Thesis for The Degree of Licentiate of Engineering

Natural Language Processing in Context

A Picture is Worth a Thousand Words

Lovisa Hagström

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Natural Language Processing in Context
A Picture is Worth a Thousand Words

Lovisa Hagström

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2023.

To my grandmother, Barbro.
Till min farmor, Barbro.

Natural Language Processing in Context
A Picture is Worth a Thousand Words

Lovisa Hagström

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Modern NLP models learn language from lexical co-occurrences. While this method has allowed for significant breakthroughs, it has also exposed potential limitations of modern NLP methods. For example, NLP models are prone to hallucinate, represent a biased world view and may learn spurious correlations to solve the data instead of the task at hand. This is to some extent the consequence of training the models exclusively on text. In text, concepts are only defined by the words that accompany them and the information in text is incomplete due to reporting bias. In this work, we investigate whether additional context in the form of multimodal information can be used to improve on the representations of modern NLP models. Specifically, we consider BERT-based vision-and-language models that receive additional context from images. We hypothesize that visual training primarily should improve on the visual commonsense knowledge, i.e. obvious knowledge about visual properties, of the models. To probe for this knowledge we develop the evaluation tasks Memory Colors and Visual Property Norms.

Generally, we find that the vision-and-language models considered do not outperform unimodal model counterparts. In addition to this, we find that the models switch their answer depending on prompt when evaluated for the same type of knowledge. We conclude that more work is needed on understanding and developing vision-and-language models, and that extra focus should be put on how to successfully fuse image and language processing. We also reconsider the usefulness of measuring commonsense knowledge in models that cannot represent factual knowledge.

Keywords

NLP, BERT, Neural network, Vision-and-language models, Grounding, Knowledge representation

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] T. Norlund, **L. Hagström**, R. Johansson, *Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it? Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (Nov 2021), 149-162.*

[**Paper II**] **L. Hagström**, R. Johansson, *What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (May, 2022), 252-261.*

[**Paper III**] **L. Hagström**, R. Johansson, *How to Adapt Pre-trained Vision-and-Language Models to a Text-only Input?*
Proceedings of the 29th International Conference on Computational Linguistics (Oct, 2022), 5582-5596.

Other publications

The following manuscripts have been published, but are not included in this work.

[**Paper a**] **L. Hagström**, R. Johansson, *Knowledge Distillation for Swedish NER models: A Search for Performance and Efficiency*
Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa) (May, 2021), 124-134.

[**Paper b**] **L. Hagström**, T. Norlund, R. Johansson, *Can We Use Small Models to Investigate Multimodal Fusion Methods?*
Proceedings of the 2022 CLASP Conference on (Dis)embodiment (Sep, 2022), 45-50.

Acknowledgment

Firstly, I would like to thank my main supervisor Richard Johansson for his steady support and guidance. I would also like to thank my co-supervisor Marco Kuhlmann for his support and educative feedback. Many thanks also to my examiner Graham Kemp who has supported me beyond the role of examiner.

I'm happy to have the privilege of working with many ambitious and intelligent people. Thank you to my PhD colleagues, Mehrdad, Christopher, Juan, Markus, Arman, Denitsa, Sólrún, Firooz, David, Hampus, Emilio, Fazeleh, Emil, Niklas, Linus and Hannes for making work at Chalmers so much more fun. A special thanks to Tobias whom I have learned a lot from through our discussions and collaborations. Not to forget my awesome office buddies Lena, Newton, Anton and Adam. Thank you also to Dag, Kolbjörn, Birgit, Morteza, Fredrik, Moa, Simon, Ashkan, Vladimir and Selpi for providing me with many opportunities to learn and develop.

I'm deeply grateful to my parents and sister. Thank you to Caroline and Jonas for your everlasting support. And to Sofia for being the best sister.

I would not have undertaken this journey without the help of my partner, Mattias. Thank you for always encouraging and inspiring me.

Lastly, many thanks to my grandmother, Barbro, for your endless patience and love. I enjoyed every conversation with you and I wish you had been able to live long enough to see all of us grandchildren blossom.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The majority of the computations for this work were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

Deep Learning and Natural Language Processing (NLP) research has recently made substantial progress (Devlin et al., 2019; Brown et al., 2020; Thoppilan et al., 2022b; OpenAI, 2022). We now have models that can e.g. classify text, extract information from text and generate plausible looking texts without the need for manual feature engineering. Instead, the models learn how to read and write from largely unannotated human-authored texts. There are several examples of downstream tasks for which these models have proven useful, ranging from text summarization to translation and chatbots (Turovsky, 2016; Gupta and Gupta, 2019; OpenAI, 2022).

Modern NLP models are derived from the same basic components. They build on language models, i.e. mathematical models that predict a probable word given a context of surrounding words (Shannon, 1948). An illustrative example can be seen in Figure 1.1. These language models are then parameterized by deep artificial neural networks trained in an unsupervised fashion to predict the most probable words based on lexical co-occurrences from large amounts of (mainly) freely available digital texts (e.g. English Wikipedia and thousands of books) (Devlin et al., 2019; Bengio et al., 2000).

While recent NLP models have made breakthroughs in many aspects of language processing, they have also exposed potential limitations of modern methods for processing and representing language. One limitation of current NLP models is that they are prone to *hallucinate*, i.e. produce outputs that are factually incorrect or simply in conflict with the context (Maynez et al., 2020). When they describe the biography of a person they might add plausible

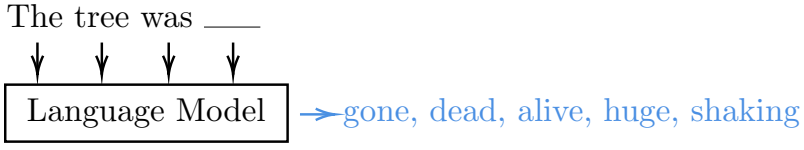

Figure 1.1: An example of how a Language Model works by predicting words from a textual context. The suggested continuations were generated by a large BERT model with 336M parameters.

but incorrect facts and when they act as a chatbot they might switch persona mid-conversation. The models also represent a biased world view, in which the Sun revolves around the Earth and women are more likely to work as receptionists (Gordon and Van Durme, 2013; Bolukbasi et al., 2016).

The aforementioned limitations are potentially not surprising since they arise from training language models exclusively on lexical co-occurrences (Bender and Koller, 2020; Bisk et al., 2020). There is nothing in the design or training of these models that specifically encourages factuality or lack of bias. Most probably, the case is rather the opposite. Also, from the perspective of the model, concepts such as 'facts', 'bananas' or 'personas' only exist in text and they are only defined by the words that surround them. There is no external information to this that further differentiates, complements or grounds the concepts. Contrariwise, a human has access to more modalities of information than text to help her differentiate between different concepts. She exists in a multimodal world that can be seen, heard, felt, experienced and read. When she reads about a banana, she knows that it refers to something external to text, i.e. the concept is grounded in information external to text.

In this work, we investigate whether multimodal information can be used to ground and improve on the linguistic representations of modern NLP models. We specifically consider vision-and-language (VL) models, language models augmented with visual information, and evaluate them for if they have acquired any additional knowledge or general language understanding. In Paper I we introduce a basic method for evaluating the color knowledge of NLP models, something we expect to improve from training on visual information. In the same paper we also propose a new VL model architecture. Paper II presents a more general task that measures the visual conceptual knowledge of a model and evaluates a set of VL models on this. However, since all of these tasks are text-only and VL models generally are trained on text-image pairs, there is the question of whether the models are evaluated slightly out-of-distribution. Therefore, Paper III further builds on the work of Paper I and II by investigating if there are better setups for evaluating VL models on text-only tasks.

Before we delve further into the contributions of the aforementioned papers, we also outline their theoretical background in Chapter 2.

Chapter 2

Background

The main focus of the work presented in this thesis is in the area of representation modelling, as introduced in Section 2.1 and expanded on in Section 2.3. BERT models are typically used in this setting and they are derived from the Transformer network, introduced in Section 2.4 and Section 2.2 respectively.

The motivation for the work in this thesis originates from observed limitations of language models, described in Section 2.7. The approach we investigate for solving some of the limitations of NLP models is to make use of context and grounding, for example in the form of multimodal models, as introduced in Section 2.8. Specifically, we focus on vision-and-language models and build on previous research on these, as presented in Sections 2.9 and 2.10. Additionally, we make use of different evaluation methods in the investigations of the vision-and-language models, as described in Sections 2.5 and 2.6.

2.1 NLP models

Modern NLP models can be segmented into three different categories depending on what format they are to be used in, as illustrated in Figure 2.1. The main formats are representation modelling, language modelling and sequence-to-sequence processing (Jurafsky and Martin, 2022). For representation modelling, an NLP model is used to generate a vector, a representation, of the text input that then can be used instead of the text for any text related task, such as sentiment classification or categorization (Peters et al., 2018; Devlin et al., 2019). The main focus of the work of this thesis lies on representation models. For language modelling, the goal is to generate a continuation to the provided input text. This format is useful for text generation, used in e.g. chatbots, question answering or story generation (Brown et al., 2020). For sequence-to-sequence processing the task is to generate a completely new sequence based on an input sequence. This format is for example used in machine translation and summarization (Lewis et al., 2020).

Neural networks are used in most modern NLP models regardless of usage format, as they are especially suitable for modelling text or text representations. Neural networks are universal function approximators (Hornik et al., 1989) in

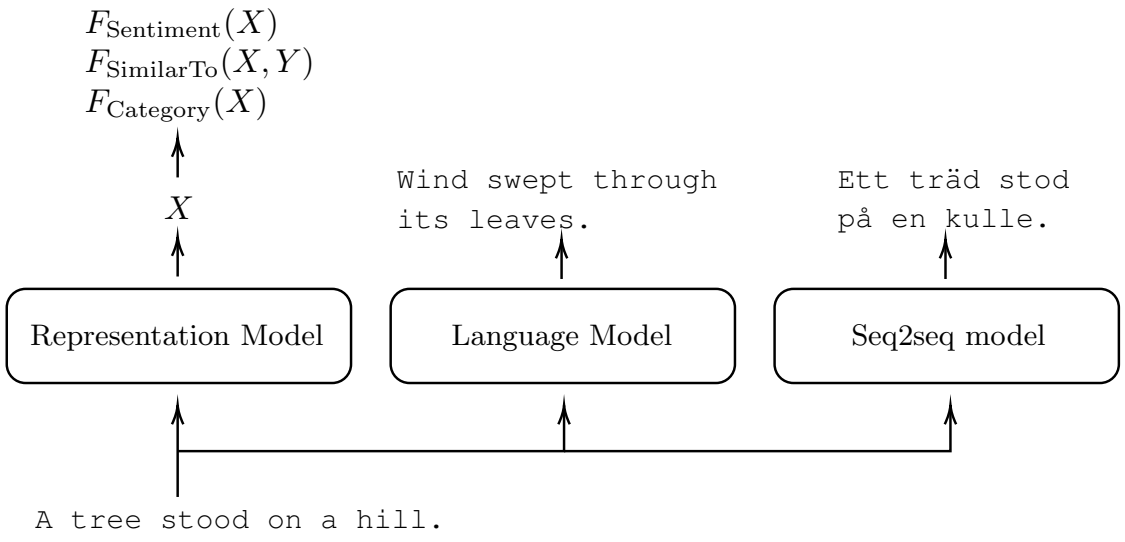

Figure 2.1: The three main usage formats of NLP models.

the sense that they may be used to model any imaginable function. In addition to this, they require no manual feature engineering and can be trained directly on unannotated text data to learn in an unsupervised fashion.

Most modern NLP models are *parametric models*, in the sense that they are fully described by and limited to their finite set of parameters. This can be described as,

$$P(x|\theta, \mathcal{D}) = P(x|\theta),$$

where θ contains the parameters that fully describe the network and $\mathcal{D}$ is the data on which the network has been trained. Consequently, θ contains all of the information that has been learned from the training data. This is very convenient in the sense that we only need to retain the neural network parameters and not the data for subsequent text processing purposes.

2.2 Transformer

The neural network architecture that typically is used in modern NLP models is the deep *Transformer* network (Vaswani et al., 2017). This network utilizes stacked *Attention* layers (Bahdanau et al., 2015) to model dependencies between words in a sequence and has proven to be very performant for language modelling. This network setup works well also for longer sequences for which important word sets are far from each other. Compared to the previous state-of-the-art NLP models based on recurrence and convolutions, the Transformer architecture largely avoids sequential computing. Thanks to its superior modelling capacity and parallelizability, the Transformer is the new state-of-the-art network for language processing.

The Transformer network was originally developed for language translation. Since translation is a sequence-to-sequence task, the Transformer consists of two networks, an encoder network to encode the input to be translated and a

decoder network that generates the translation while considering the encoded input and preceding output. Each of these network parts can and have been used separately in modern NLP models. The encoder lends itself especially useful for representation learning, while the decoder is useful for autoregressive language modeling for which we wish to generate continuations based on preceding values of a provided sequence. The only remaining sequential aspect of this model is the generation by the decoder, meaning that all other computations can be parallelized for faster training.

Both the encoder and decoder of the Transformer network build on stacks of six respective identical layers. The layer for the encoder consists of multi-head attention and a fully connected feed-forward network components. The layer for the decoder is similar to the encoder layer, while it contains additional attention over the encoder output and masks the attention over the decoder input to prevent information leakage from the tokens to be predicted. This stacked setup allows for easy re-scaling of the Transformer, since one can simply change the number of layers in the stacks.

2.3 In practice: language representations and transfer learning

Language representations are used to some extent in all NLP models. While neural networks can be used to model anything in any form we desire, it is useful to make use of some inductive bias to structure the modelling and facilitate the training of the neural networks. Therefore, representations are typically used to incrementally process text, moving from more general to more specific linguistic representations further up in the model. In this context, a representation is a vector $v \in \mathbb{R}^d$ where d is the size of the representation, that contains the essential information of the concept it represents. In the field of NLP these concepts can for example be words, sentences or documents.

An example of a language representation is the *word embedding* (Mikolov et al., 2013). It has been tuned to encode distinctive information about the word it represents, so if we for example consider the word embeddings for the words “cat” and “dog”, we would find that they are more similar than for example “cat” and “plankton”. We can also have a representation that encodes information about a full sentence. The benefit of language representations is that they allow us to modularize language processing such that we can, for example, reason about sentence representations as an aggregation of word embeddings. It also allows us to develop generic language representations that later can be reused for more specific tasks.

Language representations are typically used in *transfer learning* (Pan and Yang, 2010). For this setting, a NLP model is first trained to produce generally useful language representations for a general task, after which the representations can be further tuned for a specialized task. Transfer learning is suitable for situations in which we have several similar tasks and expect them to require similar skills to solve. We can then learn a set of generic skills that can be tuned and used for each task. This is useful from the efficiency perspective in

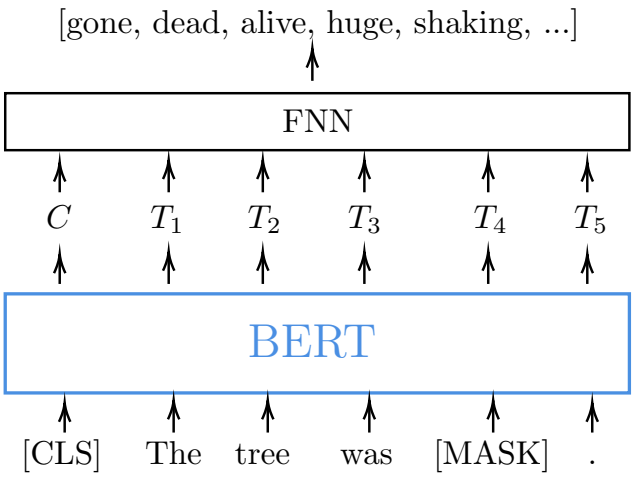

Figure 2.2: An illustrative image of BERT for masked language modelling. FNN denotes a feed-forward neural network.

not having to train several similar models from scratch, and might solve all tasks better.

2.4 BERT

A model frequently used in the work of this thesis is the Bidirectional Encoder Representations from Transformers (BERT) model. It had a large impact on modern NLP research after it was developed by Devlin et al. (2019) and showed a promising path forward for NLP. BERT is a language representation model that has been trained to generate contextualized token representations in a bidirectional fashion, also considering the words after the word of interest in a sequence. The BERT model comes in two sizes, BERT-base and BERT-large, modelled by a Transformer encoder with stacks of 12 or 24 layers respectively, as described in Section 2.2. The encoder generates representations for input tokens that can be used by a smaller network to solve some downstream task, as illustrated in Figure 2.2. The assumption is that if the encoder is sufficiently trained, it should be able to generate language representations that are useful for generic language tasks, as in transfer learning.

The BERT model is trained in two steps. The first step is a pre-training phase in which the model is tasked with Masked Language Modelling (MLM), i.e. predicting masked words in a text passage, and next sentence classification on a large text corpus. The training data of the BERT model consists of English Wikipedia and the Book Corpus (Zhu et al., 2015). The second training step is a fine-tuning phase during which the model can be specifically tuned to perform some kind of specific linguistic task, usually by adding a feed-forward neural network on top. With this setup, even low-resource tasks may be possible to solve thanks to the general language capabilities that have been obtained by the model in the pre-training step.

2.5 Evaluating for natural language understanding

The overarching goal of NLP research is to acquire models that are useful for natural language processing. To facilitate this research there are methods for assessing if a proposed model for language processing is better than another. A popular method for assessing the performance of a NLP model is to evaluate it on an open benchmark. The benchmark score of this model can then be compared to scores of previously evaluated models.

A frequently used benchmark that we also make use of is the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). As the name implies, the aim of this benchmark is to test for the general natural language understanding (NLU) of a model from evaluation on nine different NLP tasks, such as sentiment prediction and question answering. When the BERT model was proposed it was mainly evaluated on GLUE. However, shortly after BERT was released, other newer NLP models outperformed both BERT and the human baseline on GLUE, indicating that this benchmark no longer is suitable for evaluating the language capabilities of modern NLP models. In reality, the NLU of NLP models is inferior to that of a human and a proper benchmark should reflect this. Therefore, GLUE is useful for evaluating BERT-like models, but not necessarily newer and larger language models.

A successor to GLUE is SuperGLUE (Wang et al., 2019). This benchmark also evaluates for general natural language capabilities from evaluation on eight different tasks that are more challenging and diverse than those of its predecessor. Consequently, this benchmark is more suitable for evaluating successors of the BERT model that outperform the human baseline of GLUE.

2.6 Evaluating for knowledge

More than linguistic knowledge is required for successful language processing (Zhang et al., 2021). To fully understand language, commonsense world knowledge, and some factual knowledge is required. We here refer to general knowledge of an agent as when the agent takes the world to be one way and not another, leaving no room for contradictions (Brachman and Levesque, 2004). With commonsense knowledge we refer to knowledge that most humans agree on, such as “A bear has a nose.”, and with factual knowledge we refer to information that is non-questionable and can be found in e.g. certified Wikipedia pages. We can thus view commonsense knowledge as a subset of factual knowledge.

An underlying theme of this thesis is to measure the commonsense knowledge of NLP models. Zhang et al. (2021) suggest that recent success of large language models on NLU benchmarks can be attributed to the capability of these models for learning the required commonsense knowledge for solving the benchmarks. It is seemingly more difficult for NLP models to learn commonsense knowledge than syntactic and semantic knowledge, suggesting that this is the driver for successful NLU.

There is much interest in extracting and measuring the different types of knowledge that supposedly resides in NLP models. For BERT-like models, cloze statements are typically used to extract knowledge since the models are tuned to this format from their MLM pre-training. Petroni et al. (2019) use this format to test for factual and commonsense knowledge with their LAMA (LAnguage Model Analysis) probe based on Wikipedia and commonsense knowledge. They find that much knowledge is stored in language models and further hypothesize that these models have a potential use as knowledge bases. Weir et al. (2020) also test for commonsense knowledge, such as "A dog has fur.", and find that these are present in different BERT based models. West et al. (2022) experiment with extracting latent commonsense knowledge from a GPT-2 model to create a knowledge graph, and use different prompting techniques for this.

2.7 Limitations of modern NLP models

As mentioned in the introduction, recent successes have also exposed limitations of modern NLP models. In this section, we further expand on issues related to hallucination, learning spurious correlations, reporting bias and critique against the general setup and goals of current NLP research. The work of this thesis is primarily focused on mitigating the limitations related to reporting bias and general setup of NLP research.

Generative NLP models suffer from hallucination. As illustrated in Figure 2.3, even sophisticated models such as ChatGPT may generate plausible but incorrect facts. The text generated by the model states that Astrid Lindgren was born in Vimmerby, while she was born in a town close to Vimmerby. It also states that Lindgren received the Right Livelihood Award in 1979, while she actually received it in 1994. If these models are used for low-risk scenarios such as storytelling, there is no immediate danger. It is, however, a problem for situations in which correct information is important and expected.

Language models are also affected by *reporting bias* (Gordon and Van Durme, 2013) in that the majority of human written text does not contain information the writer expects the reader to already be familiar with. Text rather describes non-obvious world information that is worth writing down. For example that a banana is green and straight rather than yellow and curved. Since modern NLP models are trained solely on human written text, they learn an inaccurate world representation. Paik et al. (2021) have for example observed that the models learn a perception of color that does not agree with that of a human, but rather with the text used for training.

Another issue with NLP models, and with many deep neural networks in general, is that they may learn to make correct inferences with an incorrect method. Very few modelling constraints are necessary or imposed on the models to guide their inference process, as this is also a strength of neural networks. However, this means that the models may make use of spurious correlations or artifacts in the training data to solve the data and not the task (Feder et al., 2022). NLP models have been noted to base predictions that should be factual

Astrid Lindgren was a Swedish writer and children's book author best known for her character Pippi Longstocking. She was born on November 14, 1907 in Vimmerby, Sweden. Lindgren grew up in a family with strong storytelling traditions, and she began writing stories and poems from a young age.

After completing her education, Lindgren worked as a secretary and later as a journalist. In the 1930s, she married Sture Lindgren and had two children. In the 1940s, she began writing children's books, starting with the character Pippi Longstocking, which became her most well-known and beloved creation.

Lindgren's books have been translated into more than 70 languages and have sold millions of copies worldwide. In addition to Pippi Longstocking, Lindgren is also the author of numerous other children's books, including the Ronia the Robber's Daughter series, The Children of Noisy Village, and Karlsson-on-the-Roof.

Lindgren was also a strong advocate for children's rights and was involved in various humanitarian and political causes throughout her life. She received numerous awards for her work, including the Hans Christian Andersen Award for Children's Literature in 1958 and the Right Livelihood Award in 1979. Lindgren passed away on January 28, 2002 at the age of 94.

Figure 2.3: Hallucination in ChatGPT. The model was prompted with the query "Give me the biography of Astrid Lindgren.". Passages marked in yellow are questionable or clearly incorrect.

on the style of their prompts, meaning that they might switch their prediction when they are queried for the same fact but in a different way (Elazar et al., 2021; Cao et al., 2021; Jiang et al., 2020). Additionally, NLP models solve the grounded commonsense inference task SWAG by picking up on dataset-specific distributional biases instead of learning to perform commonsense inference (Zellers et al., 2018, 2019). It has also been observed that BERT picks up on shallow heuristics such as lexical overlap to solve a general natural language inference task (McCoy et al., 2019).

There is also critique against the general approach of learning language from lexical co-occurrences. For the development of general NLP models such as BERT, the goal was general language understanding, as evaluated by e.g. GLUE. However, Bender and Koller (2020) and Bisk et al. (2020) have questioned whether modern NLP models can acquire language understanding, since they only operate on linguistic form. The general argument is that understanding requires meaning, where meaning is the relation between linguistic form and communicative intent, while the models have no notion of communicative intent. This argument is encapsulated by the symbol grounding problem that points to the impossibility for a non-Chinese speaker to learn the meanings of Chinese words from Chinese dictionary definitions alone (Harnad, 1990). For simplicity, we will intermittently continue to refer to the language processing capabilities of a NLP model as "language understanding", while we generally agree that the word "understanding" is misused in this setting.

2.8 NLP in context: grounding and multimodal models

Much critique against modern NLP models refers to that they are not *grounded* (Bender and Koller, 2020; Bisk et al., 2020). This generally refers to that they have no connection to the actual concept(s) a linguistic form, e.g. a word, refers to. The exact definition of grounding is not yet established within NLP research, and different interpretations can be found across research papers (Chandu et al., 2021). In this paper, we define grounding in the context of NLP to refer to connecting a linguistic concept to its correspondence in some representation dimension external to text, such as sensory representations or knowledge representations. For example, if a human reads the word "tree" they know that it refers to something they can e.g. see, hear, touch, interact with and smell in the real world. They also know that they can go to the library or Google to get more information about it. Consequently, a lack of grounding for an NLP model can also be seen as a lack of context. A model that has been trained solely on text has not been provided with the necessary context to ground linguistic form in.

An obvious first step towards grounding NLP models is to provide the models with the context together with the linguistic form. For example, when we provide a model with the phrase "a dog sitting on a couch", we can also provide the model with an image of a dog sitting on a couch. There is then potential for the model to learn to connect the concepts in the image with the concepts in the text. Such a model is denoted as *multimodal* since it receives information from multiple modalities. A multimodal model can combine information from two or more modalities and is useful for tasks that require multimodal information processing, such as visual or video question answering and image captioning (Zellers et al., 2021; Goyal et al., 2017; Lin et al., 2014). A multimodal model also has the potential to produce language representations that are grounded in information from a modality other than text.

Apart from arguments related to grounding, multimodal models also have the benefit of a multi-view of concepts (Huang et al., 2021). Meaning that if some information in text is incomplete, it can potentially be found in the complementary modalities. For example, if the text suffers from reporting bias in a certain aspect, the other modalities might compensate for this and result in better latent representations of the model.

2.9 Vision-and-language models

A modality that is frequently combined with text is the visual modality. Models that process both visual information and textual information are referred to as vision-and-language (VL) models. Examples of VL models considered in our work are VisualBERT, LXMERT, OSCAR, CLIP-BERT and FLAVA (Li et al., 2019; Tan and Bansal, 2019; Li et al., 2020; Singh et al., 2022). All of these models, except for CLIP-BERT and to some extent FLAVA, have been

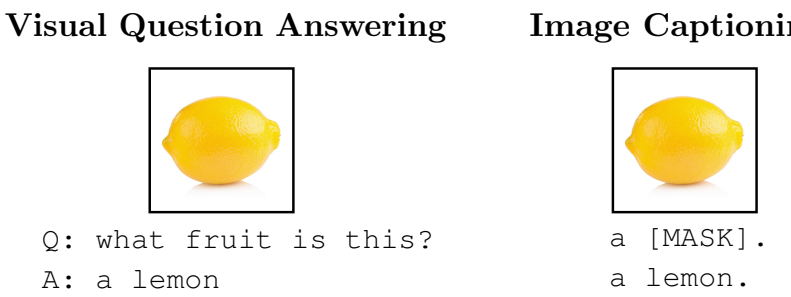

Figure 2.4: Two potential tasks for VL models. In the question answering case, the model usually generates an answer or performs a choice out of multiple options. In the image captioning case, the model can either be queried in an MLM fashion or generate a caption from scratch.

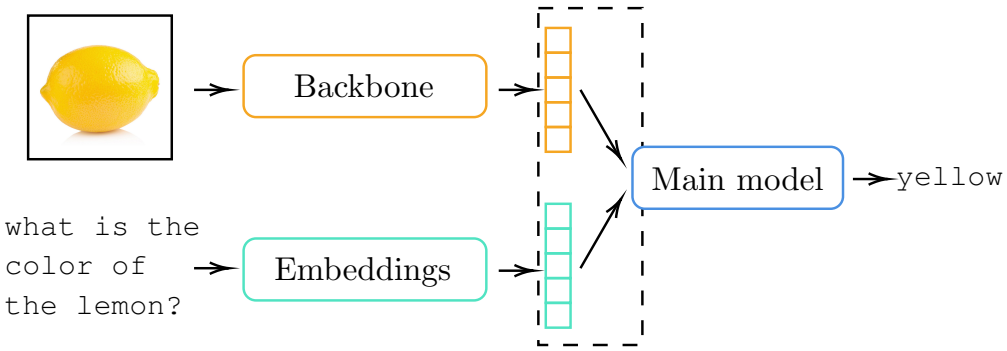

Figure 2.5: The typical setup for a VL model. Image features extracted by a backbone are given to a main model together with the text representation, usually formatted as embeddings. The dashed rectangle marks the part of the model that fuses the visual and textual information and is further described in Figure 2.6.

developed to solve predominantly VL tasks, such as Visual Question Answering (VQA) or image captioning, as illustrated in Figure 2.4. Furthermore, all of these models were developed as general purpose models and can similarly to BERT be adapted to different downstream tasks.

Most VL models are largely similar in their model setup, as illustrated in Figure 2.5. Typically, the models form initial representations for the visual input and textual input separately before the information from the different modalities is fused in the main model. Pre-trained word embeddings are typically used for the text input and a pre-trained visual model, generally referred to as *backbone*, is used to generate a representation for the visual input. VisualBERT, LXMERT and OSCAR use a frozen Faster R-CNN object detector (Ren et al., 2015) to extract detection features from the visual input, while CLIP-BERT utilizes a frozen CLIP model (Radford et al., 2021) and FLAVA utilizes a non-frozen Vision Transformer (ViT) model (Dosovitskiy et al., 2021) to generate image features. Also, all aforementioned models use Transformer encoder networks and VisualBERT, OSCAR as well as CLIP-BERT are based on a BERT model architecture.

The method for fusing the visual and textual information varies depending

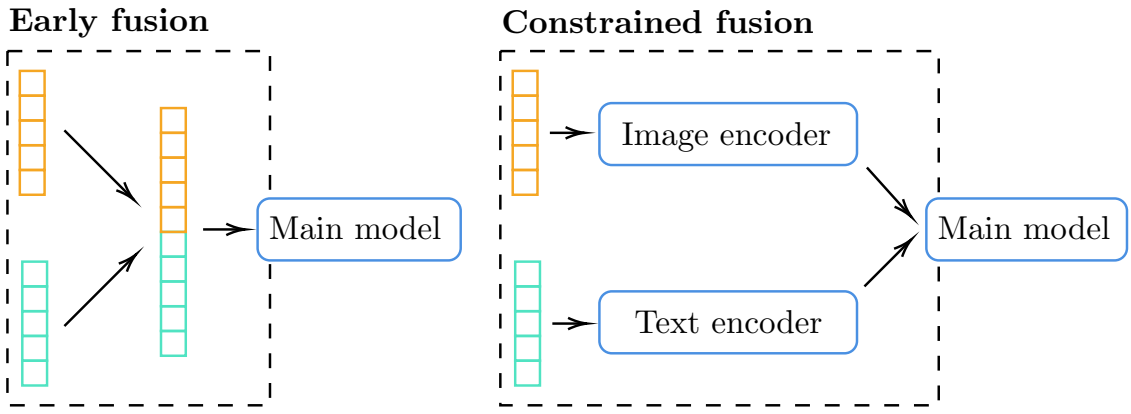

Figure 2.6: The two different fusion methods used by the VL models described in this thesis. For the early fusion, the image and text representations are simply concatenated and for the constrained fusion the representations are processed separately before the information is fused in a constrained manner through e.g. cross-attention. For the constrained fusion method the main model can also be referred to as *multimodal encoder*.

on VL model. All VL models used in the work of this thesis are *fusion encoders* with self-attention spanning across the modalities (Singh et al., 2022). Furthermore, all models except for LXMERT and FLAVA use a single transformer as main model for early and unconstrained fusion modalities, while LXMERT and FLAVA only allow for cross-attention between the modalities in specific co-attention transformer layers in the main model, as illustrated in Figure 2.6.

The models we consider are also similar in their training procedure. VisualBERT, OSCAR and CLIP-BERT are initialized from pre-trained BERT-base model weights. All aforementioned VL models are then trained on image-text datasets of varying size and information content. Common for all datasets is that they either are visual question answering datasets or image captioning datasets, as illustrated in Figure 2.4. For example, VisualBERT is trained on the image captioning dataset MS COCO and the Visual Question Answering (VQA) dataset (Lin et al., 2014; Goyal et al., 2017), while LXMERT in addition to these datasets is trained on Visual Genome, GQA and VG-QA (Hudson and Manning, 2019; Zhu et al., 2016). The tasks the models are trained on differ slightly depending on model. Examples of training tasks are MLM, image-text matching and image feature prediction. Most VL models are trained on at the least MLM and image-text matching.

A major difference between FLAVA and other VL models is that FLAVA was developed to work well for vision tasks, language tasks and multimodal vision-and-language tasks simultaneously. Most VL models have been developed to only work well for vision-and-language tasks, ignoring the unimodal situations. FLAVA, on the other hand, has been designed to perform well for language understanding, visual recognition and multimodal reasoning tasks. It builds on a modularized architecture, utilizing separate image and text encoders that also generate representations for a multimodal encoder that fuses the separate

information sources. Each separate encoder is then responsible for each task from each modality.

2.10 Analysis of vision-and-language models

Apart from research on improving the performance of VL models on VL tasks, there is also work on understanding how VL models work and how their cross-modal interactions function, i.e. how they fuse information from text with information in images. It has been found that VisualBERT, LXMERT and other BERT based VL models have a similar performance when unified from training on the same VL corpus for the same tasks, indicating that the different VL model architectures are of little importance for model performance (Bugliarello et al., 2021). For VisualBERT and LXMERT, it has also been found that visual information is used more for text processing than vice versa for visual processing, indicating that the models are not symmetrically cross-modal (Frank et al., 2021). Hessel and Lee (2020) also found indications that the VL performance of LXMERT largely cannot be attributed to cross-modal interactions, but rather unimodal information processing.

Based on the discussion on grounded NLP, there have also been investigations into the use of VL models for language processing grounded in visual information. Yun et al. (2021) investigate whether multimodal training improves on the linguistic representations of NLP models. They evaluate the representations of VisualBERT and find that they are not better than those of the text-only BERT model, indicating that more work is needed before we can use multimodal training to improve on the performance of NLP models. On the other hand, Paik et al. (2021) find that CLIP, a model that processes both images and text, mitigates issues with reporting bias and can acquire a color perception that agrees more with that of a human compared to text-only models. Iki and Aizawa (2021) also investigate the language understanding of VL models, but rather to check if performance on GLUE is preserved by a model tuned to perform VL tasks. It was generally found that VL models decrease in performance on GLUE compared to their language model counterparts.

Chapter 3

Summary of Included Papers

3.1 Paper I

In Paper I we investigate the use of visual data to complement the knowledge of large language models, as described Section 2.8. We propose a method for evaluating visual knowledge transfer to text for uni- or multimodal language models and introduce a novel text-only task, Memory Colors, querying for knowledge of memory colors, i.e. typical colors of well-known objects (Pérez-Carpinell et al., 1998). The task is in English and contains 109 object types paired with their memory color according to the knowledge of 11 human annotators. Since we are only interested in evaluating representation models, Memory Colors is formatted as a cloze-style fill-in-the-blank task, similar to the format used by Petroni et al. (2019). An example of a query from Memory Colors is "What is the color of a sunflower? [MASK]", where [MASK] should be filled in with the correct answer (yellow).

Similarly to the case for humans, we assume that a model with sufficient knowledge of visual concepts should be able to answer text-only queries about their colors without necessarily being provided with images of the concepts. To support this point, we complement Memory Colors with a human baseline from 11 human annotators that did not have access to images while answering the queries.

We also introduce a novel VL model architecture, CLIP-BERT that utilizes CLIP as backbone and BERT-base as main model. We train it on 4.4M captions and 2.7M images. After training, it can be used to make inferences in an implicit or explicit mode. In the implicit mode, the model is queried with only text and in the explicit mode it is also provided with a visual representation of the text generated by CLIP. Since CLIP has been trained to map visual and textual representations to the same space, it has a potential use for "imagining" a visual representation corresponding to text when no image is available. We also measure an upper bound for the CLIP-BERT performance by evaluating

it when it is provided with images corresponding to the text.

To separate and investigate the knowledge contributions from text versus images, we experiment with removing information about visual concepts from the text part of the training data by using different filtering methods. For example, we might remove a training example from the data if it contains an object and its corresponding color from Memory Colors. In this way we can clearly separate knowledge contributions from images and text respectively.

Finally, we evaluate CLIP-BERT on Memory Colors in the different modes with the different filterings of the training data. We also evaluate its text-only counterpart BERT-base trained on the same different filterings of the training data. We find that CLIP-BERT outperforms BERT in every filtering setting, and with a larger margin if visual information is filtered out from the text data used for training. We also find that a CLIP-BERT model in explicit mode has a larger performance margin when visual information has been filtered out from the training corpus. This indicates that our method can successfully be used to measure visual knowledge transfer capabilities in models and that our novel model architecture shows promising results for leveraging multimodal knowledge in a unimodal setting.

Contributions T. Norlund mainly contributed to the design of the study, implemented the CLIP-BERT model and code for evaluating it. He also made major contributions to the writing of the paper.

L. Hagström mainly contributed to the design of the study and developed the Memory Colors dataset. She also made major contributions to the writing of the paper.

R. Johansson provided supervision on the work and writing for the paper.

3.2 Paper II

In Paper II we concretize and further build on the hypothesis that training on a visual modality should improve on the visual commonsense knowledge in a model. Here, visual commonsense knowledge refers to obvious visual properties of concepts, such as memory colors or that bears have fur, and is less likely to be found in text. To get more robust results, we develop Visual Property Norms (VPN), a successor to Memory Colors that is approximately 60 times larger. It tests for more general visual perceptual knowledge and contains queries for up to 6,541 different conceptual features. VPN is derived from the CSLB concept property norms dataset (Devereux et al., 2014) that contains the conceptual knowledge of 30 human participants for each of 541 concrete objects. Similarly to Memory Colors, VPN is also a text-only task that evaluates for knowledge a model can express through text.

In addition to CLIP-BERT, we also evaluate LXMERT, VisualBERT and BERT baselines on VPN and Memory Colors. We find that while CLIP-BERT has the highest performance on Memory Colors, on VPN it does not outperform a BERT model that has been trained on texts that contain visual information. We also find that all models are sensitive to how they are prompted, casting

some doubts as to whether we can reason about them as having commonsense "knowledge". These results also agree with previous research (Elazar et al., 2021; Cao et al., 2021).

Contributions L. Hagström performed the main work and R. Johansson supervised.

3.3 Paper III

In Papers I and II we used textual prompts to evaluate the visual commonsense knowledge of VL models. It is however an open research question on how to use multimodal models for unimodal tasks. We handled the lack of visual information to the models by simply omitting the image input. However, most VL models have not been developed with text-only tasks in mind and generally expect image-text pairs as input. Therefore, when we evaluated the VL models, we potentially evaluated them out-of-distribution and measured a performance that was sub-par. Iki and Aizawa (2021) undertook the same issue when they evaluate different VL models on GLUE. They handled it by providing the models with a black image to replace the lacking image input, while the question remains as to whether this is a suitable method for evaluating VL models on text-only tasks.

In Paper III we investigate and compare seven possible methods for adapting three different pre-trained VL models to text-only input. The investigated adaptations can be classified into three different categories, 1) provide no visual features and potentially fine-tune the models on a text dataset, 2) provide the models with imputations for the missing visual information and 3) use the backbone of the model to 'imagine' the visual information. For the first category, we experiment with using the models zero-shot, fine-tuning on Wikipedia or fine-tuning on the training corpora of LXMERT. For the second category, we use imputations based on an average of the training images for the model, a black image, zeros or visual features that have been fine-tuned on LXMERT training corpora or Wikipedia. The third adaptation for the third category is only tested on VL models with CLIP as backbone.

The VL models we investigate in the paper are CLIP-BERT, LXMERT and VisualBERT. We also develop four text-only BERT baselines that are counterparts to the VL models. The first baseline is a default BERT model as developed by Devlin et al. (2019), the second BERT model has received further training on the training corpora of LXMERT, the third has received further training on Wikipedia corpora of the same size as the LXMERT corpora, and the fourth BERT model has been trained from scratch on the LXMERT corpora. Using these baselines, we can further disentangle effects of multimodal training from effects of 1) training on corpora with visual information, such as captions and visual questions, and 2) training on larger corpora.

We also evaluate FLAVA, a VL model that requires no adaptations since it has been developed to work well for text-only tasks. Consequently, we should be able to use it as an upper bound to compare the adaptation results against.

We evaluate all baselines and models with their different adaptations on GLUE and VPN. Generally, we find that the VL models are sensitive to adaptation for VPN, a zero-shot task, while they are less sensitive to adaptation for GLUE. We also find that the adaptation methods perform differently for different models and that unimodal model counterparts perform on par with the VL models regardless of adaptation, indicating that current VL models do not necessarily gain better language understanding from their multimodal training. Additionally, FLAVA performs worse than both unimodal baselines and VL models on the different evaluations, indicating that more work is needed for the development of truly multimodal models. Lastly, we find that all models and their adaptations are sensitive to how they are prompted, in accordance with the observations made for Paper II.

Contributions L. Hagström performed the main work and R. Johansson supervised.

Chapter 4

Discussion and Future Work

In this thesis, we investigated whether multimodal training could be used to make NLP models learn better linguistic representations in certain aspects. We mainly evaluated BERT based VL models on GLUE, Memory Colors and Visual Property Norms, of which the two latter evaluation tasks were developed by us to test for visual commonsense knowledge. Our hypothesis was that if the linguistic representations of the VL models improve from the visual training, it would firstly benefit performance on tasks that require knowledge related to visual information.

Generally, we found that the visual training signal does not improve on the linguistic representations of a VL model compared to the representations of a text-only counterpart trained on the same text data. We hypothesize that there are several potential explanations for this. One is that the evaluated VL models are not good representatives of multimodal models, suffering from e.g. simplistic fusion methods and poor backbones that have been observed to produce noisy visual representations (Frank et al., 2021). It could also be that the capabilities we evaluated for can be obtained from text-only training on visual related texts, such as captions.

Most likely, if we want models to learn grounded language representations from vision, more work is needed on resolving how the visual information should aid the textual to produce generally useful model representations and grounded outputs. Further research into information fusion methods for VL models would therefore be valuable. However, working with and evaluating VL models is difficult due to large model sizes, complex inference procedures and large training image datasets. Therefore, we have done some work on this in Paper b and are also interested in future work on making VL model investigations easier.

We noted that all evaluated models were sensitive to how they were prompted when we evaluated for commonsense knowledge. Essentially, the models changed their "world view" depending on the prompt even when they were queried for the same knowledge, contradicting the definitions of knowledge in

Section 2.6. We therefore question whether we can reason about these models as having commonsense or factual knowledge. This agrees with previous results by e.g. Elazar et al. (2022). They found that NLP models seemingly acquire shallow heuristics rather than factual knowledge.

In our future work, we will continue to focus on the intersection between factual knowledge and NLP models. One question we will further investigate is if there are alternatives to fully parametric models for representing knowledge in NLP models. If we expect a model to acquire knowledge from multimodal information, we also have the option of representing this knowledge using a non-parametric model. It would then be possible to ground the NLP model predictions in structured knowledge representations instead of implicit model parameters, as done by e.g. Thoppilan et al. (2022a).

Bibliography

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Emily M. Bender and Alexander 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ronald Brachman and Hector Levesque. 2004. *Knowledge representation and reasoning*. Elsevier.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. 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)*, pages 1860–1874, Online. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 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.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model's 'factual' predictions. *arXiv preprint arXiv:2207.14251*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Stella Frank, Emanuele Bugliarello, and Desmond 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*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taichi Iki and Akiko Aizawa. 2021. Effect of visual extensions on natural language understanding in vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2189–2196, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Dan Jurafsky and James H. Martin. 2022. Speech and language processing. `https://web.stanford.edu/~jurafsky/slp3/`. Accessed: 2023-01-02.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*. European Conference on Computer Vision.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. `https://openai.com/blog/chatgpt/`. Accessed: 2022-12-26.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Joaquín Pérez-Carpinell, MD De Fez, Rosa Baldoví, and Juan Carlos Soriano. 1998. Familiar objects and memory color. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch*

Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 23(6):416–427.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

C Shannon. 1948. A mathematical theory of communication, bell system technical journal 27: 379–423 and 623–656. *Math. Rev*, page 379.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*.

Hao Tan and Mohit 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022a. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James

Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022b. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Barak Turovsky. 2016. Found in translation: More accurate, fluent sentences in google translate. *Blog. Google. November*, 15.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of the Cognitive Science Society (CogSci)*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651. Curran Associates, Inc.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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)*, pages 1112–1125, Online. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.