well do you agree with the following statements?

1. The description contains words that correctly refer to the objects and events in the image

2. The description is referring to the relevant/important parts of the image.

3. The sentences have a correct structure and are grammatical.

4. The sentences are well-connected and form a single story.

Write your feedback in the field below if you have any (not necessary).

6.5. Look and Answer the Question: On the Role of Vision in Embodied Question Answering

6.5.1. Abstract

We focus on the Embodied Question Answering (EQA) task, the dataset and the models (Das, Datta, et al., 2018). In particular, we examine the effects of vision perturbation at different levels by providing the model with either incongruent, black or random noise images. We observe that the model is still able to learn from general visual patterns, suggesting that they capture some common sense reasoning about the visual world. We argue that a better set of data and models are required to achieve better performance in predicting (generating) correct answers. The code is available here: <https://github.com/GU-CLASP/embodied-qa>.

6.5.2. Introduction

When language generation models are employed in real-world scenarios, they need to correctly perceive the environment, understand physics between objects and reason about the events in order to produce logical and correct descriptions (Lake et al., 2017). In order to study and ultimately construct such models, several language-and-vision tasks were developed including Visual Question Answering (VQA) (Antol et al., 2015; Gordon, Kembhavi, et al., 2018) and Visual Dialogue (Das, Kottur, et al., 2017). The advantage of such models is their ability to process visual information *jointly* with language. However, several papers following have found that **vision is often dismissed** by the model and language is much more attended to. Attempts were made to influence this bias on *the dataset side* and make the contributions of both modalities more equal. For example, Goyal et al. (2017) show that coupling questions in a VQA dataset with complementary images, which lead to different responses, makes the model learn more from vision and less



Figure 6.30. Example of successive removal of context, content and structure. For each removal type, we show the first frame from the set of frames that the model takes to answer the question “What color is the stove in the kitchen?”. From left to right: **original** (nothing is removed), **shuffled** (structure and content are present, but context is incorrect), **blind** (no content and context, but structure), **random** (most disturbed representation).

from language biases. A different way of tackling the language bias in VQA datasets is to augment them with a larger variety of different question types, generated with either a template-based method or neural networks (Kafle and Kanan, 2017). Caglayan, Madhyastha, et al. (2019) note that there exists a dataset structure bias realised through short and repetitive texts, which in principle could inhibit gains from vision. On the other hand, many papers have proposed *models* capable of better fusion between vision and language. Zheng et al. (2020) introduce a method to learn better alignment between language and vision spaces based on reasoning over entities in texts and objects in images for the VQA task. Work on multi-modal machine translation looked at the model performance when images are replaced with incongruent scenes (Elliott, 2018) or leveraging the importance of vision modality by testing different fusion techniques (Raunak et al., 2019).

VQA models cannot be directly applied in the real world scenario due to challenges that require direct interaction of the model with the environment. Therefore the task of Embodied Question Answering (EQA) has been proposed by Das, Datta, et al. (2018) which is very much different from the standard VQA. It combines question answering with a preceding navigation task in the environment, first looking for a target object that the question is about. When the agent reaches the navigation endpoint, the system answers

the question based on the view from its final position. Therefore, the success of the navigation directly affects the accuracy of question answering. EQA task is much harder than VQA, because (i) the robot does not contain a human model of attention (Dobnik and Kelleher, 2016), (ii) there is no guarantee that navigation will be successful, (iii) all questions relate to home environments, which are more similar to each other than unconstrained situations in the photographs used for VQA, and (iv) questions are limited in vocabulary, scope and complexity which restricts the language and makes it even a stronger predictor. To support the latter, Thomason et al. (2019) have shown that a language-only model outperforms multi-modal or vision-only system during QA in the EQA task. This demonstrates a stronger need for the deeper analysis of how and to what extent vision can be even utilised in the EQA model.

While most of the existing research on EQA has focused on the navigation subtask (Batra et al., 2020; Wijmans et al., 2019; Yu, Chen, et al., 2019), in this work we examine **the general role of vision for the QA in the EQA task**. In particular, we investigate how EQA model is using visual information and whether it is sensitive to visual perturbations when answering the question. First, we confirm previous results, comparing models trained and tested in different uni-/multi-modal conditions showing that just as in the VQA task, the model in the EQA task tends to hallucinate and disregard vision. Second, we turn to the examination of *how different visual disturbances affect performance of the model*. We evaluate the model with images of different types exemplified in Figure 6.30. The effects of various disturbances reflected in the evaluation scores will tell us how much removing context, content and (or) structure from images impacts question answering.

Our study can be viewed as *a test bed* to understand how vision is used in the EQA task. Similar benchmarks were developed for VQA (Agrawal, Batra, Parikh, and Kembhavi, 2018) and person-centric visual grounding (Luo, Banerjee, et al., 2022). In terms of the EQA, most of the work examined what can be used *instead* of the visual features. For example, Hu, Fried, et al.

(2019) show that using route structures instead of visual representations is better for the task. Schumann and Riezler (2022) found out that the model relies on properties of the environment graph much more rather than on visual features in the EQA for outdoor scenes. Different from previous studies, here we do not completely remove visual modality or compare it against other modalities. Instead, we evaluate *the limits* of the existing EQA model when its vision is permuted. We also view the EQA task as a simple NLG task, e.g. the model is asked to map important parts in vision and language (content selection) followed by prediction of a *single* label (surface realisation). In general, the focus of this paper is to understand the interplay between different modalities used in this simple generation scenario which is also relevant for generation of longer sequences of descriptions.

6.5.3. Task Description

Models The EQA task is split into two subtasks: navigation and question answering. Below we briefly describe the models used for both subtasks, a more detailed scheme is provided in Appendix 6.5.8. The navigation starts with an LSTM-based *planner* (Hochreiter and Schmidhuber, 1997) that selects an action from a pre-defined set (turn left, turn right, forward, stop) based on the question \mathbf{Q} , last action \mathbf{a}_{t-1} , last hidden state \mathbf{h}_{t-1} and visual representation $\mathbf{V}_t = F(\mathbf{I}_t)$, where F is a convolutional network (Cun et al., 1990) pre-trained on three tasks: RGB reconstruction, semantic segmentation, depth estimation. Next, the current hidden state of the planner \mathbf{h}_t , the predicted action \mathbf{a}_t and the current visual input \mathbf{V}_t are given to the *controller* that decides how many times the action has to be executed. The visual input \mathbf{V} is updated for each reiteration of the action. The controller is a simple multi-layer perceptron that returns control to the planner once it concludes that it needs a new action. The question answering module is an information fusion network. The question \mathbf{Q} is encoded by an LSTM network, while F takes N frames from the end of the navigation I_{T-N}, \dots, I_T once the agent has decided to stop (as predicted

by the planner) or the maximum number of actions $T = 100$ has been taken. Both representations are jointly attended and passed through a multi-layer classifier to predict a probability distribution across the answers.

Dataset The EQA dataset consists of automatically generated questions and answers from rules. The questions are made over visual scenes from the Matterport3D dataset (Chang et al., 2017) from which answers are generated. The authors use Habitat (Savva et al., 2019) to render the visual scenes. Each question in the dataset is replicated 15 times with different coordinates for the initial position of the agent as there is no single navigation path to the target object. There are three types of questions in the published dataset:

- colour: *What colour is the OBJ?*
- colour_room: *What colour is OBJ in the ROOM?*
- location: *What room is the OBJ located in?*

Nearly 70% of all questions are of colour_room type, ~15% are of colour type and the rest (~15%) are of location type. Placeholders *OBJ* and *ROOM* are filled with objects from dataset annotations (e.g., chair, plant) and room types (e.g., bathroom, kitchen) respectively.

Dataset and model limitations We describe several issues related to the EQA dataset. First, the quality of the rendered scenes is often poor, negatively affecting both navigation and question answering (Appendix 6.5.9). Annotations of answers are sometimes questionable, including the ways the set of possible answers has been defined (e.g., limited set of possible colours in the scene) (Appendix 6.5.10). A different concern is the “naturalness” of questions. Some questions are highly atypical of real interactions, e.g. why would one ask “What colour is the table in the living room?”. Another problem is that house environments are visually similar, consisting of instances of the same object classes (e.g., sofas, plants) that often share the same attributes (e.g., sofas are brown, plants are green). This also leads to an unbalanced distribution of answers: some answers (“black” and “brown”) are over-represented in the

dataset, possibly allowing the model to exploit these priors, e.g. sofas are often brown. Although this dataset bias amplifies the model’s ability to answer many questions about similar objects, artificially inflating accuracy on this dataset, the same biases prevent it from correctly answering questions about objects with specific properties, which require fine-grained visual understanding. Therefore in order to truly use vision to answer questions (e.g., when sofa is red, not brown), the model must have a *deeper* understanding of *fine-grained* visual representations, but as shown by Anand et al. (2018), the EQA models often struggle to utilise visual input. In the following sections, we will examine the level of visual understanding of the EQA model and overview problems on the dataset and modelling side that make it learn so little from vision.

6.5.4. Is language really stronger in EQA?

In the first set of experiments, we change the model’s vision stream or visual input representations. **Vis-L** is the standard EQA model (Das, Datta, et al., 2018) without any perturbations on the vision side. Given the question \mathbf{Q} and N image frames, the model predicts the most probable answer a^* :

$$a^* = \arg \max_{a \in \mathcal{A}} P(a | \mathbf{Q}, \mathbf{I}_{T-N}, \dots, \mathbf{I}_T). \quad (6.29)$$

For the **Blind-L** model, we keep the vision stream in the model, but change visual representations. In particular, we replace them with arrays of zeros before they are passed to the CNN for pre-processing:

$$\mathbf{I}_t = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{I}_t \in \mathbb{R}^{3 \times 256 \times 256}. \quad (6.30)$$

Finally, in the **\emptyset -L** model, we completely remove the vision stream and train it on questions only:

Metric	Vis-L	Blind-L	\emptyset -L
↓ Overall Mean Rank (MR)	4.352	4.454	3.685
MR, Color Room Questions	3.611	3.157	3.247
MR, Color Questions	2.693	2.261	2.304
MR, Location Questions	10.137	13.667	7.611
↑ Overall Accuracy (A)	0.38	0.323	0.362
A, Color Room Questions	0.374	0.348	0.337
A, Color Questions	0.528	0.478	0.522
A, Location Questions	0.222	0	0.278
Kappa Score	-0.005	0.014	0.024

Table 6.6. Results for the models both *trained* and *evaluated* with the specified settings described in Section 6.5.4. We also report results per question type. The best scores are coloured in **blue**.

$$a^* = \arg \max_{a \in \mathcal{A}} P(a|\mathbf{Q}). \quad (6.31)$$

We run all three models for 50 epochs using the official implementation²³ and choose the checkpoints with the lowest validation loss. For evaluation, we calculate accuracy (the top answer) and the mean rank (position of the correct answer in the ranked list of answers by the predicted probability distribution). We also compute Cohen's Kappa (Artstein and Poesio, 2005) which measures the agreement between the classifier and the ground truth dataset corrected by agreement by chance which is based on the distribution of labels. A kappa close to 0 (which ranges from 0 to 1 for agreement and 0 to -1 for disagreement) indicates that most agreement can be predicted only by knowing a distribution of labels. The higher the kappa the more the classifier is utilising additional knowledge that it has learned beyond a distribution of labels.

The results are shown in Table 6.6. The **Vis-L** model has the highest overall accuracy. However, the kappa score close to 0 shows that the model

²³https://github.com/facebookresearch/habitat-lab/tree/main/habitat_baselines/il

has a similar performance to a model that has memorised the distribution of labels. The lower mean ranks for **Blind-L** and **Ø-L** show that they are better at approximating the correct answer than the **Vis-L** model. These models strongly learn from language since the lack of vision does not prevent them from learning from biases in the dataset, leading to higher ranks. The **Vis-L** model however needs to process vision, but it is not capable of doing that (Thomason et al., 2019). Thus vision interferes and obstructs it from learning from language biases, confusing the problematic model and leading to lower ranks of the correct answers. When breaking the results based on question types, colour question are generally the easiest to answer, followed by the colour_room and location questions. The location questions are the hardest to predict in terms of accuracy and ranking overall. Furthermore, they are also most affected by different model configurations. In particular, the results suggest that the location questions are better predicted from language alone (**Ø-L**). The **Blind-L** model has the worst ranks and the worst accuracy overall. Its inconsistent performance across question types is hard to explain. Possibly, irrelevant visual information (black images) makes it more unpredictable than no vision at all or complete vision. Although the **Blind-L** is not the optimal model, it is still not far off from the other two models due to the second source of information - language.

Overall, we partially replicate the results of Thomason et al. (2019) and observe that vision is not that crucial. The role of language is much stronger than the role of vision, as demonstrated by the performance of the **Ø-L** model that predicts answers from questions alone. However, Frank, Bugliarello, et al. (2021) show that diminishing the importance of vision is detrimental for language tasks. Therefore in the second experiment we investigate *how different visual perturbations are utilised by the model and what are the model's limits in learning from vision*. We are particularly interested in examining if the model is able to understand complex high-level patterns from images or does it only learn lower-level information, which is present in some form in

different visual permutations.

6.5.5. “How much” vision is required?

To understand the limits of the model when utilising vision, we ask the following question: how much information can the model extract from different visual representations? We train the model according to Eqn. 6.29, but *evaluate* it on the vision with various levels of perturbations. In the **Eval-Shuffled** set-up, the model is provided with incorrect images for a specific question. In this case, the model gets structurally plausible representations which do not contain object(s) that the question asks about since the images depict a different house or room. We give more details about shuffling in the Appendix 6.5.11. The **Eval-Blind** model has been evaluated on images which were transformed into arrays of zeros, following Eqn. 6.30. In **Eval-Random**, the model has been given arrays of random noise as its visual input. The image vectors were replaced by an array of the specified shape ($3 \times 256 \times 256$) that was populated with random samples from a uniform distribution:

$$\mathbf{I}_t = \begin{bmatrix} \mathbf{v} & \cdots & \mathbf{v} \\ \vdots & \ddots & \vdots \\ \mathbf{v} & \cdots & \mathbf{v} \end{bmatrix}, \quad \mathbf{v} \in [0, \dots, 1]. \quad (6.32)$$

Results in Table 6.7 demonstrate that each of the **Eval-** configurations results in lower performance compared to the baseline (**Vis-L**). However, the model performs better on both incongruent (**Eval-Shuffled**) and black (**Eval-Blind**) images rather than random noise (**Eval-Random**). This suggests that the model is using *visual patterns* to support its prediction in some way. The performance across question types is similar to the results for models from the first set of experiments in Table 6.6: location questions are the hardest, colour questions are the easiest. Both experiments suggest that the visual information is not used as much as one would hope - disturbing vision or completely

Metric	Vis-L	Eval-Shuffled	Eval-Blind	Eval-Random
↓ Overall Mean Rank (MR)	4.352	5.145	5.508	6.899
MR, Color Room Questions	3.611	4.157	4.562	5.512
MR, Color Questions	2.693	3.035	3.087	3.319
MR, Location Questions	10.137	12.722	13.278	18.33
↑ Overall Accuracy (A)	0.38	0.266	0.246	0.211
A, Color Room Questions	0.374	0.264	0.258	0.258
A, Color Questions	0.528	0.307	0.217	0.194
A, Location Questions	0.222	0.222	0.222	0
Kappa Score	-0.005	0.013	0.004	-0.005

Table 6.7. Results for the models trained with original data (as Vis-L), but evaluated with specified conditions, described in Sec. 6.5.5). We also report results per question type. Intensity of the blue colour indicates performance of the model for the specific metric (more intensity means better performance).

removing it has little effect on the overall performance, suggesting that the model exploits language more. In terms of accuracy, location questions (which have the lowest accuracy on the baseline) are affected the least by different visual input. One reason could be that the baseline is bad so there is not much room for decrease in performance. Another reason could be that there are only 15 *distinct* location question-answer pairs in the evaluation set, seven of which are also found in the training. This may be the reason for a more exploitable language bias for location questions compared to other types.

6.5.6. EQA: biases and limitations

Recently, Hirota et al. (2022) have discovered social and gender biases in the VQA dataset. In the EQA, on the other hand, the model acts in the house environments with household objects without any humans, meaning that there are no biases towards any social group. The nature of dataset problems in the EQA task is different from VQA. One of the primary problems of the EQA is the lack of the perfect navigation module that would select correct images as input to the QA module. In addition, even if navigation is perfect, there is a chance for an image to be badly rendered (Appendix 6.5.9). These

problems combined make the task harder and bridge it with the likes of captioning of images taken by visually impaired people (Gurari et al., 2018) instead of VQA where images are fixed and taken in perfect conditions to answer the question. Another problem is of the limited scope of automatically generated questions and distribution of answers. In our view this directly forces the model to rely on language (which is limited and predictable) and to consider only basic visual patterns.

6.5.7. Conclusion

We looked at the Embodied Question Answering task and the corresponding dataset, focusing on how much vision is exploited by the QA module. The novelty of our study is the examination of *how* and *what* does the model learn from different types of images. Our results suggest that even if vision is not properly used, the model can extract general patterns from different visual permutations that are helpful to some degree. This means that the model could be looking at incongruent images or images with homogeneous structure (black) and answer questions correctly. Overall, we show that the model captures low-level knowledge of vision but is not capable of identifying and reasoning about specific high-level visual contexts that require understanding of scenes at a fine-grained level. Future work can improve model’s vision by implementing cognitive attention (Dobnik and Kelleher, 2016; Kruijff-Korbayová et al., 2015) or splitting the QA task into more subtasks because QA involves several inference steps and is not a simple pattern matching procedure. Using pre-trained multi-modal transformers such as LXMERT (Tan and Bansal, 2019) could also tell us whether these models are able to overcome problems related to dataset construction and image selection for the QA task in the EQA. If a performance of such a model improves then it must be the case that transformers capture some common sense knowledge through pre-training, but this could also be a hallucination of a different kind:

it is hallucination because it is general V&L knowledge not the specific one arising from a particular image and text.

6.5.8. Baseline QA Model

Fig. 6.31 shows the architecture of the baseline model for question answering in the EQA task. The model consists of three parts: language encoder, vision encoder and attention across both modalities. Questions are processed by a standard LSTM network (Hochreiter and Schmidhuber, 1997) that also learns word embeddings from scratch. $B = 20$ stands for the batch size, $N = 5$ is the number of used image frames taken from the last steps of navigation, $L = 11$ is the question maximum length, and $M = 64$ is the dimension size. Note that each question representation is repeated N times. Images are represented as three-channel (RGB) 256×256 egocentric scenes from the Habitat’s image renderer. A CNN network that has been pre-trained for RGB reconstruction, semantic segmentation and depth estimation is used to process images. The fully connected layer refers to a sequence of a linear layer, a ReLU layer, and a dropout layer with $p = 0.5$. $D = 4608$ is the dimension size of the visual processing network. The output representations from the language and vision encoder are jointly attended and summed across N frames. The resulting representation is passed to a multi-layer perceptron to predict the scores across $A = 35$ possible answers. We ran all models on 4 NVIDIA GeForce GTX 1080 Ti GPUs, running time was approximately 4 hours per model. In all experiments we report results for the models with the minimal loss across 50 epochs. In our experiments we did not use any explicit tools except Habitat-Lab²³, release version 0.1.7, MIT license.

6.5.9. Image Rendering Problem

While the majority of scenes are rendered properly, some of the scenes could be of poor quality. An example is shown in Fig. 6.32, demonstrating that

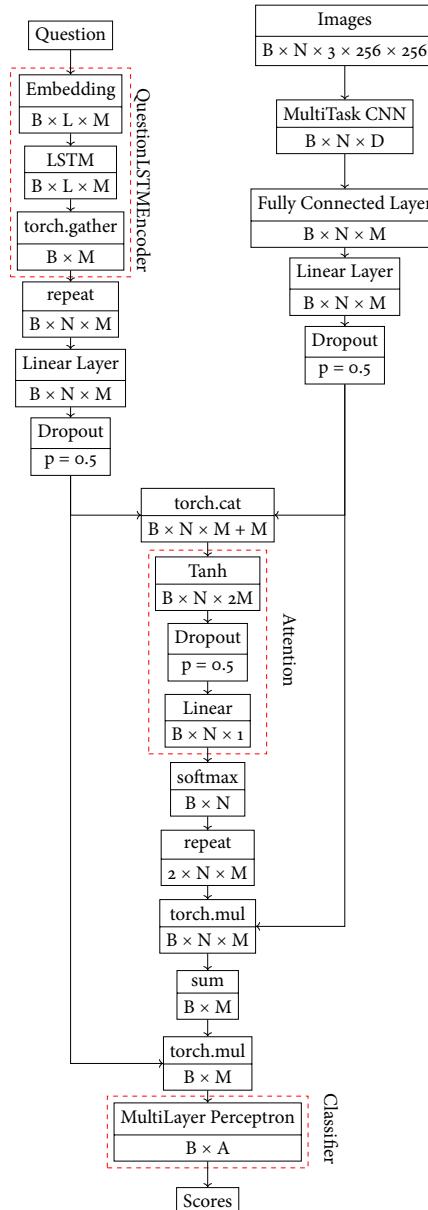


Figure 6.31. The baseline question answering model described in Das, Datta, et al. (2018) with available implementation in Habitat-Lab, link: https://github.com/facebookresearch/habitat-lab/tree/main/habitat_baselines/il. We schematically show the key components of the model: QuestionLSTMEncoder, Attention, and Answer Classifier. The stream in the top right side corresponds to the processing of visual information.

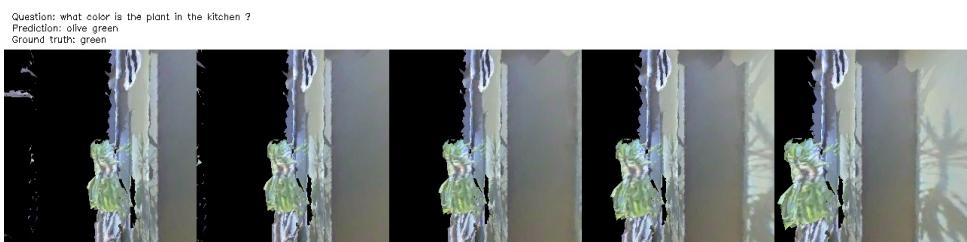


Figure 6.32. Example of a badly rendered scene from the EQA dataset.



Figure 6.33. Example of a sequence of images, the question, the predicted answer and the ground-truth answer.

the last five image frames used to answer the question include a lot of visual noise which makes the scene very confusing for a human eye. One wonders how an agent processes such poorly rendered scenes: does it rely on language information to answer the question? Note that scene annotations often include an object named “void” which is simply a black space. It is possible that the agent will encounter such confusing and uninformative space at the end of its navigation path. This could either confuse the agent or enforce better learning from the language stream. Or can the agent infer the answer from the general colours in the scene, given that the navigation often finishes at a close proximity to the target object? The example that we show is intended to demonstrate that due to the quality of the visual input, the agent might be biased to strongly learn from language and dataset biases.

6.5.10. Colour Problem

The EQA dataset has been generated automatically which means that it might contain errors. An example is shown in Fig. 6.33, where the question answering model has answered “tan” when asked about the colour of the sofa in the living room. One could say that, when looking at the image, the sofa is indeed tan, while there is an yellow armchair next to it. It could be that the model is actually correct in its prediction for a good reason and annotations are incorrect. The problem with colour annotations is also related to the set of colours used by annotators, that is coming from Kenneth L. Kelly’s “Twenty-two colours of maximum contrast” (Kelly, 1965) with two additional colours: “off-white” and “slate grey”. This set of colours has been designed to describe situations when contrast is needed (e.g., colour coding of graphs), not necessarily to depict colours in real world with natural descriptions. For example, the set introduces “buff” and “yellowish pink”, the former one is replaced with “tan” in the EQA dataset and the latter one is simply replaced with “yellow pink”, which makes the dataset even less natural. In addition, many colours in this set are easily confused under different lighting conditions (“white” and “off-white”, “grey” and “slate grey”), complicating the task for the question answering model.

6.5.11. Example Episode

An example episode structure from the EQA dataset. We display only a part of the shortest path coordinates and viewpoint lists. In **Eval-Shuffle**, shuffling is performed by modifying the original set of image frames and creating a new one. We show an example of one navigation episode from the EQA dataset below. A single episode includes a `question` field, which includes the question, answer, question type, and answer token IDs. We shuffle these question fields (line 68 in the example structure) across different episodes. Note that the authors of the dataset duplicated questions across multiple episodes, which, however, have different navigation paths to the target. This

has been implemented in order to ease the navigation task since there is no single correct navigation to the target object. We acknowledge that it could be possible that an episode with a shuffled question still has a valid set of last N image frames, but this possibility is low – for a single question, this probability is less than one percent.

```

1  {'episode_id': '640',
2   'scene_id': 'mp3d/5LpN3gDmAk7/5LpN3gDmAk7.glb',
3   'start_position': [15.50573335967819, -0.7660300302505512, 8
4     ↪ .392731789742543],
5   'start_rotation': [-5.312086480921031e-17,
6     -0.8526401643962381,
7     -0.0,
8     0.522498564647173],
9   'info': {'bboxes': [{ 'type': 'object',
10     'box': { 'centroid': [13.2358, -14.5238, 0.497693],
11       'a0': [1.0, 0.0, 0.0],
12       'a1': [0.0, 1.0, 0.0],
13       'a2': [0.0, 0.0, 1.0],
14       'radii': [0.593273, 0.243441, 1.68627],
15       'obj_id': 305,
16       'level': 0,
17       'room_id': 18},
18     'name': 'door',
19     'target': True},
20     { 'type': 'room',
21       'box': { 'centroid': [10.874245, -11.97072, 0.5380600000000001],
22         'a0': [1.0, 0.0, 0.0],
23         'a1': [0.0, 1.0, 0.0],
24         'a2': [0.0, 0.0, 1.0],
25         'radii': [3.1686549999999998, 3.26178, 1.95437],
26         'room_id': 18,
27         'level': 0},
28       'name': ['kitchen'],
29       'target': False}],
30     'question_meta': [{ 'name': 'colour', 'diffuse': 'grey'}],
31     'question_answers_entropy': 0.8303560860446519,
32     'level': 0},
33   'goals': [{ 'position': [13.2358, 0.4976929999999973, 14.5238],
34     'radius': 0.6412771421234348,
35     'object_id': 305,
36     'object_name': 'door',
37     'object_category': 'object',
38     'room_id': 18,
39     'target': True}]}

```

```

38     'room_name': 'kitchen',
39     'view_points': [{ 'position': [12.985883260576134,
40         -1.246680130110505,
41         14.494095338174798],
42     'rotation': [-2.855981544936522e-28,
43         -0.7071067811874078,
44         -0.0,
45         0.7071067811856873]},
46     ...
47     {'position': [13.089462756345679, -1.246680130110505, 13
48         ↵ .976197859327065],
49     'rotation': [-1.2227381688226952e-16,
50         -0.8910065241891411,
51         -0.0,
52         0.45399049973802935]}]}],
53     'start_room': 'R22',
54     'shortest_paths': [[[{'position': [15.50573335967819,
55         -0.7660300302505512,
56         8.392731789742543],
57     'rotation': [-5.312086480921031e-17,
58         -0.8526401643962381,
59         -0.0,
60         0.522498564647173],
61     'action': 2},
62         ...
63     {'position': [13.042462387438766, -0.7660300302505512, 13
64         ↵ .951177365325918],
65     'rotation': [-1.2227381690007914e-16,
66         -0.8910065242228339,
67         -0.0,
68         0.45399049967190386],
69     'action': 3}]]},
70     'question': { 'question_text': 'what colour is the door in the
71         ↵ kitchen?',
72     'answer_text': 'grey',
73     'question_tokens': [4, 5, 6, 7, 19, 9, 7, 10],
74     'answer_token': [0, 0, 0, 0],
75     'question_type': 'colour_room'}}
```

6.6. Context matters: evaluation of target and context features on variation of object naming

6.6.1. Abstract

Semantic underspecification in language poses significant difficulties for models in the field of referring expression generation. This challenge becomes particularly pronounced in setups, where models need to learn from multiple modalities and their combinations. Given that different contexts require different levels of language adaptability, models face difficulties in capturing the varying degrees of specificity. To address this issue, we focus on the task of object naming and evaluate various context representations to identify the ones that enable a computational model to effectively capture human variation in object naming. Once we identify the set of useful features, we combine them in search of the optimal combination that leads to a higher correlation with humans and brings us closer to developing a standard referring expression generation model that is aware of variation in naming. The results of our study demonstrate that achieving human-like naming variation requires the model to possess extensive knowledge about the target object from multiple modalities, as well as scene-level context representations. We believe that our findings contribute to the development of more sophisticated models of referring expression generation that aim to replicate *human-like* behaviour and performance. Our code is available at <https://github.com/GU-CLASP/object-naming-in-context>.

6.6.2. Introduction

The adaptability of human language presents a significant challenge for computational modelling, as it relies on both external contextual factors and internal personal beliefs and goals of the language users. The significance of the intents and goals cannot be overstated, as they dictate the specific choice of referring

expressions and object descriptions (Alikhani and Stone, 2019; Baltaretu et al., 2019; Ilinykh, Zarrieß, et al., 2018; Mädebach et al., 2022; van Miltenburg, 2017). Furthermore, these choices can vary depending on the specific task or the absence thereof. Put simply, language continues to evolve and adapt, while existing models are typically trained to generalise. Evaluating such systems proves hard, as evaluation metrics typically assume a single optimal solution, disregarding other valid alternatives (Kreiss et al., 2022). As variation in language arises due to different levels of underspecification between language units (words) (Pezzelle, 2023), addressing this problem brings valuable insights into understanding the effects of the task, contexts and how their interplay can be modelled.

But what is the “task”? And how do we define “context”? A task-oriented language use is often understood through the prism of human-human interaction, where communicative goals are important (Brennan and Clark, 1996). During these interactions, a shared understanding, known as a common ground, is established to optimise communication (Stalnaker, 1978). What ends up being in common ground is dependent on the task, and the importance of tasks and intents for modelling language has been emphasised in many recent proposals to language grounding (Andreas, 2022; Fried et al., 2023; Julianelli, 2022; Schlangen, 2022). In contrast, language can be used to simply describe objects in the world with an intent to **identify** them. These intents are typically determined by the set of instructions provided to a human e.g. “describe an image” (Lin, Maire, et al., 2014). In doing so, we perform *the object identification task* which is a communicative act, albeit a highly specific one.

The intent to simply describe things without a specific communicative goal has been one of the traditional tasks in the field of natural language generation (NLG). As referring is an important aspect of human communication (Frank and Goodman, 2012), much computational work has focused on building automatic referring expression generation systems (Krahmer and

van Deemter, 2012). The primary goal of referring expression generation is to produce a text in natural language that identifies a target object within a given context (Reiter and Dale, 2000) by making the object uniquely identifiable from the distractors. In the absence of the communicative intent, the definition of “given context” becomes extremely important as it directly influences referring (Schüz, Gatt, et al., 2023). **Visual context**, for instance, plays a crucial role in determining the content of the referring expression. This can be exemplified by multiple variables such as naturalness of the scenes where the target object appears (Kazemzadeh et al., 2014; Mitchell, Reiter, et al., 2013; van Deemter, van der Sluis, et al., 2006) or the presence of visual distractors and their position relative to the target object (Graf et al., 2016) and the typicality of the visual context as a whole (Gualdoni, Brochhagen, et al., 2022; Gualdoni, Brochhagen, et al., 2023; Gualdoni, Madebach, et al., 2022). But visual context is not the only context available in the task of referring. Humans also rely on their knowledge of the world when describing things, and their **background knowledge** influences the choice of referring given a specific visual context (Dale and Viethen, 2009). In fact, the use of various names to refer to a single entity stems from the fact that different speakers tackle underspecification in different ways. Humans use given context to fill in the missing information, but they do so differently based on individual perspectives. Therefore, investigating the effect of different contexts on the naming variation and capturing human behaviour in models is beneficial for developing a better REG architecture.

This study addresses two challenges: (i) existing models of referring are simply not learning to approximate possible names for entities and (ii) it is hard to generate a correct name if the level of semantic underspecification is high. As underspecification is correlated in humans with variation, we assume that the models that approximate human behaviour should be equally “confused” as humans when generating descriptions and should produce the same variation. For a model that is behaving this way we can be sure that

the variation is due to the way they capture semantic knowledge and context sensitivity rather than the noise (e.g., better performance on more frequent labels). **Our primary questions** are as follows: what is the set of features that enables computational model to closely capture the variation observed in human object naming? Can we combine such features to get closer to a REG model that can capture human-like object naming?

To address the questions outlined above, we investigate the effects that different context representations have on the model that is tasked with predicting an object name. We use CLIP (Radford, Kim, et al., 2021) to encode different context representations and train a simple classifier to predict target object names using the Many Names dataset (Silberer, Zarrieß, and Boleda, 2020; Silberer, Zarrieß, Westera, et al., 2020). We specifically examine how different features influence model’s ability to capture human object naming variation. Through the comparison of the model’s performance with humans across various metrics, we identify features that assist the model in making more valid and contextually motivated approximations of naming variation, reminiscent of human behaviour. We then combine different features and examine their fit for capturing naming variation. Our results demonstrate that the model that captures contextual sensitivity of object naming well (be it language or vision or both) is a good approximation of human knowledge and behaviour. We note that, unlike Silberer, Zarrieß, Westera, et al. (2020), we are testing how different types of knowledge contribute to naming variation rather than building or evaluating object naming models. While Silberer, Zarrieß, Westera, et al. (2020) also focus on typicality and whether the name is the top one or an alternative one in naming, we are interested in individual variation and the effects of context representations on the “distortions” of such typicality.

6.6.3. Problem formulation

6.6.3.1. Dataset

As our dataset, we use the Many Names dataset (Silberer, Zarrieß, Westera, et al., 2020) as it provides a suitable testbed for studying naming variation. This dataset stands out from other language-and-vision data collections that can be used for studying naming variation (Kazemzadeh et al., 2014; Krishna et al., 2017; Mitchell, Reiter, et al., 2013; Plummer et al., 2015; Yu, Poirson, et al., 2016) due to its high number of name types per object and alignment between names and objects. This way we can directly study the variation in reference to entities. The dataset was created by picking a single target object per image based on annotated data from Visual Genome (Krishna et al., 2017). Next, name annotations for each object were collected from multiple crowd-workers²⁴. There are on average 36 name tokens per object in Many Names, and their name types are sorted based on the frequency of being used to refer to objects. An example from the Many Names dataset is shown in the upper part of the Figure 6.34. In our experiments, we use the dataset splits of ManyNames v2.1 as reported in Silberer, Zarrieß, Westera, et al. (2020). Specifically, the train / val / test splits consists of 21503 / 1110 / 1072 items respectively.

6.6.3.2. Learning scheme

We approach object naming through the prism of referring expression generation. Our objective is to capture human-like variations in naming. Therefore, we shall look into the probability distribution of names that the model produces in a given context. Training a model to approximate naming distribution similar to humans should improve referring expression generation, possibly reducing deterministic nature of the models (van Deemter, Gatt, et al., 2012). However, one problem with the naming distribution in model’s output is that

²⁴For details, see Silberer, Zarrieß, and Boleda (2020).

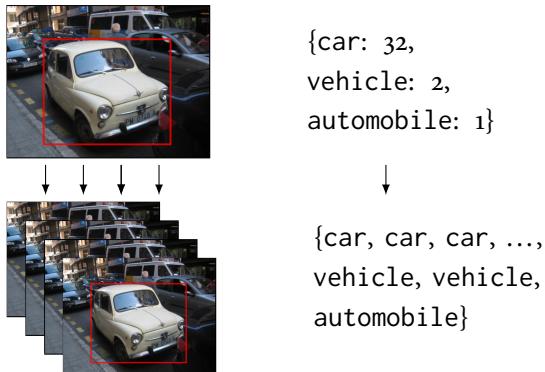


Figure 6.34. Dissecting the Many Names dataset (Silberer, Zarrieß, Westera, et al., 2020) into individual instances. The Target condition is depicted in which the model was provided with features of the object in the red box; datasets for Context-Obj and Context-Scene were built in the same way.

it may include invalid or non-human-like naming variations. To address this, we aim for our models to demonstrate shifts in the probability distribution, mirroring the changes observed in human object naming. These shifts are then learned by mapping different representations corresponding to visual context and background knowledge, rather than random noise, with the target names.

While it is possible to build different models per speaker to account for variation among these speakers (Dale and Viethen, 2009), our goal is to develop *a single function* that can approximate such variation across multiple individual describers. We deliberately chose to train such a simple model because it allows us to focus on evaluating the contribution of features to naming variation rather than the model's complexity. We ask if this function can predict the likelihood of a speaker referring to a particular object with a particular name. To answer the question, we break down the individual accumulated counts of frequencies into the number of individual referring events, each consisting of one description. This approach is similar to that of Coventry, Cangelosi, et al. (2005). The frequency of these events in the dataset

reflects the likelihood that the object would be referred to with that name. The bottom part of Figure 6.34 provides a more detailed example, which involves breaking down the counts of different name types from individual instances. This mirrors how humans describe an image, where each person may use different names for the same object. By learning from these individual instances, the network is expected to learn the variations in naming and, therefore, capture speaker uncertainty. During training, the model is repeatedly presented with input–“car” pair 32 times, while inputs mapped with “vehicle” and “automobile” are shown to the model 2 and 1 time, respectively. This variability in selection is akin to the diverse choices humans make in object naming. By using such training scheme, we encourage the model to learn *uncertainty* inherent in human naming, which is important for capturing variation. In the next section, we will describe how we represent different inputs to the name prediction model.

6.6.3.3. Input representation

The dataset consists of the following elements: for the j^{th} sample, there is an image i_j , a target object t_j with a bounding box t_j^{bb} obtained from Visual Genome, and a dictionary V_j containing names and their frequencies assigned to t_j by crowd-workers. Our initial proposal is to use each feature independently as input to a simple classifier to evaluate individual contribution of features. Next, a combination of different features can be explored. In terms of the features, we examine different types of representations which differ in the level of contextual information available. These include features that solely focus on the target object (Target), features that incorporate information about surrounding objects but exclude the target object (Context-Obj), and features that cover knowledge about the entire scene (Context-Scene). For each feature type, we consider three representation modes: visual, linguistic, and their combination. We encode each feature type with CLIP (Radford,

Kim, et al., 2021)²⁵, a pre-trained multi-modal transformer that learns strong multi-modal representations through its contrastive learning on large amount of image-text pairs. Our motivation for selecting different modalities and combining them is as follows. Text features can be seen as representations of the background knowledge in terms of the meaning of a word in the contexts that were given to the pre-trained model, e.g. CLIP. This knowledge is acquired through extensive pre-training, and CLIP, in particular, possesses rich contextual information about entities and objects. Hence, textual features encode *general* knowledge about the interaction of these objects, not related to particular events (although it is possible that due to naming variation of labels some specific local context is also captured). An example of this type of world knowledge includes the typical contexts in which bananas appear (kitchen, food, nature, market), how they are typically used (eaten, consumed), and who typically uses them (humans, animals). On the other hand, vision features contain information about the immediate context of the target object. Their purpose is to encode the situation in which the object appears in a specific case. Here is an example of this type of feature: a more detailed and specific understanding of the situations in which bananas appear could involve a market with various fruits of different colours and a better understanding of how bananas fit into this specific context. By integrating both these feature types, we take a step toward modelling the information sources that humans employ for object naming. These features include world knowledge about how objects interact in the world and specific visual information about these objects.

In the Target condition, our aim is to examine the effect of the knowledge about the target object in the process of object naming. We seek to determine whether a model can effectively capture naming variation in the absence of contextual information, relying solely on the appearance and/or common sense knowledge of the target object. To represent common sense

²⁵We use a pre-trained ViT-L/14@336px based on the code from the official CLIP GitHub repository: <https://github.com/openai/CLIP>.

knowledge²⁶, we use labels that have been assigned to objects (both target and context) by the annotators of the Visual Genome dataset (Krishna et al., 2017). By encoding these labels with CLIP, we can leverage strong signals and extensive additional knowledge about the objects. It is important to note that this type of information is not typically available to a conventional referring expression model. In fact, any identification system that uses this information would be considered cheating in predicting names. In our experiments, we incorporate this knowledge to evaluate its contribution to generating a variety of names, but it is important to acknowledge that this feature may or may not be available in individual tasks.

With the Context-Obj condition, we measure how well a target’s name can be predicted from surrounding objects alone. In other words, can we “guess” a name based on the visual and/or common sense knowledge about context objects? Finally, with the Context-Scene condition, we focus on attention and search: given visual and/or common sense knowledge about the scene as a whole (e.g., all objects treated equally, no difference between context or target objects), can we model human naming variation?

Target We represent visual \mathbf{v}_j^v and linguistic \mathbf{v}_j^ℓ information about the target object as follows:

$$\mathbf{v}_j^v = f_{\text{CLIP}}(t^{\text{bb}} j), \quad (6.33)$$

$$\mathbf{v}_j^\ell = f_{\text{CLIP}}(t_j^{\text{VisGen}}). \quad (6.34)$$

Here, t_j^{VisGen} represents the label of the target object from Visual Genome.

Context-Obj Another type of feature that can be explored is the knowledge of context. In this particular setup, the input representations do not contain

²⁶In this study, we use the terms “linguistic” and “common sense” interchangeably, as they both refer to the knowledge and understanding of language-related information and general knowledge about the world.

any information about the target object, whether visual or common sense-related. This setup can be viewed as a “guessing game” where the model is given a context representation and tasked with predicting the name of an object likely to appear in that context. To model this scenario, we use Visual Genome annotations to represent the context of the target object. Specifically, we extract a list of bounding boxes for all objects that are *not* the target object, denoted as $\mathbf{R}_{\setminus t_j} := (r_1, \dots, r_K)$, where K is the number of objects in i_j . Then,

$$\bar{\mathbf{v}}_j^v = f_{\text{CLIP}}(\mathbf{R}_{\setminus t_j}), \quad (6.35)$$

$$\bar{\mathbf{v}}_j^\ell = f_{\text{CLIP}}(\mathbf{L}_{\setminus t_j}), \quad (6.36)$$

where $\mathbf{L}_{\setminus t_j}$ is the list of object descriptions, where each element is a simple phrase consisting of a name and up to five attributes from Visual Genome annotations, e.g. “car black big”, and $\bar{\mathbf{v}}$ is the average of the objects or their descriptions. We also apply L2 normalisation on the resulting vector to obtain a more robust context representation. This normalisation helps enhance the discriminative power of all feature vectors and disregards the influence of differences in magnitude and scale²⁷. The motivation behind this design choice is further described in Appendix 6.6.8.

Context-Scene In the third experiment, our focus is to examine the predictability of naming variation from the context *as a whole*. We use perceptual features of the entire image that have been encoded with CLIP and incorporate object-relation triplets that describe the content of the scene. These triplets are sourced from the Visual Genome dataset, where each image is annotated with relationships. We note that these relationships are generated by different crowd-workers, ensuring a diverse range of annotations for our experiment. While the number of relations may differ from image to image, they collectively provide an overview of the objects present in the scene and their

²⁷In each experiment where we need to create a single vector from a list of vectors, our approach is to first compute the average vector from the list and then normalise it.

associated events. By leveraging these relationships, we can create language input features for the Context–Scene model:

$$\mathbf{v}_t = f_{\text{CLIP}}(< S, P, O >), \quad (6.37)$$

where $< S, P, O >$ represents a single string comprising the subject, predicate, and object names of a specific relationship triplet. Since annotated scene contexts in Visual Genome are not predetermined and vary across images, textual descriptions can be constructed in various ways. To generate textual scene descriptions, we shuffle and randomly extract a varied number of relationship strings. We then employ different methods to feed these strings to the CLIP model in order to obtain language features. Subsequently, we evaluate the Context–Scene model using each type of text representation to identify the one that demonstrates optimal performance. The selected model is then used in our primary experiments. More details on how the best Context–Scene model that uses text was chosen can be found in Appendix 6.6.9.

6.6.4. Model

In this study, we adopt a simple approach by constructing a `CLS` (classification) model. The objective is to approximate a function that can predict naming variation. The success of this function approximation provides insights into the suitability of the features as predictors of naming variation. The approach is akin to the use of generalised linear models in statistical testing, where we aim to capture the relationships between the features and the predicted labels. To maintain a close connection to linearity, we build a single-layer feed-forward network as our model. We specifically examine the probabilities assigned to all the labels predicted by the model and evaluating their degree of variation against the probabilities assigned by humans.

The model is trained following the scheme outlined in Section 6.6.3.2 and takes input representations described in Section 6.6.3.3. The model takes \mathbf{x} which is either a concatenation of visual and linguistic features $\mathbf{x} = (\mathbf{v}_v \oplus \mathbf{v}_\ell)$

or a uni-modal feature, e.g. $\mathbf{x} = \mathbf{v}_v$ or $\mathbf{x} = \mathbf{v}_\ell$, where $\mathbf{x} \in \mathbb{R}^{1 \times 768}$. The model is trained to predict a target name \mathbf{y} from the set of all possible names that are available: $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$, where $N = 1642$ is the number of all possible names. N is determined by the set of unique names across all data splits. The model is defined as follows:

$$\hat{\mathbf{y}} = \sigma((f_2(f_1(\mathbf{x})))), \quad (6.38)$$

where

$$f_1(\mathbf{x}) = \text{ReLU}(\text{BN}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)), \quad (6.39)$$

$$f_2(\mathbf{x}') = \text{Dropout}(\mathbf{W}_w \mathbf{x}' + \mathbf{b}_2) \quad (6.40)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d_2}$, and $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times 1}$ is output linear layer that produces the list of logits $\tilde{\mathbf{Z}} \in \mathbb{R}^{1 \times N}$. The model applies **softmax** σ over the last dimension of $\tilde{\mathbf{Z}}$ to transform unnormalised scores into name probabilities. We adjust d_1 depending on the type of the experiment: if we test features from a single modality, then $d_1 = 768$, otherwise $d_1 = 1536$. We set $d_2 = 512$ and Dropout = 0.1.

All models were trained using a batch size of 64 and standard cross-entropy loss. The Adam optimiser (Kingma and Ba, 2015) with a weight decay of $1e-5$ was used, and the learning rate was set to $4e-3$. During training, the gradients were clipped by their norm per single batch, with a maximum norm set to 3. The models were trained for a total of 200 epochs, and the best model was selected based on the validation loss at the epoch level. Additionally, we used a scheduler, reducing the learning rate if there was no improvement in the loss for three consecutive epochs during validation.

6.6.5. Evaluation metrics

To evaluate the general performance of the model, we use multiple metrics. We note that during evaluation, we do not differentiate between top and

alternative names. Our model learns that each possible name is valid but to varying degrees based on the frequency of being assigned to an object. The model is never presented with multiple names and their frequencies simultaneously. This means that it does not make comparative judgments about one name being more or less valid than another. Therefore, our results should be interpreted as an assessment of how often the model would use a specific name to describe an object, without considering its relation to other alternatives.

Firstly, we measure the model's ability to predict the top name (e.g., the most frequent name) by looking at accuracy @1. Other degrees of accuracy are also useful to consider, as they indicate whether the top name occurs in the top- k predictions generated by the model, where k is the number of name types used to describe a specific target in the specific image. The final accuracy scores are reported as averages over the total number of samples. We also compute the mean rank of the ground-truth label among the model's predictions and report the average mean rank (AMR) across all items. Additionally, we measure the perplexity of the models as an indicator of overall predictive performance. Unlike accuracy, which solely focuses on comparing the top name, perplexity allows us to compare the variation in the predictions of different names. However, perplexity does not measure semantic equivalence or similarity between the predicted names and the human-generated names. We note that since we have previously evaluated the success of the model with accuracy, we can assume that such noise is minimised. We compute perplexity PP by taking the logarithmic base of the entropy and raising it to the power of entropy, e.g. $\text{PP} = \exp^H$.

To evaluate the suitability of features for predicting naming variation, we calculate the entropy (Shannon, 1948) of each model and humans. Entropy helps us quantify uncertainty, and we anticipate that the best model will demonstrate a similar level of uncertainty as humans. To assess the degree of association between the entropy of each model and human responses, we

Condition	Mode	Accuracy (%) ↑			AMR ↓	PP ↓	H ↓	ρ
		@1	@5	@10				
1	TEXT	69.15	87.68	89.94	41.45	4.745	0.210	0.540*
2	Target VISION	56.70	81.09	86.34	52.87	7.199	0.266	0.485*
3	VISION-TEXT	70.02	90.99	92.30	33.77	3.740	0.178	0.574*
4	TEXT	40.90	67.58	76.73	52.13	14.924	0.365	0.343*
5	Context-Obj VISION	49.14	75.14	83.20	40.79	10.360	0.315	0.328*
6	VISION-TEXT	46.48	72.98	81.04	45.87	11.531	0.330	0.321*
7	TEXT	4.09	16.85	31.80	59.00	51.111	0.531	-0.024
8	Context-Scene VISION	47.93	73.51	81.42	60.73	9.116	0.298	0.410*
9	VISION-TEXT	53.34	77.91	83.98	38.87	8.281	0.285	0.424*
Human					1.623		0.065	1.000

Table 6.8. Evaluation of different features (models 1-9) against human scores. We highlight the top three models **per condition** in each metric, with colour intensity reflecting their performance (stronger indicates better). Human scores are provided as a reference. The values of Spearman correlation ρ with * denote a very high level of significance, e.g. p-value ≤ 0.001 .

compute Spearman’s rank correlation coefficient (Spearman, 1904). This metric measures the monotonic relationship between the two, and it serves as our primary evaluation metric. The way entropy is calculated is slightly different between the model and humans in terms of the probabilities that we use. For the model, we take the degree of belief that the object should be assigned a particular label by the neural network, represented by logits \tilde{Z} . These logits are transformed into probabilities using the **softmax** function: $P_m = \sigma(\tilde{Z})$. For humans, we consider the probability (derived from frequencies) that a human would assign a particular label to the object, representing a collective likelihood. For each test item, we collect all available ground-truth human responses (m) and their corresponding frequencies (x_1, x_2, \dots, x_m) . These frequencies are then transformed into probabilities:

$$p_i = \frac{x_i}{\sum_{j=1}^m x_j}, \quad \text{for } i = 1, 2, \dots, m. \quad (6.41)$$

Next, we construct a new vector $\mathbf{P}_h \in \mathbb{R}^{1 \times N}$, where values in positions corresponding to the positions of each response in the model's dictionary \mathcal{V} (with $|\mathcal{V}| = N$) are replaced with their respective probabilities p_i , and the rest are set to 0. To compute entropy \mathbf{H} of \mathbf{P}_m and \mathbf{P}_h , we use the following operation:

$$\mathbf{H}_{m \setminus h} = - \sum_{k=1}^{|\mathbf{P}_{m \setminus h}|} p_k \log p_k. \quad (6.42)$$

We normalise the maximum attainable entropy by $-\log \exp(N)$ to ensure comparability between different models, resulting in entropy values ranging between 0 and 1, where 1 represents the highest possible entropy. All metrics are reported as averages across the test set. We anticipate that the model probabilities will show greater variation across labels due to noise compared to humans, as the model may assign low probabilities to labels that are not applicable. On the other hand, humans tend to produce “cleaner” labels as they are direct judgments. To address this issue, we compare the ranks of entropies using correlation coefficients. This choice is relevant because the vector \mathbf{P}_h contains many zero values, which motivates us to focus on the ranks of the values rather than the values themselves. When describing an object, humans select from a limited set of “valid” names, whereas the model considers both “valid” and “invalid” names (a total of 1642 possible name types). By examining the ranks of the model’s predictions, we mitigate this issue. We would like to emphasise the general importance of statistical testing to determine the extent to which the model’s performance is influenced by either the network design or the features themselves. In this paper, we employ Spearman correlation to measure the relationship between input features and target variables. This test is appropriate because we are interested in whether the simple neural network can approximate a function between input features and the resulting naming variation. This correlation shows whether there is a linear relation between the model’s prediction and human scores and,

therefore, whether those input features are associated with human scores. We believe that future work can focus on measuring the effects not only of features but also of the model’s design on naming variation.

6.6.6. Results

Table 6.8 demonstrates the results of our experiments, which focused on evaluating different feature representations (modes) for various feature types (conditions) in modelling naming variation. Firstly, we examine differences within each condition and analyse different modes to identify the best features for representing specific condition. Next, we explore the differences between conditions and consider the potential of combining them to achieve a more human-like performance in the object identification model. We conclude by emphasising features that need to be encoded by an REG (Referring Expression Generation) model to effectively capture human-like object naming variation.

6.6.6.1. Best feature per condition

Representing targets In the Target condition, multi-modality proves to be crucial as it achieves the highest performance in predicting the correct answer, exhibiting the lowest mean rank and perplexity. Additionally, language-and-vision features significantly reduce uncertainty and bring it closer to human levels, as indicated by entropy and correlation measures. Notably, language appears to contribute more to the fusion of modalities, as it offers greater informativeness compared to visual information. This observation aligns with previous studies conducted on various multi-modal tasks Agrawal, Batra, Parikh, and Kembhavi, 2018. The contribution of the text mode can be attributed to the degree of semantic similarity that an object label from Visual Genome and a target name share with each other. For example, the Visual Genome label for the target object in Figure 6.34 is “sedan”, which is very

similar in meaning to the target names, while context labels (“street”, “human”) might be less useful in reducing uncertainty for naming. Additionally, encoding it with CLIP that is expected to understand relations between “car”, “sedan” and “vehicle” might provide even more informative representations, reducing ambiguity about the choice of the name. Nonetheless, the vision representation in the Target condition demonstrates good performance, as it does not lag far behind the performance of the text features. One possible explanation for this result is that the knowledge in text is simply not very effective, either due to noise or its challenging nature to learn from, or it may not be very informative. We emphasise that it is important to evaluate the quality of knowledge types in the Limitations section. Interestingly, incorporating visual appearance of the target object further enhances the correlation between the predicted and human naming variation. We conclude that for effectively representing the target object, the most optimal feature representation involves combining visual information with common sense knowledge of the target object.

Representing context as objects In the Context-Obj condition, the vision-only model demonstrates the best performance in predicting a single correct name and achieves the lowest mean rank of the correct name in its predictions. It also has the lowest entropy among the different modes considered. However, it is important to note that the vision-only model does not exhibit the highest correlation with human naming variation. The highest correlation is observed when the model relies solely on textual features, despite having the highest entropy among all three modes. This observation is interesting as it emphasises the significance of world knowledge in capturing naming variation. Understanding what objects might co-occur in a given context provides valuable information to the model (Dobnik, Ilinykh, et al., 2022). For instance, having the context labels “counter”, “fridge”, and “oven” might assist the model in predicting the target name “pot” more accurately than relying

Condition	Accuracy (%) ↑			AMR ↓	PP ↓	H ↓	ρ
	@1	@5	@10				
3+9	71.02	88.59	90.62	37.66	3.773	0.179	0.580*
3+4	70.55	88.76	90.62	43.02	4.187	0.193	0.568*
3+9+4	71.41	89.73	91.42	38.96	3.995	0.187	0.578*

Table 6.9. Evaluation of different combinations of the best-performing features from Table 6.8. The meaning of colour intensity and * is described in Table 6.8. The numbers in condition correspond to the features from Table 6.8.

solely on visual features of these context objects. Interestingly, contrary to the Target condition, combining linguistic and visual information leads to the lowest correlation score. Based on these results, we conclude that representing context in a model that aims to capture naming variation is best achieved through the textual labels of the context objects.

Representing context as a scene When representing context as a single image with or without relationship triplets, combining language and vision yields the best performance across various metrics, including correlation with humans. There is a notable reduction in uncertainty and an increase in correlation when the model has access to the visual appearance of the context alone, represented by the image as a whole. This improvement can be attributed to the model’s ability to better contextualise the target object as text knowledge provides only general information about what context objects are and lacks details on how the objects actually look. In contrast, uncertainty in the model is significantly high when the model is provided with relationship triplets alone. In fact, this condition shows no correlation with human naming at all. The text-only model stands out with exceptionally high perplexity and significantly higher entropy compared to any other model in any of the conditions. We believe this highlights the importance of choosing appropriate representations for conveying textual knowledge about the scene. Exploring the performance of models using other types of representations,

such as scene categories, captions, or more coherent scene descriptions, is left as a topic for future investigation. Considering that the task involves mixed representations of targets and context without explicit labelling, the Context-Scene model approximates correlation most effectively when there is a fusion of modalities.

Overall, the findings indicate the importance of the text modality in learning about the target object. However, combining text with vision is necessary to achieve lower entropies and higher correlations with human naming. This demonstrates that predicting a name solely from text is challenging because the model lacks knowledge about the appearance of objects and struggles to determine what to focus on. Access to visual representations allows the model to differentiate between targets and contexts, possibly due to factors such as the perspective and location of the objects, which are relevant for naming. In the next section, we focus on identifying the optimal feature combination for better object naming. Our goal is to assess the correlation with human naming when multiple conditions are combined, thereby determining the best possible combination of features.

6.6.6.2. Combining best-performing features

Here we test different feature combinations to replicate human-like naming variation. We acknowledge that without testing of *all* possible combinations, we cannot really conclude which feature combination is the best. However, here we have chosen feature combinations based on our intuition regarding what is commonly found in models and what yields the best performance when considering individual features. Table 6.9 presents the results of combining features that have shown the highest correlation with humans across different conditions. For each condition, we progressively combined features that showed the highest correlation with humans by concatenating them together. As a result, the input vector size for the 3+9+4 condition became 5×768 , representing the combination of two modalities for the target, two modalities

for the context as a scene, and one modality for the context as objects. The best model, which incorporates visual and common sense knowledge about the target (3 in Table 6.8) along with multi-modal knowledge about the scene (9 in Table 6.8), achieves the lowest entropy and improves the correlation with humans compared to the previously best model, the Target model. This indicates that combining the appearance of an object, including its label, with the shared context and thematic representation of the scene as a whole can be beneficial. Interestingly, combining different features with each other generally yields better results than using them individually, except for the combination of the best Target and Context-Obj models. The optimal combination is found to be the integration of knowledge about the target with knowledge about the scene as a whole. Notably, the 3+9 combination achieves lower accuracies, suggesting that it may be more focused on capturing variation rather than predicting the most probable name. These findings have implications for the representation of context. While the visual appearance of objects is important, it also needs to be presented in a consistent and comprehensive manner, such as using a whole image where the relationships among context objects are clear, and the fit of the target within the overall context can be easily extracted.

6.6.7. Conclusions

Naming and language in general is semantically underspecified (Frison, 2009; Pezzelle, 2023). To fill in the missing gaps in reconstructing meaning, language users rely on contextual information, be it perceptual information or background knowledge. In this study we examined different types of context representations for capturing human object naming variation. We have found that to capture naming variation it is important to have a lot of knowledge about the target object. We also have shown that the way context is represented matters: object-level visual representations might narrow down the gap in uncertainty between models and humans, but they might not correlate the most with humans in object naming. Future work on this topic should focus

on using encoders other than CLIP, building more complex classifiers and investigating the effect of different ways to represent common sense knowledge (e.g., not relationship triplets, but captions or another type of image descriptions). Also, looking at object naming in a task context with communicative goal is another important direction.

Limitations

Information fusion This work uses averaging to generate a single vector when combining multiple language and/or vision features. It should be acknowledged that adopting an alternative fusion method, such as multiplication or summation, could potentially affect the final scores of the models, particularly when the differences between them are relatively minor. We recognise that the results reported in this study are specific to the particular technical setup employed, involving L₂ normalisation with averaging. Hence, further investigation is warranted to determine whether the reported findings remain consistent when using a different fusion method. Some of our ideas for information fusion are presented in Appendix 6.6.8. In addition, fusing different features from different conditions by multiplying them or learning a function to fuse them can be an alternative to a simple concatenation that we use in this study.

Knowledge representations We note that in the context of a standard REG task, knowing the label of the target is practically impossible. Hence, it is expected that a model with linguistic knowledge about the target would perform well. Also, adding more features (visual, linguistic, others) appears to hinder performance due to the increased number of parameters and a larger hypothesis space. Therefore, the objective of learning should be to strike a balance between model size and feature informativeness. It is also important to seek a knowledge representation that closely resembles how humans name objects.

6.6.8. Fusing features

In our approach, when it is necessary to combine multiple uni-modal or multi-modal representations into a single vector, we use averaging of features. This averaging process is followed by an L₂ normalisation step, which normalises the features based on the Euclidean distance between individual points. Additionally, we have experimented with using multiplication for feature fusion, particularly in cases where we want to emphasise joint features or attributes and assign more importance to overlapping information. Multiplication is expected to highlight specific features that are shared across objects, such as in the case of visual features. However, we have observed that multiplication often leads to many zero values in the resulting features, and in some cases, it even leads to inf or NaN values due to the sparsity of visual representations. This sparsity can make the resulting vector difficult to learn from, especially depending on the number of objects being multiplied. Although summation of features is a straightforward approach, we have concerns that using this method results in a diluted final vector. As a result, we decided to use averaging followed by L₂ normalisation as it tends to be a more effective and stable approach for feature combination.

6.6.9. Representing language for Context-Scene

Table 6.10 presents the performance of various variations of the Context-Scene model, which incorporates the textual modality. The text representation can be either a single string containing 10 or 5 relations present in the image (10-string and 5-string), or a list of different relations (10-list and 5-list). The best model is selected based on the loss and average mean rank score, both computed on the test set. The best-performing model is highlighted in bold in the table.

Condition	Text Format	Accuracy (%) ↑			AMR ↓	Loss ↓
		@1	@5	@10		
Context-Scene + Text	10-list	4.04	17.98	30.43	62.63	4.774
	10-string	4.09	16.85	31.80	59.00	4.676
	5-list	3.83	16.69	31.70	63.32	4.756
	5-string	3.58	16.99	30.80	125.50	5.722
Context-Scene + Vision-Text	10-list	52.40	75.58	83.09	45.49	2.490
	10-string	53.27	77.27	83.24	43.38	2.463
	5-list	52.44	76.68	83.11	45.12	2.475
	5-string	53.34	77.91	83.98	38.87	2.403

Table 6.10. Performance of different Context-Scene models, which use textual modality as part of their input.

6.7. Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation

6.7.1. Abstract

This paper describes insights into how different inference algorithms structure discourse in image paragraphs. We train a multi-modal transformer and compare 11 variations of decoding algorithms. We propose to evaluate image paragraphs not only with standard automatic metrics, but also with a more extensive, “under the hood” analysis of the discourse formed by sentences. Our results show that while decoding algorithms can be unfaithful to the reference texts, they still generate grounded descriptions, but they also lack understanding of the discourse structure and differ from humans in terms of attentional structure over images.

6.7.2. Introduction

What are the properties of the well-generated text? This question has been in the centre of many debates in the natural language generation community (Dale and White, 2007; Gatt and Krahmer, 2017). While human evaluation has always been the gold standard in the quality assessment of generated texts, the field is often reluctant to run such evaluation due to the lack of standardisation in evaluation reports and generally high cost (Howcroft et al., 2020). Therefore, a number of simpler and cheaper *automatic metrics* were introduced, specifically in the field of machine translation, although their validity has been questioned (Reiter and Belz, 2009).

As computer vision and NLP started to merge, automatic metrics became an important part of the evaluation process of image descriptions. In general, image descriptions are evaluated with means of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE (Lin, 2004),

CIDEr (Vedantam, Lawrence Zitnick, et al., 2015) and WMD (Kusner et al., 2015). However, Kulkarni, Premraj, Dhar, et al. (2011) and Elliott and Keller (2014) have demonstrated that such metrics only weakly correlate with human judgements in the context of image description generation task. The discrepancy between human and automatic evaluation is deeply rooted in the differences between the fields of machine translation which originally introduced aforementioned metrics and image captioning, which adopted them. In principle, text-only evaluation is highly constrained: the key requirement for high-quality translation is the perseverance of semantics between two parallel texts. In comparison, evaluation of texts generated in multi-modal tasks is influenced by many factors as the generated texts might mention a different set of objects, attributes and relations which are not described in reference texts. Such generations would cause low values from reference-based metrics, although they could be completely plausible and truthful to the image. As such, the tasks of machine translation and image captioning are inherently dissimilar in terms of evaluation. To mitigate this problem, metrics that directly compare texts against image objects have been proposed (Hessel et al., 2021; Jiang et al., 2019; Madhyastha et al., 2019; Wang, Yao, et al., 2021). They are typically better than BLEU in that they assign a more accurate score to image-correct descriptions. A relatively recent trend has been to develop a set of metrics that would evaluate goal-oriented captions, produced with specific communicative intent (Inan et al., 2021) or for a specific group of users (Fisch et al., 2020), for example, if an image of a snowdrop is described as “the spring flower”.

A notable feature of the aforementioned metrics is their sole focus on evaluation of image captions. Different from captions, **multi-sentence** image descriptions impose additional challenges for generation systems including understanding of the textual *discourse* in the multi-modal context. Analysis of discourse has been in the focus of both text-only (Poesio, 2004; Poesio et al., 2004) and language-and-vision tasks (Dobnik, Ilinykh, et al., 2022; Takmaz

et al., 2020). However, given a huge interest in generation of longer image descriptions, e.g. image paragraphs (Ilinykh and Dobnik, 2020; Kong et al., 2014; Krause et al., 2017), recipes (Nishimura et al., 2019) and stories (Huang et al., 2016), we believe it is important to gain a deeper insight into how humans and models *structure* and *realise* discourse in such descriptions. In this paper, we understand discourse as a match between linearisation of the semantic knowledge (e.g., a fit of non-linear concepts into linguistic linear order) and underlying planning (Reiter and Dale, 1997). We build on previous intuitions about evaluation in NLG and look under the hood of how different decoding algorithms build discourse in image paragraphs. We compare a number of decoding strategies for correspondence with how humans distribute and describe objects in longer texts. The main purpose of this study is to gain insights into whether decoding strategies generate texts *similar* to humans and whether these texts exhibit the corresponding discourse structure. There is a limitation on what and how things can be communicated and decoding algorithms have a direct control over it. The choice of decoding algorithm also has an effect on how information is expressed in the communicative channel (Shannon, 1948) and how successful its reconstruction by the perceiver will be (Lazaridou et al., 2017). Our results shed more light on the differences between decoding algorithms in terms of (i) the discourse structure, (ii) faithfulness to the reference texts, (iii) groundedness into the image and (iv) attentional structure.

6.7.3. On the importance of decoding

It is impossible to neglect the impact of the choice of the decoding on the structure of the generated texts²⁸. Discourse in multi-modal descriptions can be affected by many factors, including scene structure (Linde and Goguen, 1980), the desire to have more accurate or more diverse texts (Massarelli et al.,

²⁸For a broader overview of the factors that influence inference in generation we refer the reader to Zarrieß et al. (2021).

2020; Zhang, Duckworth, et al., 2021) and aspects of the task (Kiddon et al., 2016; Narayan et al., 2022). Other constraints include adherence to a specific topic as in poetry generation by controlling for content and form (Hopkins and Kiela, 2017) and incorporating pragmatic reasoning when describing images with text (Cohn-Gordon et al., 2018; Vedantam, Bengio, et al., 2017) or optimising model’s predictions for a specific metric (Gu et al., 2017; Rennie et al., 2017; Zarrieß and Schlangen, 2018) in the spirit of reinforcement learning. Notably, Balakrishnan et al. (2019) have shown that using tree-structured semantic representations, similar to those used in traditional rule-based NLG systems, helps to evaluate generated texts during decoding for the specific discourse. In this work, we describe analysis on *what* and *when* different algorithms generate, comparing their outputs with the human gold standard.

6.7.4. Task and model

As our modelling task, we choose the task of image paragraph generation and the Tell-me-more corpus described in (Ilinykh, Zarrieß, et al., 2019b). In this task, a human is given an image and five (5) text fields. The describer writes sentences about the image so that they help a potential listener to identify it within a set. The describer is also asked to write sentences in a sequence, keeping in mind that after each sentence the listener needs more information to identify the image, e.g. thus, tell-me-more. Ilinykh, Zarrieß, et al. (2019b) show that collected multi-sentence descriptions have a fixed intentional structure, in the sense of Grosz, Joshi, et al. (1995), but attention structure demonstrates a different behaviour as supported by the analysis in (Dobnik, Ilinykh, et al., 2022).

As our model, we use the architecture of the object relation transformer proposed by Herdade et al. (2019)²⁹. This is a two-stream multi-modal transformer, which consists of three self-attention blocks, operating on the image, text and across modalities. Each block has the standard parts of the

²⁹https://github.com/yahoo/object_relation_transformer

transformer (Vaswani et al., 2017): multi-head self-attention followed by a feed-forward network, residual connection and layer normalisation.

On the vision side, the model takes the set of pre-extracted visual features of detected objects, which we receive by using the object detector released by Anderson, He, et al. (2018)³⁰ and pre-trained on Visual Genome (Krishna et al., 2017). Specifically, every object o_j in the the set of detected image objects $\mathbf{O} = (o_1, \dots, o_{|\mathbf{O}|})$ has a visual feature $v_n \in \mathbb{R}^{1 \times D}$, where $|\mathbf{O}| = 36$ and $D = 2048$. In addition, we store other outputs of the object detector, including object labels, attributes and confidence scores. They will be used in later stages to link paragraphs with objects in the image. The benefit of the object relation transformer is its ability to encode complex geometric relations between bounding boxes. Thus, we also extract the set of geometric features $\mathbf{G} = \{x, y, w, h\}$, which are fused with visual features inside the model³¹.

On the textual side, the model generates a paragraph word by word in auto-regressive fashion. Specifically, it takes the current token w_j and constructs its representation based on previously generated tokens w_1, \dots, w_{j-1} . All the future tokens in the paragraph $w_{j+1}, \dots, w_{|\mathbf{W}|}$ are replaced with the `MASK` token, framing the task as the classic next word prediction task. The generation starts with the `START` token and ends when either the maximum length of the paragraph \mathcal{L} is reached or when the `END` token is generated. As the last step, representation from two self-attention blocks are processed by the cross-attention which outputs the probability of all tokens from the vocabulary \mathcal{V} .

In terms of model’s parameters, we keep all of them untouched, thus they correspond to the original set of parameters described in Herdade et al. (2019). We train the model on the full Tell-me-more dataset, consisting of 3590 image-paragraph pairs in the train set and 410 pairs in both validation and test sets. The analysis in this paper is performed on the test set only.

³⁰<https://github.com/peteanderson80/bottom-up-attention>

³¹We refer the reader to (Herdade et al., 2019) for more details.

6.7.5. Decoding algorithms

Given the model vocabulary \mathcal{V} and \mathcal{L} as the maximum length of the generated sequence, the space of possible sequences has $|\mathcal{V}|^{\mathcal{L}}$ members, thus, becoming intractable. Rather than traversing through such space, a number of different decoding methods are used to find the most likely sequence. The most straightforward heuristics is to take the most probable word w at timestamp j until either the maximum length of the generated sequence \mathbf{w} is reached ($\mathcal{L} = 100$) or the `END` token is generated. We employ standard **greedy search**:

$$w_j = \underset{w'_j}{\operatorname{argmax}} \log p(w'_j | \mathbf{w}_{<j}, \mathbf{O}; \theta), \quad (6.43)$$

where $\mathbf{w}_{<j} = (w_1, \dots, w_{j-1})$ is the sequence of previously predicted words, $\mathbf{O} = (o_1, \dots, o_{|\mathbf{O}|})$ is the set of detected image objects and θ is the set of model parameters. Despite its simplicity and low complexity, greedy search is known for its sub-optimality on the global sentence level (Chen, Li, Cho, et al., 2018; Gu et al., 2017), often leading to generation problems such as the garden path sentence issue (Gibson, 1991).

A more popular and standardized approach is to use **beam search**, a version of the breadth-first search, that tracks multiple candidate sequences $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ and chooses the one with the highest cumulative probability score, frequently computed as summation of word scores in each sequence. Typically, the most probable sequence is picked as the final one, but other sequences can also be considered. The search starts with the word sequence $\mathbf{w}_1 = \{\text{START}\}$ and continues until the length of every predicted sequence reaches the maximum length \mathcal{L} or all of them are completed with the `END` token:

$$\mathbf{w}_j = \underset{\substack{w'_j \subseteq \mathcal{B}_j, \\ |\mathbf{w}'_j| = k}}{\operatorname{argmax}} \log p(\mathbf{w}'_j | \mathbf{w}_{j-1}, \mathbf{O}; \theta). \quad (6.44)$$

In beam search, the parameter k denotes the number of desired sequence candidates and \mathcal{B} stands for the set of sequences currently under generation. Beam search is computationally more expensive, but it is also more efficient in finding the optimal sequence due to more sophisticated exploration of the word space. However, bigger k often leads to “safe” and generic texts and candidate generations themselves can resemble each other a lot, lacking diversity (Li, Galley, et al., 2016) or becoming repetitive (Holtzman et al., 2020).

The problems of beam search have been addressed by many different approaches, mostly focused on increasing intra-set diversity of generated sequences (Kulikov et al., 2019; Meister, Forster, et al., 2021). In one of such approaches, Vijayakumar et al. (2018) propose to extend beam search by incorporating a *dissimilarity* term in the objective function. Specifically, **diverse beam search** splits beam sets into G groups W^1, \dots, W^G and at each word generation timestamp j for every sequence in the current group $\mathbf{w}_j^g \in W^g$, it encourages diversity with sequences from previous groups $W^h, h \leq g$ using a metric of dissimilarity Δ :

$$\begin{aligned} W_j^g = \operatorname{argmax}_{k \in [B']} & \sum_{k \in [B']} \log p(\mathbf{w}_{k,[j]}^g) \\ & + \lambda \sum_{h=1}^{g-1} \Delta(\mathbf{w}_{k,[j]}^g, W_{[j]}^h), \end{aligned} \quad (6.45)$$

where B' is the number of beams in each group, λ is the parameter that controls the diversity, Δ is the Hamming distance, which negatively penalises sequences sharing identical n-grams. Diverse beam search has been specifically designed to boost diversity in the multi-modal description generation task, where focus is to mimic human texts with shifts between many objects, relations and specific details. However, as reported by the authors, the best results in terms of diversity are achieved by using a simple n-gram-based heuristics, which does not take the multi-modal nature of the task into account. In addition,

diversity is encouraged between beam sets on the group level rather than between sentences within a single group, limiting the scope of diversity on the sentence level. Finally, the look-up over groups is constrained to the current word position at each generation step, shrinking the context window for the currently generated word and possibly capping the number of satisfactory generations at this timestamp.

A very different method to encourage more diverse output is to sample from the word distribution. For obvious reasons pure sampling leads to incoherent and grammatically incorrect texts. Therefore, **top- k sampling** has been proposed by Fan et al. (2018): the method cuts the probability distribution and keeps the distribution p' consisting of top k tokens with the highest probability:

$$w_j \sim \log p'(w_j | \mathbf{w}_{<j}, \mathbf{O}; \theta). \quad (6.46)$$

A known issue with the top- k sampling algorithm is that it is hard to find the optimal value for the parameter k since setting it too low could remove highly probable words or, on the contrary, keep the less probable words if it is too high.

Instead of relying on pre-defined number of tokens, **nucleus sampling** (Holtzman et al., 2020) takes words from the subset of the vocabulary in which the defined probability mass is concentrated:

$$p' = \sum_{w_j \in \mathcal{V}'} \log p(w_j | \mathbf{w}_{<j}, \mathbf{O}; \theta) \geq p, \quad (6.47)$$

where \mathcal{V}' is the top- p part of the vocabulary \mathcal{V} , in which only the words that accumulate most of the probability mass are kept. Parameter p is typically used to define the maximum value of accumulated probability. The original distribution is then re-scaled and the next word is sampled from the new distribution P :

$$P = \begin{cases} \log p(w_j | \mathbf{w}_{<j}, \mathbf{O}; \theta) / p' & \text{if } w_j \in \mathcal{V}' \\ 0 & \text{otherwise.} \end{cases} \quad (6.48)$$

The main advantage of nucleus sampling is its ability to track the shape of the probability distribution, allowing for dynamic control of the number of candidates at each timestamp. A different, but related method to introduce controlled randomness is to use **temperature scaling**. The diversity is achieved by controlling the peaks in the distribution and dividing it by the parameter τ :

$$p(w_j | \mathbf{w}_{<j}, \mathcal{O}; \theta) = \frac{\exp(\varphi_j/\tau)}{\sum_{w_j \in \mathcal{V}} \exp(\varphi_j/\tau)}, \quad (6.49)$$

where φ_j is the logit for a word w_j in the vocabulary. Lower temperatures are known to enforce the high probability events and choosing a proper value for this parameter can lead to better texts in terms of quality and diversity (Caccia et al., 2020).

We note that in this work we mainly focus on the most frequently used decoding strategies, excluding analysis of the result of more direct manipulations with texts such as length normalisation and coverage penalty (Wu, Schuster, et al., 2016), n-gram blocking or introduction of the noise model (Hill et al., 2016; Lample et al., 2018).

For our experiments with decoding algorithms, we set the following set of parameters. We set the beam size $k = 2$. Vijayakumar et al. (2018) argue that setting setting $G = k$ leads to the best results in terms of generation with diverse beam algorithm, therefore, we set $G = k = 2$ and λ equals 0.5. For top- k sampling, we try multiple values for k , aiming to investigate the impact of this parameter on generation. Specifically, we generate texts with k being the value from the following set: $\{25, 50, 75, 100\}$. For nucleus sampling, we

Metric	g	b ₂	s ₂₅	s ₅₀	s ₇₅	s ₁₀₀	st ₅₀	n ₂₅	n ₅₀	n ₉₅	db ₂
BLEU-1	37.16	30.79	33.82	34.57	33.84	33.91	36.48	34.11	34.36	33.61	37.08
BLEU-2	23.90	19.86	18.54	19.20	18.54	18.29	22.20	18.70	19.07	18.46	23.85
BLEU-3	15.53	13.13	10.07	10.77	10.09	9.99	13.67	10.30	10.67	10.25	15.51
BLEU-4	9.54	8.02	4.81	5.40	4.95	5.00	7.89	5.29	5.62	5.15	9.52
METEOR	14.22	12.97	12.53	12.79	12.53	12.46	14.00	12.67	12.80	12.58	14.20
ROUGE-L	<i>30.64</i>	30.71	<i>23.86</i>	<i>23.77</i>	23.56	<i>23.29</i>	<i>28.48</i>	<i>23.15</i>	<i>23.75</i>	<i>23.79</i>	<i>30.55</i>
CIDEr	16.62	12.30	10.48	11.30	9.78	9.54	16.51	10.54	10.76	10.56	16.80
WMD	39.80	39.10	38.40	38.41	38.17	38.06	40.26	38.28	38.34	38.33	39.84

Table 6.11. Scores of automatic metrics for different inference algorithms. The best scores per metric are in **bold**, while second best scores are in *italics*. The notation for searches should be read as follows throughout the paper: “g” - greedy, “b₂” - beam search with the width $k = 2$, “s_k” - sampling, where k is the top tokens from which the prediction is sampled, “st₅₀” - sampling from the full probability distribution with temperature scaling $\tau = 0.5$, “np” - nucleus sampling with p denoting the part of the vocabulary with the most probability mass, “db₂” - diverse beam search with the width $k = 2$.

set p to one of the following values: $\{25, 50, 95\}$. We also run pure sampling with $k = 100$ and temperature scaling with $\tau = 0.5$. Our parameters for different inference algorithms are chosen based on experiences from the corresponding research that introduces these algorithms. They also reflect our goal of evaluating how results generated by different searches can be affected by a single hyperparameter.

6.7.6. Linking

In the context of the image paragraph generation task, discourse structure in texts is affected by both text and image. To evaluate such structure, we require a mapping between object descriptions and objects in the image. While images in the Tell-me-more corpus were originally annotated with objects as part of the ADE20k corpus of house environments (Zhou, Zhao, et al., 2017), the descriptions were collected separately, hence, there are no annotations between texts and images. We decided to map noun phrases and image objects automatically, using *linking*, which is based on similarity

between object labels and noun phrases in texts³².

Primarily, linking is performed by taking both attribute and object label from the object detector and merging them into a single string, e.g. “white couch”. Next, spaCy (Honnibal et al., 2020) is used to extract noun phrases from image paragraphs, and we seek to connect each noun phrase with one of the objects in the image $o_n \in \mathbf{O}$ by embedding them both with a Sentence Transformer (Reimers and Gurevych, 2019) and comparing them based on the cosine similarity with the threshold of 0.5³³. If there are multiple similarity values that exceed these threshold for a single noun phrase, we map this phrase with the object that has the highest similarity value. Otherwise, if the noun phrase is in plural form, we map multiple objects that also share the same lemma. We perform linking for both reference texts and texts generated by each of the decoding algorithms.

6.7.7. Automatic evaluation

Table 6.11 shows scores for the most common metrics in multi-modal automatic evaluation. As we can see, greedy search and diverse beam perform the best. The worst performance is demonstrated by a variety of sampling algorithms and, somewhat surprisingly, nucleus sampling. Beam performs relatively well, achieving the highest score in ROUGE-L. When looking at the example generations in Table 6.12, we see that beam search generates very short sentences with fewer mentions of different objects, which definitely has an effect on the performance with n-gram-based metrics. Top- k sampling generally performs worse when the sampling size is increasing: CIDEr score drops to 9.54 with sampling from the full distribution. Interestingly, setting k to 50 improves the performance, indicating that this value might be the optimal one for this parameter. Nucleus sampling has a very stable performance

³²We use the linking from Dobnik, Ilinykh, et al. (2022).

³³Different methods of linking noun phrases and object labels have been evaluated in Ilinykh and Dobnik (2022).



ref	there is a very large well made bed on the left. this looks like a hotel room. two lamps are on a table on the right side. a window that you can barely see out of can be seen if you look straight ahead. there is a picture on the wall to the left and another on the right side wall.
g	this is a living room with a white couch and love seat. there is a white couch with a white and white pillow. there is a coffee table in front of the couch. there is a glass coffee table in the middle of the room. there is a picture hanging on the wall.
b2	this is a living room. there is a white couch. there is a white couch. there is a white coffee table. there is a couch.
s25	this is a tv in UNK. the couches match the couches and bedspread is white. there are two green UNK and there is a sofa table with pillows. there are two different decorative beige feet couch.
s50	a large room with a sofa for a coffee table. a four drawer rectangular coffee table sits full of glass roses loveseat. a fireplace is on the left hand wall. a large telephone is near the couch. there are two brown chairs sitting at the far end of the couch.
s75	its a formal living room. a glass sofa and ottoman. a glass table is behind the couch. the tv is behind the sofa. a patterned sofa is on the right side of the one sofa.
s100	living room with a black couch with white print pillows underneath. coffee table piece of 2 candles hang on the back wall. green coffee coffee table lobby is sitting in the middle of the couch by chair and books. decorative and an condition painted on the floor in color. light brown seat at the back end of the room.
st50	this is a living room. the couches are white. there is a white couch. there is a coffee table. there is a white couch.
n25	this room is very neat and space. the couches have orange felt UNK around the mantel. on the right side of the picture contains red flowers. theres a picture on the wall facing the sofa. theres a sofa on the wall.
n50	its a living room with a grey couch and upholstered sofa. theres a tv on the right side next to the couch. two dining chairs UNK facing the couch. theres a coffee table in front of the couch and chair. theres a glass coffee table cloth and robe hanging in the middle.
n95	there is a stone seat in the photo. the sofa is white with UNK upholstery. a beige chair and orange chair chair a round coffee table but its not sailboat.the couches fabric cover match the white accent pillows with a picture on the wall alongside them as the black chair and tan carpet.
db2	this is a living room with a white couch and love seat. there is a white couch with a white and white pillow. there is a coffee table in front of the couch. there is a glass coffee table in the middle of the room. there is a picture hanging on the wall.

Table 6.12. Example of the image and paragraphs generated with different inference algorithms.

		g			b2			s50			st50			n50			db2		
		P	S	K	P	S	K	P	S	K	P	S	K	P	S	K	P	S	K
R	BLEU_1	0.23	0.18	0.13	0.3	0.28	0.22	-0.01	-0.06	-0.03	-0.06	-0.03	-0.02	0.25	0.21	0.15	0.27	0.19	0.15
	BLEU_2	0.21	0.17	0.12	0.34	0.28	0.2	-0.04	-0.16	-0.1	-0.13	-0.15	-0.11	0.14	0.1	0.06	0.3	0.19	0.14
	BLEU_3	0.14	0.16	0.1	0.29	0.22	0.17	-0.05	-0.12	-0.07	-0.15	-0.2	-0.14	0.11	0.1	0.07	0.27	0.21	0.16
	BLEU_4	0.01	0.1	0.07	0.26	0.24	0.18	0.04	-0.1	-0.04	-0.12	-0.16	-0.12	0.19	0.11	0.08	0.2	0.22	0.16
	METEOR	-0.21	-0.18	-0.13	0.14	0.12	0.09	-0.21	-0.22	-0.16	-0.22	-0.32	-0.22	-0.05	-0.09	-0.06	-0.26	-0.24	-0.19
	ROUGE_L	0.18	0.15	0.1	0.22	0.23	0.16	0.06	0.02	0.02	-0.19	-0.22	-0.15	0.19	0.16	0.1	0.28	0.21	0.15
	CIDEr	0.02	0.15	0.1	0.33	0.17	0.12	-0.06	-0.17	-0.1	-0.15	-0.18	-0.11	0.23	0.23	0.17	0.16	0.19	0.14
C	WMD	-0.0	0.0	-0.0	0.2	0.16	0.1	-0.14	-0.09	-0.06	-0.12	-0.14	-0.1	-0.09	-0.09	-0.06	-0.14	-0.12	-0.09
	BLEU_1	0.14	0.13	0.09	0.11	0.12	0.09	0.02	0.01	0.0	0.06	0.11	0.07	0.22	0.19	0.13	0.19	0.18	0.12
	BLEU_2	0.12	0.08	0.06	0.15	0.15	0.12	-0.05	-0.12	-0.09	0.09	0.13	0.09	0.13	0.1	0.06	0.21	0.18	0.12
	BLEU_3	0.02	0.05	0.03	0.09	0.11	0.08	-0.09	-0.12	-0.09	0.11	0.13	0.08	0.12	0.1	0.06	0.18	0.18	0.13
	BLEU_4	-0.12	-0.02	-0.03	0.02	0.09	0.06	0.0	-0.13	-0.04	0.1	0.14	0.09	0.22	0.15	0.11	0.17	0.22	0.16
	METEOR	-0.15	-0.13	-0.08	0.09	0.08	0.06	-0.19	-0.17	-0.13	-0.09	-0.1	-0.07	-0.16	-0.24	-0.16	-0.27	-0.2	-0.2
	ROUGE_L	0.13	0.19	0.14	0.06	0.08	0.05	-0.07	-0.09	-0.07	-0.02	0.03	0.02	0.22	0.19	0.14	0.16	0.17	0.11
	CIDEr	0.03	0.12	0.09	0.14	0.1	0.05	-0.0	-0.01	-0.01	-0.07	0.09	0.08	0.22	0.26	0.17	0.12	0.21	0.16
F	WMD	-0.02	-0.03	-0.02	0.16	0.13	0.1	-0.22	-0.17	-0.12	-0.09	-0.07	-0.05	-0.22	-0.28	-0.21	-0.1	-0.09	-0.08
	BLEU_1	0.41	0.37	0.27	0.42	0.4	0.31	-0.22	-0.24	-0.19	0.01	0.0	0.01	0.13	0.08	0.06	0.32	0.32	0.24
	BLEU_2	0.39	0.36	0.28	0.38	0.29	0.23	-0.18	-0.27	-0.21	-0.01	-0.04	-0.03	0.07	0.05	0.03	0.31	0.31	0.22
	BLEU_3	0.29	0.32	0.23	0.35	0.25	0.19	-0.22	-0.25	-0.18	0.01	-0.0	0.0	0.12	0.07	0.05	0.3	0.3	0.22
	BLEU_4	0.15	0.24	0.18	0.23	0.2	0.14	-0.01	-0.17	-0.12	0.03	0.05	0.0	0.19	0.06	0.04	0.22	0.24	0.17
	METEOR	-0.07	-0.07	-0.08	0.12	0.09	0.06	-0.11	-0.14	-0.1	-0.0	-0.06	-0.03	-0.12	-0.2	-0.16	-0.01	-0.01	-0.01
	ROUGE_L	0.31	0.29	0.22	0.24	0.24	0.18	-0.06	-0.1	-0.08	-0.08	-0.07	-0.05	0.16	0.12	0.09	0.28	0.29	0.19
	CIDEr	0.16	0.27	0.2	0.36	0.27	0.21	-0.35	-0.37	-0.28	0.01	0.02	0.02	0.02	0.1	0.07	0.13	0.29	0.23
	WMD	0.04	0.03	0.02	0.19	0.22	0.14	-0.14	-0.16	-0.12	-0.03	-0.02	-0.03	-0.02	-0.03	-0.03	0.1	0.05	0.03

Table 6.13. Correlation scores between automatic metrics and human judgements across three criteria. R, C and F on the left side stand for relevance, correctness and composition (flow), corresponding to the type of questions that the crowdworkers were provided with. P, S and K stand for Pearson’s, Spearman’s and Kendall’s correlations. We report correlation scores per search and per correlation metric. The scores coloured in red have $p < 0.05$.

with n50 showing the best scores. We note that temperature scaling has a huge positive impact on the scores of sampling algorithm, pushing it towards the performance of greedy and diverse beam search. This might indicate that sampling and its randomness can be successfully controlled with the proper value for temperature.

The reason for a high performance of greedy search could be its ability to generate the “safest” words combined with the simplicity of the images and the lack of surprisal in them. For example, images in the dataset correspond to standard room types and thus contain standard objects. This is a blessing if we care about model’s ability to generalise over the house environments, but also a curse since occasionally a highly salient and surprising object might appear in the images and searches will not be able to describe it. We believe that our results show the inadequacy of automatic metrics in measuring the sensitivity of inference algorithms to the type of objects and their salience.

6.7.8. Human evaluation

To support our hypothesis that automatic metrics are not enough to measure fine-grained differences between various decoding algorithms, we conduct a human evaluation on Amazon Mechanical Turk. We randomly sample 10% of images from the test set, which equals 41 items. For each of these images, we take generated texts from the top-6 decodings based on the CIDEr score. We get 287 different image-text pairs to evaluate. During the evaluation, we provide workers with an image and its description and ask them to answer 3 (three) different questions, aiming to evaluate (i) relevance: does the text describe relevant and essential objects, (ii) correctness: does the text describe objects correctly (e.g., using correct words), (iii) composition: do object descriptions naturally follow each other. The example item for human evaluation is shown in Appendix 6.7.14. Each judgement is a score on a scale between 1 and 5, where 1 is the lowest rank. We collect three different judgements per item and average them. We pay 0.17 US dollars for a single assignment and restrict the location of the workers to the US, the UK, Canada, Ireland or Australia. We also ran our experiments with Master workers only (25 different human participants). We follow Kilickaya et al. (2017) and compute three different correlation scores: Pearson’s correlation, Spearman’s rank correlation and Kendall’s correlation.

The correlation scores are presented in Table 6.13. In general, sampling-based methods do not significantly correlate with automatic metrics or correlate but negatively. More controlled decodings, such as greedy or beam search, correlate with automatic metrics more, especially for the composition question (F). This indicates that automatic metrics correlate more with decodings that introduce less randomness. Future work will need to examine whether randomness and diversity in such searches as top- k sampling is a suitable type of diversity since it is unclear from correlation scores alone. In terms of the relevance of objects, sampling with temperature generally has negative scores (similar to other sampling-based methods). Still, a significant negative

	ref	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	2.9	0.9	0.3	1.0	1.1	1.1	1.0	0.9	1.1	1.1	1.1	0.9
s2	1.6	1.7	1.5	1.7	1.7	1.8	1.8	1.7	1.8	1.8	1.7	1.8
s3	1.4	1.6	1.4	1.7	1.8	1.8	1.8	1.6	1.8	1.8	1.8	1.6
s4	1.3	1.6	1.4	1.8	1.8	1.8	1.9	1.6	1.8	1.9	1.8	1.6
s5	1.2	1.7	1.4	1.8	1.8	1.8	1.7	1.6	1.7	1.8	1.7	1.7

Table 6.14. Average number of noun phrases generated by different inference algorithms. The numbers are given per sentence.

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	200.0	175.0	227.9	215.0	233.3	231.0	208.0	227.9	209.2	228.2	200.0
s2	205.2	215.8	207.6	215.8	218.1	220.1	207.5	228.2	231.1	206.1	206.5
s3	210.8	205.2	213.8	215.0	233.3	203.8	210.3	202.6	219.3	207.1	207.2
s4	197.4	196.0	216.4	206.5	200.0	208.9	208.9	205.9	201.3	216.4	197.5
s5	198.0	212.5	202.6	205.1	211.3	214.4	197.3	209.6	215.8	208.7	200.0

Table 6.15. Average proportion of noun phrases (in percent) when *more* are generated than present in the references.

correlation is found only with Spearman’s rank correlation for METEOR. Beam, however, might produce more relevant objects as demonstrated by high correlation in terms of BLEU_2 and CIDEr. We do not observe any correlation for the correctness criterion. On the contrary, text composition (flow) shows that more controlled decodings correlate considerably more with human judgements, especially when looking at n-gram metrics. This might demonstrate that more specific automatic metrics better reflect whether the object descriptions naturally follow each other. Overall, we show that while most of the automatic metrics are not sufficient in providing us with information about the salience and correctness of object descriptions for many different decoding algorithms, their scores, somewhat surprisingly, might still tell us about the sentence-level discourse and flow of object descriptions.

6.7.9. Non-grounded evaluation

Next, we will look at the surface level of noun phrases and examine faithfulness of generated texts to the reference ones. Noun phrases in image descriptions

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	20.3	8.5	18.8	18.9	19.6	18.6	18.9	19.6	18.8	20.7	20.3
s2	45.3	44.2	42.6	42.9	40.4	43.1	45.9	39.8	40.0	39.7	45.5
s3	45.9	45.6	42.0	45.9	44.4	44.5	45.5	42.7	41.2	44.4	47.1
s4	46.5	46.8	43.9	45.8	43.3	43.4	47.0	45.0	43.3	42.9	46.6
s5	49.3	43.7	41.4	42.5	37.6	40.7	46.2	38.3	40.0	39.9	49.0

Table 6.16. Average proportion of noun phrases (in percent) when *fewer* are generated than present in the references.

typically depict image objects, thus we believe that direct comparison of noun phrases in different texts can help us to understand how much each decoding algorithm learns on the surface of descriptions. Table 6.14 shows the average number of noun phrases in each sentence across different searches and references. We see that there is a gradual decrease in the number of noun phrases in references throughout the paragraph. Such decrease is not observed in texts generated by all algorithms. On the contrary, the first sentence typically has the fewest number of noun phrases generated with other sentences containing mostly the same number. This could be a sign that on the surface level decoding algorithms do not capture discourse structure, reflected in gradual decrease of the number of noun phrases. Instead, search algorithms tend to generate the same number of noun phrases across sentences, treating each sentence equally.

We also observe that the algorithms generate more noun phrases per sentence than required rather than generate fewer of them. Specifically, across all image-paragraph pairs a fewer number of noun phrases is generated for 757 sentences, a bigger number for 955 sentences and the exact number as in the references was produced for 493 sentences. To closer identify the impact of over- and under-generation of noun phrases, we compute proportion of noun phrases for both cases. As Table 6.15 demonstrates, all searches tend to generate nearly two times more noun phrases than required in each sentence. The picture changes when the searches under-generate. According to Table 6.16, while most of the sentences lack at least half of the required noun

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	0.18	0.10	0.10	0.13	0.11	0.08	0.14	0.12	0.10	0.13	0.18
s2	0.17	0.17	0.13	0.13	0.13	0.14	0.17	0.12	0.15	0.13	0.17
s3	0.13	0.12	0.10	0.10	0.09	0.11	0.11	0.10	0.11	0.10	0.13
s4	0.10	0.09	0.10	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.10
s5	0.10	0.10	0.07	0.07	0.08	0.07	0.08	0.07	0.07	0.08	0.10

Table 6.17. Dice similarity coefficient between the set of objects described in reference texts and texts generated by different decoding algorithms. The values are provided per sentence and averaged across all image-paragraph pairs.

phrases (in terms of quantity), the first sentence is affected the most by under-generation. Coupled with the results in Table 6.14, we conclude that decoding algorithms do not learn the structure of discourse on the simplest surface level of descriptions reflected in the differences in the number of noun phrases. This result indicates that searches might generate a discourse that is different from the one observed in references. In the following analysis, we will move from the surface level to the grounding level, in which we will examine if the noun phrases that are generated can be linked with image objects. We will also compare whether the objects described by different searches overlap with the ones found in reference texts.

6.7.10. Grounded evaluation

Table 6.17 shows the degree of overlap between two object sets: the first set includes objects described in references, while the second set contains objects mapped with noun phrases in generated texts from different decodings algorithms. We use Sørensen–Dice coefficient $\frac{2|A \cap B|}{|A| + |B|}$ to measure the overlap. The closer the result to 0, the less overlap is present. The results demonstrate that searches describe a very different set of objects rather than the one mentioned in the references. The highest overlap is observed with greedy search and diverse beam. The scores indicate that either a different and correct set of objects is described or the noun phrases cannot be linked with objects because

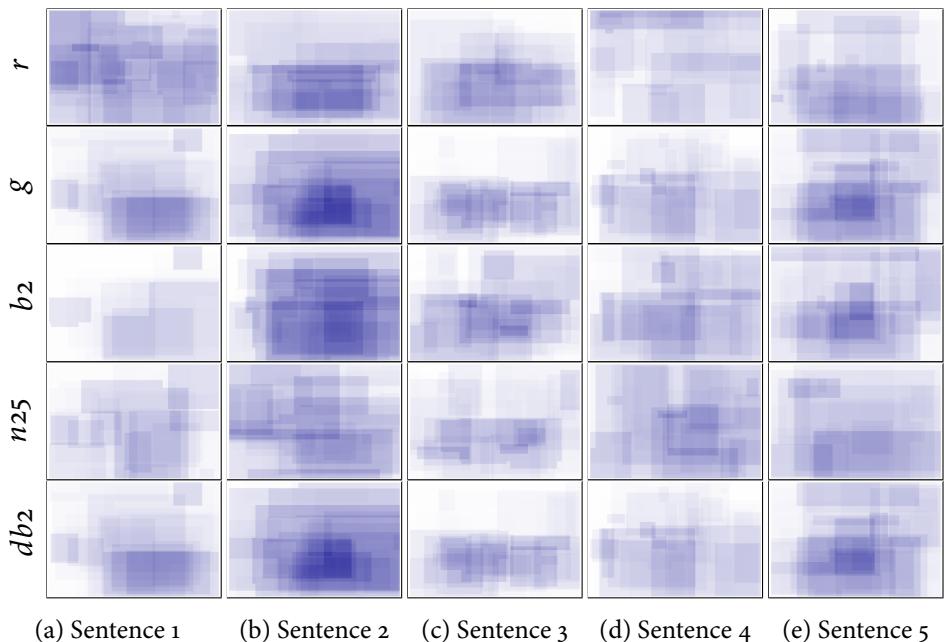


Figure 6.35. Attention heatmaps over objects, described in texts according to the results of linking. Results are shown per sentence and per search. The first row denotes attention in reference texts. We aggregate heatmaps across all images into the single image, therefore, darker colour denotes higher focus on the specific area in the image.

they are incorrect (could also be because of high randomness, leading to the lack of grammaticality).

We examine whether noun phrases in generated texts can be linked with any of the objects in the image. Table 6.18 shows the proportion of successful linking once we link noun phrases with image objects using cosine similarity. We set the similarity threshold to 0.5: if the similarity between the object label and noun phrase is higher than this value, we decide that this noun phrase is faithful to the image and can be grounded.

The results demonstrate that half and more of the generated noun phrases can be linked with objects in the image. In general, sampling algorithms generate fewer number of grounded noun phrases, possibly due to the increased randomness. Greedy search, beam and diverse beam generate the highest

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
s1	69.5	72.7	46.5	46.3	43.2	46.1	66.5	51.1	45.2	50.5	69.5
s2	65.6	65.1	47.2	49.0	43.8	46.6	58.8	44.7	50.3	47.7	65.5
s3	61.6	59.5	43.6	46.9	40.5	45.1	53.1	40.1	40.6	44.7	60.7
s4	55.4	57.6	43.7	42.7	44.7	41.0	52.5	45.0	43.7	38.3	55.7
s5	60.5	57.4	47.6	43.2	43.4	43.7	53.3	39.2	38.9	44.4	59.3

Table 6.18. Average proportion of successful linking (in percent) between noun phrases in generated texts and image objects.

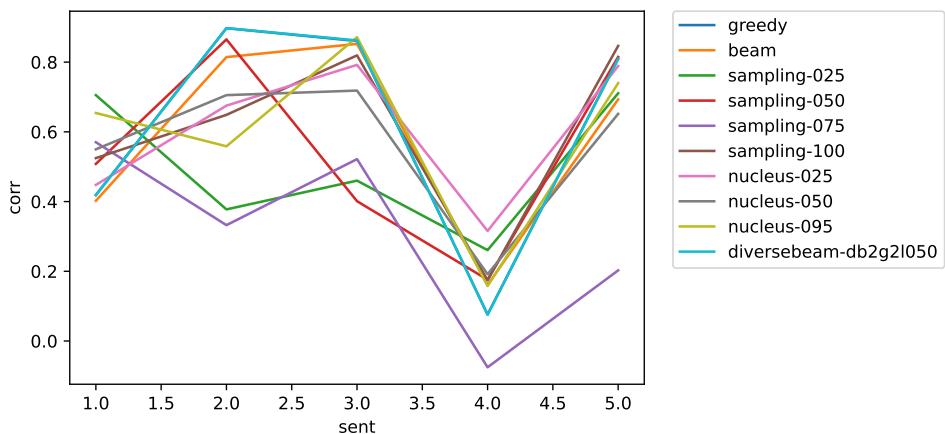


Figure 6.36. Correlation between heatmaps for different searches and reference paragraphs. X-axis is sentence in the paragraph (1-5), Y-axis is the correlation coefficient, pixel-by-pixel correlation between attention heatmaps.

number of noun phrases which are truthful to the image. We believe that while reference-correctness of generated texts can get worse, inference algorithms are still able to generate alternative descriptions of images which can be grounded. However, the structure of discourse reflected on the surface level and the level of grounding might not necessarily correspond to the one observed in references. In the next experiment, we look at the problem under the angle of attentional structure and examine spatial arrangement of linked objects and how these arrangements differ between decoding algorithms.

6.7.11. Attentional structure of discourse

Figure 6.35 demonstrates a number of the attention heatmaps across areas in the image for different sentences. At first glance, different inference algorithms look at similar locations in the image and also focus on parts which are attended by humans. However, there are relatively more areas described in the first sentence of the references, while a much smaller and fewer areas are described in generated texts. This could be directly related to the fewer number of objects and under-generation discussed previously. The second and third sentences describe specific areas of the image in all cases, mostly central ones. Interestingly, greedy and diverse beam have highly similar attention across the image. In sentence 4, human attention disperses over the full scene, while it is unclear whether the same pattern happens in generated texts. This could signal a possible *topic shift*, happening in later parts of the paragraph and inability of searches to capture that. To understand the differences on the level of sentences better, we measure the correlation between flattened heatmaps pixel by pixel. We use Pearson product-moment correlation coefficient which can be applied to images across the channels. The results are shown in Figure 6.36. As we can see, in sentence 4 attentional structure on the image differs between searches and references, supporting the idea of topic shift. Sampling methods have the lowest correlation with the references, while nucleus with $p = 25$ is affected the least in sentence 4. Note that the correlation in the first sentence is lower than in the second and the third one for most of the searches. This could be related to the importance of the first sentence and a bigger number of noun phrases in it, which are not generated during the decoding stage.

6.7.12. Conclusion

In this paper we described our analysis of how decoding strategies structure discourse in multi-modal longer image descriptions. We performed evaluation using intuitions from different evaluation perspectives: automatic,

surface-based (non-grounded), image-based (grounded) and attention-based. The results suggest that for the task of image paragraph generation decoding algorithms diverge from humans in generating specific type of discourse. Although they might generate reference-incorrect but image-correct descriptions, it is unclear what kind of discourse is generated in the end. In general, algorithms which are less random construct discourse similar to the one in human references, while sampling-based methods generate a different type of discourse, which is hard to control for. We plan to use the insights described in this paper and build a metric that would evaluate the structure of longer image paragraphs, reflected in *both* object and relation descriptions as this is currently a much needed evaluation measure.

6.7.13. Limitations

There are several directions which can support the analysis in this paper. First, the automatic linking is not a perfect mechanism, prone to errors. The method that we use works better for shorter phrases which share the same lemmas and thus are less ambiguous. Second, using more models (Li, Zhu, et al., 2019) or more datasets (Krause et al., 2017) would potentially give us a broader picture of the type of discourses formed by humans and quality of representations used during decoding phase. We also consider our analysis preliminary with the opportunity of developing a separate metric to evaluate discourse in longer image descriptions.

6.7.14. Appendix A

First, read the instructions ("Instructions" in the top-left corner).
Have a look at both short and detailed instructions.
If you don't follow them, we have a right to reject your submission.

this is a living room with tan walls and white ceiling . there is a large black couch and brick wall . there is a golden hutch in the front part of the couch. there is a brown night table in the back room with a white chair at the center of the room . there are plants on the ceiling by a chandelier hanging above the couch .

How well do you agree with the following statements?

Relevance: does the text describe relevant and important objects?

Correctness: does the text describe objects correctly (e.g., using correct words)?

Composition: do objects descriptions naturally follow each other?



Figure 6.37. The example item for the workers on AMT for human evaluation.

6.8. Describe Me an Auklet: Generating Grounded Perceptual Category Descriptions

6.8.1. Abstract

Human speakers can generate descriptions of perceptual concepts, abstracted from the instance-level. Moreover, such descriptions can be used by other speakers to learn provisional representations of those concepts . Learning and using abstract perceptual concepts is under-investigated in the language-and-vision field. The problem is also highly relevant to the field of representation learning in multi-modal NLP. In this paper, we introduce a framework for testing category-level perceptual grounding in multi-modal language models. In particular, we train separate neural networks to **generate** and **interpret** descriptions of visual categories. We measure the *communicative success* of the two models with the zero-shot classification performance of the interpretation model, which we argue is an indicator of perceptual grounding. Using this framework, we compare the performance of *prototype*- and *exemplar*-based representations. Finally, we show that communicative success exposes performance issues in the generation model, not captured by traditional intrinsic NLG evaluation metrics, and argue that these issues stem from a failure to properly ground language in vision at the category level.

6.8.2. Introduction

Grounded language use links linguistic forms (symbols) with meaning rooted in various perceptual modalities such as vision, sound, and the sensory-motor system (Harnad, 1990). But grounding is not merely a solipsistic mapping, from form to meaning; rather, it results from a communicative context in which linguistic agents act on — and have goals in — the real world (Chandu et al., 2021; Giulianelli, 2022; Larsson, 2018). Large language models trained on vast amounts of text have been criticised for lacking grounded representations



Figure 6.38. A simple learning scenario in which one speaker learns a visual concept from another speaker’s description; the learner is then able to use their provisional representation to classify an entity as belonging to the concept.

(Bender and Koller, 2020; Bisk et al., 2020), and the fast-growing field of multi-modal NLP has been working to address this problem (Beinborn et al., 2018; Bernardi et al., 2016). However, multi-modal models have several areas for improvement. Recent work suggests that these models are affected by the distribution of items in training data, often over-representing specific scenarios and under-representing others (Agrawal, Batra, Parikh, and Kembhavi, 2018). This, in turn, affects their ability to find a true balance between the levels of granularity in descriptions for novel concepts, as these models are expected to generalise (Hupkes et al., 2023). As a result, these models rely excessively on text and have to be supplied with various mechanisms, enforcing and controlling their attention on modalities such as vision (Ilinykh, Emampoor, et al., 2022; Lu, Xiong, et al., 2017; Thomason et al., 2019). This raises questions about the nature of the relationship these models learn between linguistic and non-linguistic information.

Exploiting statistical regularities in multi-modal datasets can cause models to *hallucinate*. According to Rohrbach et al. (2018), neural image captioning

systems can accurately describe objects in images but struggle to understand the overall situation, often relying on common contextual patterns associated with specific objects that co-occur. Similar problems are common for other multi-modal models, datasets (Alayrac et al., 2022), and tasks such as Visual Question Answering (Antol et al., 2015) and Embodied Question Answering (Das, Datta, et al., 2018). These examples, along with many others, illustrate that perceptual grounding cannot be achieved in the abstract but must be considered in a *communicative context*, which includes speakers' prior *common ground, joint perception, and intentions* (Clark and Wilkes-Gibbs, 1986). One important type of common ground is shared perceptual world knowledge, which need not necessarily rely on the immediate perceptual context. For instance, if someone mentions that *red apples are sweeter than green ones*, this communicates something, even to someone who is not concurrently looking at or tasting apples. We can acquire and use a (provisional) perceptual concept based on a natural language description produced by a conversation partner, a process referred to as *fast mapping* (Carey, 1981; Gelman and Brandone, 2010). *Can multi-modal language models generate a description of a perceptual category that similarly communicates the concept to an interlocutor?*

In this paper, we propose **perceptual category description**, which emphasises *category-level* grounding in a communicative context. This framework models a *simple* interactive scenario (Figure 6.38) where (1) a describer, referred to as **GEN**, generates a description of one or more visual categories, (2) an interpreter, **IPT**, learns from the generated descriptions, and (3) classifies among both the seen classes, which it already has knowledge of, and the unseen classes described by **GEN**. During training, the **GEN** model has access to images and class labels from both the seen and “unseen” sets, but only receives supervision on ground-truth *descriptions* from the seen set. This ensures that during testing the generator is evaluated based on its ability to use category-level representations of the unseen classes, rather than memorising descriptions from the training data. The **IPT** model only has

access to instances from seen at train time and performs zero-shot image classification on unseen instances using descriptions produced by GEN as auxiliary class information. Zero-shot learning from text descriptions is not a novel task; our focus in this work is on the generation of perceptual category descriptions, using “communicative success”—the performance of the IPT model—as a semi-extrinsic evaluation metric. The proposed evaluation method differs from many standard automatic generation evaluation metrics, such as BLEU (Papineni et al., 2002), which are not designed to capture the level of communicative usefulness of the generated texts. In contrast to many language-and-vision tasks, we explore the ability of multi-modal models to perform grounding on class-level representations, distinct from instance-level representations, e.g. images.³⁴ Additionally, we highlight the issue of mismatch between intrinsic evaluation (generation metrics) and task-based evaluation, as indicated by the performance of the IPT. Our results reveal challenges involved in developing better models with the ability to ground at the class level. We believe that our fine-grained analysis of the task, data and models sheds light on the problems associated with both generating and interpreting class-level image descriptions. We also contribute insights into the extent to which current evaluation methods for generated texts consider communicative context. The framework that we propose can be used for evaluating existing models of language grounding and can also aid in building new multi-modal models that perform grounding in communication. To support research in this direction, we have made our code and data available here: <https://github.com/GU-CLASP/describe-me-an-auklet>.

6.8.3. Background

Prototypes and exemplars Cognitive theories of categorisation are psychologically-motivated accounts of how humans represent perceptual concepts

³⁴See, for example, Bernardi et al. (2016) which presents a survey of image description techniques that rely heavily on the image as part of the input.

and use them for classification. Such theories have challenged the assumption that categories can be defined in terms of a set of necessary and sufficient features. In contrast, they try to account for phenomena like *prototypically effects*, in which certain members of a category are perceived as more representative of the class than others. In *prototype theory*, cognitive categories are defined by a **prototype**, an abstract idealisation of the category. Membership in the class, then, is judged in reference to the prototype (Rosch, 1975a). In *exemplar theory*, (e.g., Medin and Schaffer, 1978; Nosofsky, 1984), concepts are still defined in relation to an ideal, but this time the ideal is an **exemplar**, which is a particularly representative *member* of the very category. Put another way, an exemplar is *of the same kind* as the other members of the category, whereas prototypes, in general, are not. Experimental evidence suggests that humans employ both exemplar and prototype-based strategies (Blank and Bayer, 2022; Malt, 1989).

Perceptual categories play a role in natural language interpretation and generation. In fact, *classifier-based meaning* has been proposed as a way to ground language in perception (Schlangen et al., 2016; Silberer, Ferrari, et al., 2017). There are both formal and computational interpretations of this approach that support compositional semantics for lexical items with classifier-based perceptual meanings (Kennington and Schlangen, 2015; Larsson, 2013). In this paper, we explore how classifier-based meaning facilitates the generation of class-level descriptions by testing three different GEN model architectures: one motivated by prototype theory, one by exemplar theory, and one that uses a hybrid approach.

Zero-shot language-and-vision generation and classification In the perceptual category description framework, both models operate with textual descriptions: one generates them, and the other interprets them. The interpretation model performs zero-shot classification, with (in this case) vision

as the *primary modality* and text as the *auxiliary modality*.³⁵ In zero-shot learning scenarios that use text as auxiliary data, the quality and relevance of the text has been shown to improve model performance. For example, perceptually more relevant texts might help better learning of novel concepts (Paz-Argaman et al., 2020). Bujwid and Sullivan (2021) show that Wikipedia texts can be used as class descriptions for learning a better encoding of class labels. In a similar vein, Desai and Johnson (2021) demonstrate that for a nominally non-linguistic task (e.g. classification), longer descriptions yield better visual representations compared to labels. Image classification can be further improved with a better mapping between visual and linguistic features (Elhoseiny, Zhu, et al., 2017; Kousha and Brubaker, 2021).

Innovative language use can be resolved by taking familiar representations and mapping their components to a new context (Skantze and Willemsen, 2022). Suglia et al. (2020) and Xu, Kordjamshidi, et al. (2021) develop models that recognise out-of-domain objects by learning to compose the attributes of known objects. Also, the descriptiveness and discriminativeness of generated class description influences their utility for interpretation purposes (Chen, Ji, et al., 2018; Vedantam, Bengio, et al., 2017; Young et al., 2014). We partially explore this phenomenon in our experiments; see Section 6.8.5.2.

Language games in a multi-agent setup Our setup with two neural networks is somewhat analogous to the idea of a multi-agent signalling game (Lewis, 1969). While the idea of multiple agents developing their language to solve tasks has been extensively studies in NLP (Choi et al., 2018; Lazaridou et al., 2017), our work differs in that we do not have a direct learning signal between the models, e.g. the agents are not trained simultaneously. Therefore, our models do not cooperate in a traditional sense. Instead, we focus

³⁵This means that the model has supervised training with visual examples of seen classes, and then the model receives text descriptions (one per class) corresponding to the unseen classes. The model is then evaluated in the generalised zero-shot setting. I.e., to classify new visual examples among both seen and unseen classes. See Xian et al., 2020 for an introduction to different zero-shot learning setups and a recent survey of the field.

on developing a more natural and complex multi-network *environment* by incorporating insights from research on human cognition, perceptual grounding, and communication. In particular, we (i) explore the ability of neural language models to learn high-level representations of visual concepts, (ii) generate and evaluate concept descriptions based on these representations, and (iii) assess the performance of a separate network in interpreting these descriptions for zero-shot classification.

In related work, Zhang, Hare, et al. (2018) train an interpreter and a speaker to perform continuous learning through direct language interaction. In contrast, our setup is more straightforward as the describer does not receive feedback from the interpreter. Another study by Elhoseiny, Saleh, et al. (2013) proposes learning novel concepts without visual representations. They use encyclopedic entries as alternative information sources when perceptual input is unavailable. Our approach presents a greater challenge as humans often lack access to textual corpora when interacting in the world. Patel and Pavlick (2022) investigate the ability of pre-trained language models to map meaning to grounded conceptual spaces. We are similarly interested in grounding in a structured space of related concepts, but our setup is different, proposing the semi-interactive task of grounded category description, rather than probing models for their ability to generalise.

6.8.4. Models

At a high level, GEN and IPT each have two connected modules: an image classifier, and a grounded language module. Both networks learn visual representations which are shared between the classification and language tasks. During training, IPT learns to *interpret* textual descriptions of seen classes by mapping them into its visual representation space. If it generalises well, textual descriptions of unseen classes should then be mapped to useful visual representations at test time, even though no images of unseen classes were

available during training. Contrariwise, GEN is trained to *generate* descriptions of seen classes based on its visual representation of those classes. At test time, GEN must extrapolate to generating descriptions of unseen classes, for which no ground-truth descriptions were provided during training.

6.8.4.1. Label embedding classifier

Both models use a *label embedding classifier* that represents classes as embeddings. The embedding matrix $\mathbf{V} \in \mathbb{R}^{N \times D}$, stores visual concept representations, with $N = 200$ being the number of classes and $D = 512$ indicating the size of each single class representation vector.³⁶ The class embedding parameters (\mathbf{V}_G for GEN model and \mathbf{V}_I for IPT) are shared between the classification module and language module within each model (no parameters are shared between GEN and IPT). Both models use ResNet visual features, with a size of 2048 provided by Schönenfeld et al. (2019) as inputs to the classifier. These features were extracted from the standard ResNet-101 trained on the ImageNet 1k dataset (Russakovsky et al., 2015). In the following, $\mathbf{x} = \text{ResNet}(\mathbf{x})$ is the encoding of the input image \mathbf{x} .

The classifiers are simple two-layer feed-forward networks trained on the multi-class classification task. Visual features of the input, \mathbf{x} , are concatenated with each class vector \mathbf{v}_i from \mathbf{V} before being passed through the network. Consequently, the network produces N scores that are transformed into class probabilities $\hat{\mathbf{y}}$ using a **softmax** function σ applied along the label dimension:

$$\hat{\mathbf{y}} = \sigma((f_2(f_1(\mathbf{x}) \oplus \mathbf{v}_i))_{i \leq N}), \quad (6.50)$$

³⁶We also initialise GEN with $N = 200$ for convenience, but the labels corresponding to the 20 unseen classes are quickly disregarded during supervised training since they never appear in the training data.

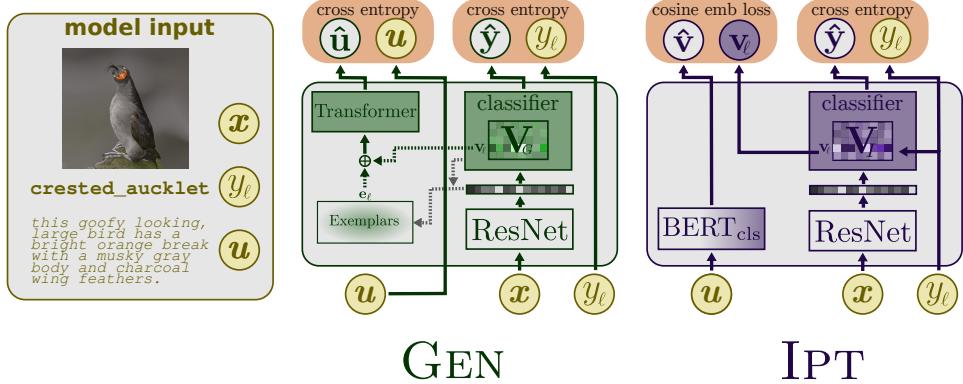


Figure 6.39. Training inputs and describer (left) and interpreter (right) model architectures. In GEN, the dotted lines indicate options for the type of class representation ($\mathbf{c}_\ell^{\text{prot}}$, $\mathbf{c}_\ell^{\text{ex}}$, or $\mathbf{c}_\ell^{\text{both}}$). If exemplars are used, they are updated based on the classifier at the end of each epoch (gray dotted lines), as described in equation 6.54.

where

$$f_1(\mathbf{x}) = \text{ReLU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1), \quad (6.51)$$

$$f_2(\mathbf{x}') = \mathbf{W}_3(\text{ReLU}(\mathbf{W}_2 \mathbf{x}' + \mathbf{b}_2)) \quad (6.52)$$

where $\mathbf{W}_1 \in \mathbb{R}^{2048 \times h_1}$, $\mathbf{W}_2 \in \mathbb{R}^{(h_1+D) \times h_2}$, and $\mathbf{W}_3 \in \mathbb{R}^{h_2 \times 1}$ is the classification output layer.

Both GEN and IPT use $h_1 = 256$ and $h_2 = 128$.

6.8.4.2. Generation model

The generation model has two modules: the classifier described in §6.8.4.1, and a *decoder* that generates text from a class representation. Given a label y_ℓ , the decoder generates text by using the class representation, \mathbf{c}_ℓ , corresponding to the label. The class representation is computed differently depending on whether the model uses prototype class representations, exemplars, or both:

GEN-PROT simply takes the corresponding row of the label embedding \mathbf{V}_G , which is also used for classification.

$$\mathbf{c}_\ell^{\text{prot}} = \mathbf{v}_\ell = \mathbf{V}_G[\ell] \quad (6.53)$$

GEN-Ex keeps an additional cache of exemplar image features (one per class) which change after each training epoch. The exemplar image for class ℓ is computed as the image that is most certainly part of that class, according to the classifier:

$$\mathbf{c}_\ell^{\text{ex}} = \mathbf{e}_\ell = \arg \max(\{\hat{\mathbf{y}}[\ell] \mid \mathbf{x} \in X\}) \quad (6.54)$$

GEN-BOTH uses the concatenation of the prototype and exemplar representations:

$$\mathbf{c}_\ell^{\text{both}} = \mathbf{v}_\ell \oplus \mathbf{e}_\ell \quad (6.55)$$

We train a standard transformer decoder to generate class descriptions (Vaswani et al., 2017). GEN models differ only in the type of input representations provided to the decoder. At each timestep, t , the model’s input is updated with previously generated tokens (w_1, \dots, w_{t-1}) and the current token w_t is predicted. We use a standard setup for the transformer: six self-attention layers with eight heads each. The model is trained for 20 epochs in a teacher forcing setup. The learning rate is set to 4×10^{-4} . The best model is chosen based on the CIDEr score (Vedantam, Lawrence Zitnick, et al., 2015) on the validation set using beam search with a width of 2.

Both the classifier and the decoder are trained jointly with the standard cross-entropy loss:

$$\text{loss}_G = \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbb{1}(y_\ell)) + \sum_{i < t} \text{CrossEntropy}(\hat{\mathbf{u}}_i, \mathbb{1}(u_i)), \quad (6.56)$$

$\hat{\mathbf{y}}$ is the output of the classifier, y_ℓ is the ground-truth label, $\hat{\mathbf{u}}_i$ output of the decoder at position i , and u_i is the ground-truth token. For inference, we explore multiple decoding algorithms which we describe below.

6.8.4.3. Decoding algorithms

In our describer-interpreter setup, the quality of the generated texts, particularly their content, is of importance. Quality generation depends heavily on the decoding algorithm used to select tokens. “Safer” algorithms may generate more accurate texts, but with poor discriminativity, while other algorithms introduce some degree of randomness, which promotes diversity (Zarrieß et al., 2021). We examine *two* decoding algorithms, with introduce different conditions for text accuracy and diversity. While greedy search can generate accurate descriptions, it is sub-optimal at the sentence level, e.g. longer generation become repetitive and ”boring“ (Gu et al., 2017).

Beam search is often used as a standard decoding algorithm because it suffers much less from the problems occurring during long-text generation. At each generation step i , it keeps track of several candidate sequences $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_k)$ and picks the best one based on the cumulative probability score of generated words per sentence:

$$\mathbf{c}_i = \underset{\substack{\mathbf{c}'_i \subseteq \mathcal{B}_i, \\ |\mathbf{c}'_i| = k}}{\operatorname{argmax}} \log p(\mathbf{c}'_i \mid \mathbf{c}_{i-1}, \mathbf{v}_i; \theta). \quad (6.57)$$

The parameter k is used to control the depth of the search tree, and \mathcal{B} is the set of candidate sequences. While beam search generally outperforms

greedy, higher k can lead to texts with low diversity (Li, Galley, et al., 2016). To evaluate whether "more diverse" means "more descriptive" in the context of our two-agent set-up, we generate texts with **nucleus sampling** method (Holtzman et al., 2020) which samples tokens from the part of the vocabulary defined based on the probability mass:

$$p' = \sum_{w_i \in \mathcal{V}'} \log p(w_i | \mathbf{w}_{<i}, \mathbf{v}_i; \theta) \geq p, \quad (6.58)$$

where p determines the probability mass value, while \mathcal{V}' is part of the vocabulary \mathcal{V} which accumulates the mass at the timestamp i . Next, a new distribution P is produced to sample the next token:

$$P = \begin{cases} \log p(w_i | \mathbf{w}_{<i}, \mathbf{v}_i; \theta)/p' & \text{if } w_i \in \mathcal{V}' \\ 0 & \text{otherwise.} \end{cases} \quad (6.59)$$

With nucleus sampling, we aim to generate more diverse texts than those generated with beam search. By evaluating the interpreter with texts generated by different algorithms, we consider the impact of generation on the success of information transfer from the describer to the interpreter.

6.8.4.4. Interpretation model

The IPT model has two modules: a label embedding classifier with a weight matrix $\mathbf{V}_I \in \mathbb{R}^{N \times D}$, and an interpretation module that maps texts to vectors of size D . IPT uses [CLS] token vectors extracted from BERT as text features. In preliminary experiments on the ground-truth test data, we observed significant improvements in the performance of IPT by using features from a BERT model (Devlin, Chang, et al., 2019) which was fine-tuned on descriptions from the seen portion of the training set. We fine-tuned the final layer with a learning rate of 2×10^{-5} and weight decay of 0.01 for 25 epochs using the

Adam optimiser (Kingma and Ba, 2015). The model was fine-tuned using a text classification task involving the seen classes. Since BERT is not visually grounded, we speculate that the pre-training task may assist the model in attending to visually relevant information within the descriptions, leading to a more informative [CLS] representation. Given a text description \mathbf{u} , we use \mathbf{u} to denote the [CLS] features (with size 768) extracted from the fine-tuned BERT model.

The interpretation module is defined as follows:

$$\hat{\mathbf{v}} = \text{Tanh}(\mathbf{W}\mathbf{u} + \mathbf{b}) \quad (6.60)$$

where $\mathbf{W} \in \mathbb{R}^{768 \times D}$ and $\mathbf{b} \in \mathbb{R}^D$.

Given a training example $(\mathbf{x}, y_\ell, \mathbf{u})$, the classifier makes a class prediction \hat{y} from \mathbf{x} and the interpreter predicts the class representation $\hat{\mathbf{v}}$ from \mathbf{u} . Our objective is to improve both on the class predictions and class representations produced by the IPT model. To evaluate the class prediction, we compare it to the ground-truth class label y_ℓ . As for the class representation, the training objective encourages the model to predict a position in the vector space with is close to the target class, ℓ , and far from randomly selected negative classes. We employ the following sampling strategy. We draw a vector \mathbf{v}_k from \mathbf{V}_I so that with a frequency of 0.5, it is a negative sample (i.e., $k \neq \ell$) and the other half the time $k = \ell$.

The two modules are trained jointly. The loss term for the classifier is computed with the standard cross-entropy loss and the term for the interpreter is computed with the cosine embedding loss, a variation of hinge loss defined below. The overall loss is computed as follows:

$$\begin{aligned} \text{loss}_I = & \text{CrossEntropy}(\hat{y}, y_\ell) + \\ & \text{CosineEmbLoss}(\hat{\mathbf{v}}, \mathbf{v}_k), \end{aligned} \quad (6.61)$$

where

$$\text{CosineEmbLoss}(\hat{\mathbf{v}}, \mathbf{v}_k) = \begin{cases} 1 - \text{Cos}(\hat{\mathbf{v}}, \mathbf{v}_k) & \text{if } k = \ell \\ \max(0, \text{Cos}(\hat{\mathbf{v}}, \mathbf{v}_k) - \delta) & \text{if } k \neq \ell \end{cases} \quad (6.62)$$

Like hinge loss, the cosine embedding loss includes a margin δ , which we set to 0.1. Intuitively, δ prevents the loss function from penalising the model for placing its class representation prediction close to the representation of a nearby negative class, as long as it isn't too close. After all, some classes *are* similar. The best IPT model is chosen based on the zero-shot mean rank of true unseen classes in the validation set.

6.8.5. Experiments

6.8.5.1. Data

We use the Caltech-UCSD Birds-200-2011 dataset (Wah et al., 2011, hereafter CUB), a collection of 11788 images of birds from 200 different species. The images were sourced from Flickr and filtered by crowd workers. In addition to class labels, the dataset includes bounding boxes and attributes, but we do not use those features in the current study, since our focus is on using natural language descriptions for zero-shot classification, rather than from structured attribute-value features.

We also use a corpus of English-language descriptions of the images in the CUB dataset, collected by Reed et al. (2016). The corpus contains 10 descriptions per image. The descriptions were written to be both precise (annotators were given a diagram labelling different parts of a bird's body to aid in writing descriptions) and very general (annotators were asked not to describe the background of the image or actions of the particular bird). This allows us to treat the captions as *class descriptions*, suitable for zero-shot

	seen	unseen	Total
Train	8482	948	9430
Test	1060	119	1179
Val	1060	119	1179
Total	10 602	1186	11 788

Table 6.19. Number of CUB corpus images by data split.

classification. We split the dataset into 180 seen and 20 unseen classes and train, test, and validation sets of each (Table 6.19).

A single training example is a triple $(\mathbf{x}, y, \mathbf{d})$, consisting of an image, class label, and description. Since there are 10 descriptions per image, this gives us 84 820 seen training examples for the interpreter. The generator is additionally trained on the 9480 unseen training examples, but with the descriptions omitted. To mitigate the possibility that the unseen split represents a particularly hard or easy subset of classes, we test 5 folds, each with disjoint sets of unseen classes. The results reported are the mean values across the five folds.

6.8.5.2. Evaluation metrics

Generation and classification We evaluate the performance of GEN with BLEU (Papineni et al., 2002) and CIDEr (Vedantam, Lawrence Zitnick, et al., 2015) — the latter has been shown to correlate best with human judgements in multi-modal setup. As is standard in classification tasks with many classes, the interpreter is evaluated with different notions of accuracy: *accuracy @1*, *@5* and *@10*, where a prediction is considered successful if the true label is the top, in the top 5, or in the top 10 labels, respectively. We also consider the *mean rank* of the true class to examine how close the model is in cases where its prediction is incorrect.

teacher	GEN train data	CE loss	mean rank	acc@1	acc@5
random baseline		5.30	100.5	0.5	2.5
ground truth	seen	2.50(18)	5	36.4(41)	75.1(28)
	unseen	4.17(46)	30(6)	19.1(41)	44.1(56)
best GEN	seen	2.36(12)	6(1)	43.7(25)	74.1(31)
	unseen	5.28(54)	46(9)	8.6(58)	25.1(39)

Table 6.20. Zero-shot classification results for the IPT model. The results are computed as macro-averages, averaged first over class, then over the fold. Only results of the unseen classes will be of interest as an evaluation metric for the GEN model, but here we report both, since it is important to see that the IPT model still performs well on seen after learning provisional unseen class representations. The ground truth results report zero-shot performance after learning from one randomly sampled ground truth description for each unseen class. The best GEN results report zero-shot performance after learning from the best GEN model (GEN-Ex with beam-2 decoding).

Discriminativity Our generation model is trained to minimise the cross-entropy of the next token, given the class label. This learning objective may encourage the model to generate “safe” descriptions, as opposed to descriptions that mention features that would help to identify birds of the given class. To measure this tendency, we define a notion of the *discriminativity* of a class description, which evaluates how helpful the description is in picking out instances of the class it describes. To compute the metric, we first extract textual features from the descriptions, where each feature consists of the noun and the set of adjectives used in a noun phrase. We define the discriminativity of a feature with respect to a particular class as the exponential of the mutual information of the feature and the bird class, as measured on the test set; that is,

$$\text{disc}(x_i) = \exp(H(Y) - H(Y|x_i)),$$

where x is a feature and Y is the bird class.

The maximum discriminativity of a feature (i.e., a feature that uniquely picks out a particular class) is equal to the number of classes, 200. For example, $\text{disc}(\{\text{'bill'}\}, \{\text{'long'}, \text{'curved'}\}) = 22.9$, whereas $\text{disc}(\{\text{'bill'}\}, \{\text{'short'}, \text{'pointy'}\}) = 2.9$, reflecting the fact that more kinds of birds have short pointy bills than

class repr.	decoding	Bleu1	Bleu4	CIDEr	discriminativity		mean rank	accuracy	
					mean	max		@1	@5
both	beam	0.68(12)	0.55(9)	1.83(30)	1.58(40)	2.32(135)	94(28)	0.0	2.1(11)
	nucleus	0.69(3)	0.32(5)	1.40(7)	5.22(162)	12.48(547)	111(14)	0.8(15)	4.1(48)
exem	beam	0.64(3)	0.58(2)	1.92(8)	1.95(41)	3.29(124)	46(9)	8.6(58)	25.1(39)
	nucleus	0.65(4)	0.36(8)	1.42(14)	5.10(144)	12.07(470)	70(7)	6.8(26)	18.3(31)
prot	beam	0.61(9)	0.55(10)	1.80(32)	1.65(22)	2.46(79)	73(17)	2.7(27)	13.6(63)
	nucleus	0.70(4)	0.38(5)	1.48(8)	5.51(183)	13.14(414)	75(12)	4.1(31)	15.1(50)

Table 6.21. Generation results on the unseen set. The models differ only in terms of the input and decoding methods. BLEU and CIDEr scores are reported as micro averages over n-grams produced in all 200 class descriptions. Mean rank and accuracy refer to the unseen test set performance of the IPT model trained on the corresponding zero-shot split, having learned provisional representations from the descriptions provided by the GEN model.

long curved bills. We define two metrics for the discriminativity, disc_{\max} , and disc_{avg} , which are the maximum and mean discriminativity of the features included in a given description.

6.8.6. Results

Our primary objective is to examine if we can learn models capable of grounding on the category level in the zero-shot image description generation setup. First, we address part of this question by exploring the performance of the IPT model when classifying new classes given GEN-generated descriptions (Table 6.20). We evaluate the performance of the interpreter on the unseen set using both ground-truth descriptions and descriptions generated by the best GEN model. See Table 6.21 for a full comparison of the generation models, including resulting IPT performance. Since multiple descriptions exist per class in the ground-truth texts, we randomly select one for each unseen class in each zero-shot fold.

Our first observation is that the model is moderately successful on the zero-shot classification task. When learning from the ground truth descriptions, the model performs well above the random baseline. While 0.19 is not very high for classification accuracy in general, it is not out of line for unseen results in zero-shot learning. It must be noted that classifying a bird

as belonging to one of 200 species based *only* on a textual description would be a difficult task for some humans as well. That the model can use the ground truth text descriptions to learn class representations that are *somewhat* useful for image classification is encouraging for the prospect of using it to evaluate the GEN models. However, we note that the performance of the model using descriptions generated from the best GEN model is quite a lot worse than the ground truth. This suggests that while the descriptions generated by the best GEN models are not totally useless, they are nevertheless not as communicatively successful as they could be. We observed intriguing results regarding seen classes: generated texts can be more useful than ground-truth descriptions for the IPT. This suggests either (i) lower quality of generated texts and the interpreter relying on common bird features and spurious correlations, or (ii) the possibility that human-generated texts are not as informative as initially assumed. Ultimately, human descriptions were primarily intended for interpretation by other humans, which could explain why they may have omitted significant information that listeners already possessed prior knowledge useful for interpretation.

Next, we compare different GEN models, as shown in Table 6.21. We can see that GEN-Ex outperformed the others on most intrinsic metrics (except for BLEU-1) and also in terms of communicative success. Beam search performed better than the nucleus on the intrinsic metrics, and particularly excelled in the case of GEN-Ex. Interestingly, nucleus-generated texts nevertheless scored much higher in terms of discriminativity. GEN-PROT and GEN-BOTH performed similarly on the intrinsic metrics, but GEN-BOTH performed extremely poorly (worse than random baseline in some cases) in terms of communicative success.

6.8.7 Discussion and conclusion

One of the motivations behind adopting this task was its reliance on *class-level* representations, which distinguishes it from other image-specific language-

and-vision tasks. We wanted to see how well the models can ground language in the absence of image pixels. Our results revealed several interesting directions for further exploration in modelling, feature representation, and the evaluation of generation and interpretation. Strikingly, the top-performing models were the GEN-Ex models, which effectively reintroduce specific images by picking out exemplars to generate from. Of course, the models we used in this study were relatively simple, and more sophisticated neural models may yield different results. But this raises an interesting question for future work — what *does* it take to learn grounded representations that are useful in this particular communicative context?

More generally, why do the GEN model descriptions fall short in comparison to ground-truth for the zero-shot performance on the IPT model? There are two possible explanations. One is that the generated descriptions may lack the necessary visual information required for successful classification of unseen classes. Secondly, the texts produced by the GEN model may not be interpretable by the IPT model. Recall that the IPT model was trained on ground truth descriptions from seen. These descriptions have a structure to them — certain regularities in conveying visual information. If the GEN descriptions deviate from this structure, IPT may struggle to effectively utilise them, even if they do in some sense “contain” visual information. Indeed, there is some evidence that this is what is happening. We see that nucleus sampling resulted in higher discriminativity scores, including for the GEN-PROT and GEN-BOTH models. Although the generator produces sequences with adjective-noun phrases that identify the correct class, IPT cannot properly use them, perhaps because they appear in texts that are “ungrammatical” for the distribution IPT was trained on. As both the GEN and IPT models are simple approximation functions of the data on which they are trained, they may rely too heavily on patterns and regularities, which can hinder their ability to learn to recognise and generate better category-level descriptions. This points to an important research direction, as it might highlight the limitations of many

current existing multi-modal models which are built on top of the transformer architecture. Such models might still be useful in various domains but often face challenges in learning higher-level concepts about the world.

A different question is whether generation metrics reflect the communicative power of texts as measured by the interpreter’s performance. In the case of GEN-Ex, IPT performs best with texts generated using beam search (Table 6.21). However, these texts overall score very low on discriminativity. Indeed, we discovered that beam search generates sentences with features common to multiple classes, e.g. “a bird with wings”. At the same time, IPT benefits more from nucleus-generated texts produced by the GEN-PROT model. These texts are more diverse, possibly describing a larger set of class features and our interpreter is able to learn better from that diversity. Intrinsic generation metrics rate nucleus-generated texts generally lower, suggesting a mismatch between task-based evaluation (e.g., interpreter’s performance) and intrinsic evaluation (e.g., generation metrics). These findings suggest that the “groundedness” of class descriptions and their use for the task might not be adequately captured by the set of NLG metrics and one might want to use the generated texts “in context” (aka interpretation) to get a clearer picture on how much important information such texts carry.

In future work, we will focus on improving interpretation performance by emphasising fine-grained differences between class features. Inspecting how the generation of more descriptive and more discriminative class descriptions can be achieved is also important. Additionally, we will examine the impact of a more interactive context on the task, which could be studied in a reinforcement learning setup (Oroojlooy and Hajinezhad, 2021).

6.8.8. Limitations

This paper focused on proposing the task of visual category description, and testing different cognitively-inspired representations for standard neural network architectures. We did not expand our experiments to more complex

models. Our analysis can also be performed in the context of different encoder-decoder combinations. Secondly, the dataset has fine-grained descriptions of categories. However, these descriptions can be so specific that they lack in generality, which depends on the domain and even personal preferences and background of those who interact. While this does correspond to the situation in certain real-life situations (bird classification being one of them), the results may look different in a more open-domain setup, such as in the dataset from Bujwid and Sullivan (2021).

Moreover, the way that the descriptions were collected may mean that they differ somewhat from how a human would produce visual category descriptions. Experimentation with more datasets of different levels of specificity and with different kinds of ground truth classes descriptions would be necessary to draw more general conclusions. Given that we employ a pre-trained transformer model to encode texts, we note that there might be an impact of BERT’s prior knowledge about birds on the interpreter’s performance. We recognise this as a promising starting point for exploring the model’s performance across other domains, allowing us to assess the general effectiveness of the setup we have introduced.

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 of 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_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_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. Schluter, 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_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/repo/rt.pdf>.
- Botvinick, M. M. (2008). “Hierarchical models of behavior and prefrontal function”. In: *Trends in Cognitive Sciences* 12.5, pp. 201–208. URL: <https://doi.org/10.1016/j.tics.2008.01.003>.

- //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. Krahmer, 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. Gülcöhre, 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. Somashekharappa. 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_cvpr_2018/html/Das_C_Embodied_Question_Answering_CVPR_2018_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%5C_Desai_2021_CVPR_3173.html

- 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. Ghaniifard, 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 Judgements 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://semidial.org/anthology/Z16-Dobnik_semidial_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://semidial.org/anthology/Z21-Dobnik_semidial_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/W17-3207>.

- 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/10.111/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/10.1002/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. Strobelt, and S. Gehrman (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_t%5C_CVPR%5C_2019/html/Hudson%5C_GQA%5C_A%5C_New%5C_Data_Set%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://semodial.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. Bejing, 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. Kurabayashi, 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=Hk8N3Sclg>.
- 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.

- eurips.cc/paper_files/paper/2021/file/505259756244493872b
7709a8a01b536-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/tacl%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://d4mucfpksywv.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. Zarrieß, 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.ac-1-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. Sripana. 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. Willemse (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/2d6cc4b2d139a53512fb8ccb3086ae2e-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 RedisCOVERS 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/3f5ee243547dee91fb053c1c4a845aa-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_Conensus-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>.
- Wiegreffe, 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/g2021.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/375c71349b295fbe2dcda9206f20a06-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>.
- Zarrieß, 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>.
- Zarrieß, 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 ADE2oK 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). “Texxygen: 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.