Thesis for The Degree of Licentiate of Engineering

Improving Language Models Using Augmentation and Multi-Modality

Tobias Norlund

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

Improving Language Models Using Augmentation and Multi-Modality

Tobias Norlund

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 wife, Jorun

Improving Language Models Using Augmentation and Multi-Modality

Tobias Norlund

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

Abstract

Language models have become a core component in modern Natural Language Processing (NLP) as they constitute a powerful base that is easily adaptable to many language processing tasks. Part of the strength lies in their ability to embed associations representing general world knowledge. However, the associations formed by these models are brittle, even when scaled to huge sizes and using massive amounts of data. This, in combination with other problems such as lack of attributability and high costs, motivate us to investigate other methods to improve on these aspects.

In this thesis, we investigate methods that augment language models with additional contextual information, for the purpose of simplifying the language modeling problem and increasing the formation of desirable associations. We also investigate whether multi-modal data can assist in forming such associations, that could otherwise be difficult to obtain from textual data only.

In our experiments, we showcase augmentation to be effective toward these ends, in both a textual and multi-modal case. We also demonstrate that visual data can assist in forming knowledge-representing associations in a language model.

Keywords

natural language processing, language models, contextual augmentation, multimodal language modeling

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **T. Norlund**, A. Stenbom, *Building a Swedish Open-Domain Conversational Language Model*
In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 357–366, Reykjavik, Iceland (Online), 2021.*

[**Paper II**] **T. Norlund**, L. Hagström, R. Johansson, *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*, pages 149–162, Punta Cana, Dominican Republic, 2021.*

[**Paper III**] R. Raj, K. Andreasson, **T. Norlund**, R. Johansson, A. Lagerberg, *Cross-modal Transfer Between Vision and Language for Protest Detection*
In Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 56–60, Abu Dhabi, United Arab Emirates (Hybrid), 2022.*

[**Paper IV**] **T. Norlund**, E. Doostmohammadi, R. Johansson, M. Kuhlmann, *On the Generalization Ability of Retrieval-Enhanced Transformers*
To appear in Findings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia, 2023.*

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to content overlapping that of appended publications or contents not related to the thesis.

[**a**] L. Hagström, **T. Norlund**, R. Johansson, *Can We Use Small Models to Investigate Multimodal Fusion Methods?*
In Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 45–50, Gothenburg, Sweden. Association for Computational Linguistics, 2022.*

[**b**] A. Stenbom, M. Wiggberg, **T. Norlund**, *Exploring communicative AI: Reflections from a Swedish newsroom*
Digital Journalism*, 2021.*

Acknowledgment

The decision to do a Ph.D. was not a light one, and came only after years of careful consideration. However, since starting I have not regretted it a single time, and that is very much thanks to the people I surround myself with.

Firstly, I owe a lot of gratitude to my main supervisor Richard Johansson. I could not have asked for a better supervisor, who is always encouraging, keen to jump into discussions, and supportive in all other regards. Secondly, I would like to thank my industrial supervisor Staffan Truvé for giving me this opportunity, for your support, and for your endless optimism that I find so inspiring. I'm also grateful for my co-supervisor Marco Kuhlmann for providing such direct and constructive feedback, which very much helps me develop as a researcher.

While doing a Ph.D. partly from remote can be challenging at times, my setup does allow me to get close to *a lot* of great people. I would like to thank all my current and former colleagues at Recorded Future for your encouragement and care, and in particular Aron, Simon, Joakim, Fredrik, Danila, Kajsa and Cathrine from Text Analytics but also Anna, Anders, Mats, Josefin, Johan H, Johan F, Johan G, Ria, Jesper, Lars and everyone else that I cannot fit on this page. This also extends to my closest Ph.D. project partners, Lovisa and Ehsan. I am very happy for our respective collaborations, and for sharing this journey with you.

I am also so grateful to AI Sweden for sharing their office space with me. Thank you Ariel, Amaru, Magnus, Tim and Francisca for providing such an exciting working environment, for the community you are building and for all the interesting discussions we are having.

Last but not least, I want to thank my wife, Jorun. I would not have undertaken this exciting but demanding challenge without your unconditional support.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

Language is core to human communication and development. Having an effective and efficient means of sharing information between individuals and across generations, has led humanity to the world-dominating position it now occupies. Language might also be a critical component in the development of intelligence, as it imposes an abstract structuring of the world that allows for complex manipulation and reasoning.

Thus far, humans exclusively possess the capability to use complex natural languages. However, with the recent developments in Artificial Intelligence (AI) and Machine Learning (ML), rapid progress has been made toward having systems capable of both meaningful natural language understanding as well as generation. The field of Natural Language Processing (NLP) studies methods for computers to perform various tasks involving natural language, typically in the form of text. Examples include tasks such as question-answering, commonsense reasoning, and natural language inference. More recently, tasks involving generation of language are gaining traction, such as machine translation and automatic summarization.

At the center of today's state of the art in NLP are so-called *language models*. A language model is a general model of language, flexible enough to adapt to tasks of different shapes and forms. Whereas traditional methods in NLP are typically very task-tailored, language models can generalize across many tasks. As we will see, language models can be very powerful when trained in a self-supervised way on large quantities of data. In fact, a core driver of progress within NLP in recent years has been to massively scale up the data and computing power that goes into training these general language models.

Common to most NLP tasks is the problem of extracting the underlying meaning from some language input, to be able to produce a desired output. For example, if faced with the question "What is the capital of Sweden?", a non-trivial interpretation has to be performed to retrieve the correct answer from some memory. Foundational to modern language representations are theories from distributional semantics and the so-called Distributional Hypothesis (Sabbettai Harris, 1954; Firth, 1957). The distributional hypothesis states that *the meaning of words is characterized by the contexts they occur and are used*

in. All language models build representations that are fundamentally grounded in context. Depending on how one chooses to quantify context in language modeling, we should expect different qualitative behaviors in the resulting model. In this thesis, we will study different types of contexts in language modeling, and how context affects the resulting language model behavior.

Of particular interest is the relationship between context and memorization. Many NLP tasks require information not directly available in the input, such as the name of Sweden's capital in the question above. For a system to solve such tasks, the information needs to be available and memorized somewhere. What information needs to be memorized is very dependent on the task, and in this thesis, we will use "knowledge" to refer to any information required to solve a specific task that is not immediately available as input. We are interested in two research questions related to knowledge representation and memorization:

RQ 1: **How should we represent and store knowledge in NLP systems?**

RQ 2: **From what data can we acquire knowledge?**

Regarding the first research question, it has been shown that language models are capable of memorizing information through the training process, in which the model parameters act as storage (Petroni et al., 2019). However, this *parametric* memory type has several shortcomings as we will discuss, including being brittle as well as costly. As an alternative, a proposal is to externalize memory from the language model parameters, and to equip the model with an explicit recall mechanism. This mechanism then *augments* the input (or context) to the language model with additional information. Such systems have the potential to address many of the shortcomings of their parametric counterparts.

The second research question is related to the data from which to acquire knowledge and learn to perform language tasks. Specifically, we hypothesize that there is a benefit in learning from not just a textual modality, but also from a visual modality, e.g. images, in certain NLP tasks. We treat this as independent of the first question, but will also study contextual augmentation together with multi-modal learning.

The thesis is structured as follows. In Chapter 2, the relevant technical background is introduced on language models and specifically related to memory and context augmentation. In Chapter 3, we relate the included papers to the overarching research questions and present their specific perspectives. Finally, in Chapter 4 we discuss the conclusions and potential avenues for future work.

Chapter 2

Background

2.1 Language models and Transformers

In this section, we will introduce language models on a more technical level, and the different variants that are relevant to this thesis. We will also introduce the neural network architecture that is most commonly used to implement language models, the Transformer.

2.1.1 Language model formalization

A language model is a probabilistic model of language, typically in the form of text. The text is discretized into a sequence of symbols or *tokens* from a fixed-size vocabulary, through a process called tokenization. Let's consider a sequence of discrete random variables $X_1, ..., X_n$, where each X_i can take values from a vocabulary $x_i \in \mathcal{V}$. In the original formulation, a language model models the *joint* probability of this sequence of random variables:

$$P(X_1 = x_1, ..., X_n = x_n)$$

We will use a simpler notation, $P(x_1, ..., x_n)$, with the random variables dropped to mean the same as the above. The joint distribution can then be factorized using the chain rule of probability:

$$\begin{aligned} P(x_1, ..., x_n) &= P(x_1)P(x_2|x_1) \cdots P(x_n|x_1, ..., x_{n-1}) \\ &= P(x_1) \prod_{i=2}^{n} P(x_i|x_1, ..., x_{i-1}) \end{aligned}$$

We can now parameterize and approximate the factors using a neural network f_θ, with some parameters θ.

$$P(x_i|x_1, ..., x_{i-1}) \approx f_\theta(x_1, ..., x_{i-1})$$

In this formulation, we get an *auto-regressive* language model, that learns to predict the next token in a sequence. This distribution can be sampled iteratively to generate text from the model.

A similar formulation is the *conditional auto-regressive* language model. Here, we also model the distribution of a sequence of tokens $x_1, ..., x_n$, but now conditioned on a separate sequence of tokens $y_1, ..., y_m$.

$$P(x_1,...,x_n|y_1,...,y_m) = \\ = P(x_1|y_1,...,y_m)\prod_{i=2}^{n} P(x_i|x_1,...,x_{i-1},y_1,...,y_m)$$

Similarly, the factors are approximated by a neural network, g_θ.

$$P(x_i|x_1,...,x_{i-1},y_1,...,y_m) \approx g_\theta(x_1,...,x_{i-1},y_1,...,y_m)$$

In a third formulation, we model the distribution of a subset of $(l < n)$ tokens, conditioned on the others:

$$P(\{x_{k_i}\}_{i=1}^{l}|x_1,...,x_n \setminus \{x_{k_i}\}_{i=1}^{l}) \qquad k_i \in \{1,...,n\} \\ \approx h_\theta(x_1,...,x_n \setminus \{x_{k_i}\}_{i=1}^{l})$$

We will denote the resulting model, h_θ, as an *auto-encoded* language model. Auto-encoded language models can be used for predicting masked or missing words in a text, for example, to probe for knowledge in language models.

Training models to predict next or missing tokens, conditional or not, are all different so-called *pre-training tasks*, used to create general models that are more easily adaptable to different downstream tasks. For example, pre-trained auto-encoded language models are also commonly used to create general-purpose vector representations of words or texts, useful for e.g. text or token classification problems. More on this in Section 2.1.3.

2.1.2 The Transformer architecture

The dominating neural network architecture for language models is the *Transformer* (Vaswani et al., 2017). At its core, a Transformer takes as input a sequence of vector embeddings and processes these through a series of layers. Each layer consists of so-called *attention* and *feed-forward* blocks. Input tokens are mapped to trainable vector embeddings before being fed as input to the Transformer. In each attention block, all embeddings are updated with information from the others, and the network can learn to attend its focus to only certain others. This inductive bias has proven particularly powerful for the processing of text. Intuitively, one can imagine this to be useful for representing co-references within a text, that the network implicitly learns to resolve through the attention mechanism. The feed-forward block has been shown to play a central role in parametric memorization (Geva et al., 2021; Meng et al., 2022a), and can be seen as a linear associative memory (Kohonen,

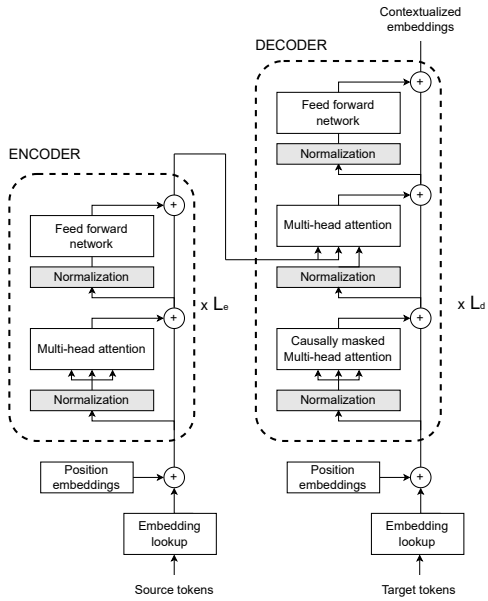

Figure 2.1: An illustration of a Transformer, consisting of an encoder and decoder. The output embeddings are called "contextualized" as they depend on source tokens and previous target tokens through attention layers.

1972). The output of the Transformer is *contextualized* embeddings, as the embedding of each token now depends on the other embeddings via the attention blocks. In a language model, we map the contextualized embeddings to a probability distribution over vocabulary items, typically through a linear projection followed by normalization.

A Transformer can consist of an encoder, decoder, or both. In a Transformer encoder, the attention is unconstrained in the sense that all embeddings can attend to all other embeddings. Auto-encoded language models are implemented as Transformer encoders. In a Transformer decoder, the attention is constrained such that an embedding can only attend to its preceding embeddings. This makes them suitable for autoregressive language models, as they should only depend on preceding tokens. In an encoder-decoder Transformer, illustrated in Figure 2.1, we feed some input sequence $y_1, ..., y_m$ to the encoder, and $x_1, ..., x_n$ to the decoder. To be used as a conditional auto-regressive language model, the decoder now incorporates an additional attention block in each layer. In the so-called *cross-attention*, decoder embeddings are updated by attending to the contextualized encoder embeddings. This makes the decoder predictions also be conditioned on the encoder sequence.

Throughout this thesis, we will use the term "language model" interchangeably to refer to any of the three presented types, depending on context.

2.1.3 Large-scale pre-training

In all language model formulations in Section 2.1, we are estimating the probability of some tokens given other tokens. We can formulate self-supervised objectives for each formulation, and train corresponding models using only raw text data. As has been shown in the last couple of years, pre-trained language models (PLM) transfer easily to many downstream NLP tasks. For example, fine-tuning a PLM on labeled data for some task typically performs better than training the same model from scratch. Furthermore, auto-regressive PLMs, and in particular large Generative Pre-trained Transformers (GPT) (Radford et al., 2019; Brown et al., 2020), demonstrate strong few-shot and zero-shot performance on downstream tasks, without any fine-tuning at all.

Scaling the PLM by increasing the number of trainable parameters and training it on more data, also typically increases downstream performance (Kaplan et al., 2020; Hoffmann et al., 2022). This applies to all types of language models, but in particular for the GPT family of models (Radford et al., 2019; Brown et al., 2020; Hoffmann et al., 2022; Rae et al., 2021; Chowdhery et al., 2022). More recently, instruction finetuning has proven effective to further increase few and zero-shot performance (Wei et al., 2022).

For auto-encoded language models, the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019) and others (Liu et al., 2019; Lan et al., 2020) have become ubiquitous as general language representation models for a breadth of applications such as document classification, textual similarity search, and named entity recognition.

For conditional autoregressive language models, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) are examples of pre-trained variants applied for conditional generative tasks, such as summarization and dialogue response generation.

All in all, PLMs provide a foundation for progress in NLP today, much due to their strong transfer-learning properties.

2.2 Knowledge representation and memorization in language models

Natural languages are used to communicate ideas, experiences, and knowledge about the world we live in. Thus, language is necessarily tightly coupled with the information it carries. For any NLP task, knowledge about the world is always required to some extent, explicitly or implicitly.

More technically, a language model is trained to *associate* some input (tokens) to some output (a masked or next token). Performing this associative task with high performance requires different types of knowledge. Consider the sentence: *The typical color of a banana is yellow.* If we mask the word "of", and ask an auto-encoded language model to predict it, the model has to have some syntactic knowledge to succeed. On the other hand, if we mask the word "yellow", it requires more semantic (or factual) knowledge. Knowledge in general is thus very tightly coupled with the language modeling task.

Language modeling, as seen from the perspective of learning associations, is very dependent of the context provided as input. Consider the following question, taken from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016):

> Which name is also used to describe the Amazon rainforest in English?

To be able to answer this question, the model has to have *a priori* formed an association to the correct answer. However, if we treat this as a reading comprehension task (which SQuAD actually is), an extract of a related Wikipedia page is also provided:

> The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America.

In this setting, the task is reduced to being able to interpret the question, locate the correct answer in the provided context, and output it. The additional context reduces the need for associations being formed a priori in the model. However, while the direct association to "Amazonia" is no longer required to be pre-existing in the model, the task still requires other pre-existing associations. For example, to only output English names, the model has to implicitly infer that other languages distinct from English exist, and that only English names are valid answers to the question. The point is however that we can reduce the amount of pre-existing associations required, by providing more informative context to the model. By engineering a mechanism to "augment" the context of a language model with relevant information or features, we can effectively achieve an *externalization* of certain types of associations. This is also sometimes referred to as externalizing the *memory* from a language model (Khandelwal et al., 2020; Yogatama et al., 2021; Borgeaud et al., 2022).

So what should a language model actually model? Assuming we want a general NLP system in possession of general knowledge about the world that is helpful to solve a breadth of tasks, how should we design the system to incorporate such knowledge? Should we design this system as a language model monolith, with high parametric capacity to form all kinds of associations, or should we try to restrict associations formed parametrically?

Seen from a machine-learning perspective, this externalization has significant implications for the learning problem of the language model. By introducing additional relevant features, the model will likely learn to use information provided as context to a greater degree. An interesting question is then whether externalizing memory can reduce the need for parametric capacity, such that we can shrink the model size with kept associative performance.

In the following sections, we will try to dissect the pros and cons of having a fully parametric versus externalized memory in language models. As guiding principles, we will relate the two alternatives to a set of desirable properties that we believe are important for many NLP applications.

1. **Generalization** Natural languages are highly flexible in that e.g. statements can be paraphrased or expressed in many different ways, depending on context. Ideally, language models should generalize to output text of different semantically equivalent forms, and not just verbatim repeat from the memory.

2. **Grounded language model output and attributability** A fundamental problem with current language models is their tendency to output text that is false or otherwise unintended, a problem also known as *hallucination*. In particular for generative tasks, we typically want statements in the output to consistently adhere to some source. This is related to attributability, i.e. the ability to trace back a statement or claim to its source, for reference and contextualization.

3. **Memory management and updatability** An important property of *factual* knowledge is that it is changing over time. For many NLP applications, it is therefore important to be able to control for this, and update the system whenever needed. However knowledge is stored, it should ideally support adding, deleting, and modifying operations.

4. **Resource efficiency and costs** Current large language models can cost millions of dollars to train, and several cents for using it to generate a single piece of text. The high costs inevitably lock out many use cases and applications. To maximize the applicability of language models, it is beneficial to optimize their resource efficiency.

2.2.1 Parametric memorization in language models

As described in section 2.1.3, language models can be pre-trained using some type of self-supervised learning. This typically entails learning to "complete" a text in some way. In autoregressive models, a proceeding token is predicted, while in auto-encoded models, a masked token at any position can be predicted. For certain texts, predicting the next or a masked token will require a knowledge recall. As an example, we can rephrase the question from before as a statement: *"In English, the Amazon rainforest can also be called [MASK]"*. For the model to predict the correct word, the knowledge needs to be memorized somehow. In conventional language models, the only place where such knowledge can be "stored" is in the trainable parameters of the model. Therefore, using such self-supervised objectives incentivizes language models to memorize knowledge in their parameters, as part of the pre-training. We refer to a language model as *parametric* when only its trainable parameters act as implicit memory, and no contextual augmentation is taking place.

The ability of pre-trained language models to recall factual and commonsense knowledge can for example be studied using the LAMA benchmark (Petroni et al., 2019). In Petroni et al. (2019), it is demonstrated that BERT is able to recall relational knowledge zero-shot, i.e. without any fine-tuning, at a level competitive with non-neural and supervised alternatives. In another study, a conditional auto-regressive language model (T5) is fine-tuned to

answer questions in a "closed-book" setting, i.e. without conditioning on text containing the answer (Roberts et al., 2020). Also here, it is shown that parametric memorization is a viable approach to knowledge memorization and recall. However, the performance on factual recall benchmarks is known to fluctuate depending on the exact prompting of the model (Jiang et al., 2020), indicating there is room for improvement in the robustness of knowledge representations.

The memory capacity also seems to scale with model size. GPT-3, which is an auto-regressive language model with an order of magnitude more parameters, significantly outperforms the results from Roberts et al. (2020), despite being zero-shot (Brown et al., 2020).

Using only a textual prompt, the model can be tasked to perform just about any NLP task zero-shot, often at competitive performance to a smaller but fine-tuned baseline. The model can also adhere to the textual format and style of the prompt, suggesting its language understanding goes beyond pure lexical memorization of its training data and that more abstract concepts are implicitly represented and memorized in the parameters.

For many applications, source attribution of generated output is a necessity to contextualize claims and assess their reliability. However, as we currently lack good tools for explaining the predictions of parametric language models, it is hard to know on what grounds the model's statements are made.

The non-transparency of a parametric memory also makes it difficult to support modifying operations such as adding, deleting, or replacing. Despite this, methods for *knowledge-editing* parametric language models have been proposed (Mitchell et al., 2022; De Cao et al., 2021; Meng et al., 2022a), and for example MEMIT (Meng et al., 2022b) display promising results including few side effects on the rest of the language model's knowledge.

As the parametric memory is built throughout the training of the language model, it requires passing all knowledge-containing data through one or several forward and backward passes. The computational cost for a forward/backward pass grows with the size of the model, and the model size grows with the size of the knowledge, as the parametric capacity needs to be adjusted for the amount of knowledge to be memorized. This becomes very resource intensive with scale. Methods such as scaling model size without additional compute by sparsely activating parts of the network have for example been proposed (Fedus et al., 2022) to reduce compute, but all knowledge-containing data still need to pass through training.

To summarize, implicit parametric memorization yields language models with some capacity for generalizing e.g. factual recall to novel prompts not seen during training, but currently has weaknesses relating to attribution, updatability, and resource efficiency. In this thesis, we are interested in whether the same weaknesses can be remedied by means of contextual augmentation, such that knowledge is to a large extent externalized into a memory representation better fit for purpose.

2.2.2 Contextually augmented language models

In contrast to parametric memorization, a contextually augmented language model has an additional mechanism to dynamically manipulate the input (or context) to the language model. For example, this augmentation could be to look up documents related to a question in which the answer might be found, and append to the input somehow. From the perspective of the language model, this would reformulate the problem above from a closed-book setting to a reading comprehension setting. Intuitively, reformulating tasks that involve knowledge recall into reading comprehension should reduce the need for parametric memorization capacity in the language model, as relevant context is explicitly provided as input.

This effectively results in an externalization of memory. However, many questions arise relating to how this can be realized in practice. Firstly, as knowledge is tightly connected with language, it is not clear exactly what should be externalized and what should be parametrically learned. Secondly, how should the augmentation mechanism be designed to provide the language model with what it does not know? Thirdly, how should we effectively train a language model with context augmentation? These are all questions that we will address in the papers accompanying this thesis.

Context augmentation has several intriguing theoretical properties compared to parametric language models. For example, if the language model grounds its output in the information it receives as input, the input can be used for interpretation and attribution. Secondly, the external memory can in principle be designed to trivially support adding, deletion, and substitution, by for example using a database or search engine. Thirdly, as is for example demonstrated in RETRO (Borgeaud et al., 2022), the number of trainable parameters can be reduced with kept language modeling performance, by means of retrieval.

Augmentation can happen at different levels of granularity and integrate with the language model in different ways. At the highest level, the context can be augmented at the prompt level. In question answering for example, the input is the question which is then augmented to also include some retrieved documents (Guu et al., 2020). Augmentation can also happen at the token level (Khandelwal et al., 2020; Yogatama et al., 2021). For example in kNN-LM (Khandelwal et al., 2020), which is an autoregressive language model, the contextualized embedding from the Transformer is used to retrieve tokens with a similar context, and use statistics of their next-tokens to adjust the output distribution. In this case, the integration is at the top of the network, rather than at the bottom (input). Finally, in e.g. RETRO (Borgeaud et al., 2022) retrieved context is augmented at a *chunk*-level, such that it is shared between m consecutive tokens.

Context augmentation is not restricted to just retrieval. As we will see, we can instead use a separately trained model that augments a language model with "visual knowledge" represented as embeddings. While this separately trained model is fully parametric, we distinguish this setting from a purely parametric language model as they are separately trained. Parametric models

are otherwise commonly distinguished from *semi-parametric* models. In a semi-parametric language model, the language model is contextually augmented, but the augmentation is constrained to be non-parametric. Non-parametric refers to memory types that are not represented by trainable parameters, i.e. a relational database or similar. As we are not only interested in non-parametric memory types, we decline from using the term semi-parametric in this thesis, and instead focus on the contextual augmentation regardless of memory type.

2.3 Acquiring knowledge from multiple modalities

A second theme in this thesis focuses on the data from which to acquire certain types of knowledge. Humans learn about the world by perceiving and interacting with it. Knowledge can for example be formed through communication using language, either by reading or listening. But knowledge can also be formed by directly perceiving the world, without any linguistic signal. For example, if you are asked what is the color of your parent's house, you can likely respond even if this was never stated verbally in the family. Humans have the capability to (verbally or textually) express knowledge that was acquired non-linguistically.

Today's most popular language models are all trained solely on text, and thus base all their knowledge on what is described linguistically in this data modality. Using only text for language learning and as the single ground for knowledge acquisition has been criticized (Bisk et al., 2020; Bender and Koller, 2020). It is for example claimed "today's best systems still make mistakes that arise from a failure to relate language to the physical world it describes and the social interactions it facilitates". A proposal is thus to broaden the types of data and to include other modalities such as images or video, with the hypothesis to for example reduce hallucination.

Additionally, using solely text may suffer from the problem of reporting bias. Reporting bias is the bias humans implicitly adhere to when deciding what to write or report on (Gordon and Van Durme, 2013). For example, humans tend to only write or report on information that they believe a reader can learn from. All written documents assume some level of pre-existing knowledge, and may only focus on parts that are novel, sensational, or surprising in some way. This has implications for what types of knowledge are to be found in written form. There is potential in reducing reporting bias by extending sources to include non-linguistic modalities as well.

In this thesis, we will investigate how knowledge can be acquired from a visual modality in addition to text, and to what extent this knowledge is complementary to what can be acquired from just text. We will treat this as orthogonal to the way of incorporating the knowledge in a language model, i.e. RQ 1. In the following section, we provide some background on vision-and-language modeling.

Multi-modal Language Models

Training models using multiple modalities can be beneficial in several regards. Models capable of processing multiple modalities are of course applicable to tasks requiring multiple modalities, but training models using multiple modalities can also be beneficial towards unimodal tasks. For example, language can assist in creating high-quality visual representations, such as in CLIP (Radford et al., 2021). This was for example demonstrated with state-of-the-art performance in zero-shot image classification tasks. It is however less clear whether also the opposite holds. Given the success of self-supervised pre-training of language models using large text corpora, can a visual modality assist in further improving language representations?

We are interested in multi-modal language modeling to investigate whether a visual modality can complement the textual to create better language representations. And specifically for knowledge representation, how does parametric memorization compare to context augmentation in the multi-modal case?

In multi-modal language modeling, we have text paired with data of an additional modality, such as an image. Depending on the task, there are different modeling setups. In image captioning, we have a conditional auto-regressive language model, where the conditioning is now with respect to an image instead of discrete tokens. The model is then trained to predict text that describes the given image. If we flip the modalities, we instead get text-to-image generation (Ramesh et al., 2021; Wang et al., 2018). In tasks such as visual question answering, the model takes both a textual question and an image, and produces an answer based on the image (Antol et al., 2015; Hudson and Manning, 2019)

The modalities can also be integrated in different ways. In VisualBERT (Li et al., 2019) for example, a representation of the image is appended to the input of the Transformer model. CLIP (Radford et al., 2021) instead consists of a separate textual and visual encoder, such that the text and image representations are independent.

As self-supervised pre-training is a successful approach to creating general-purpose language models, the same can be applied to multi-modal ones. VisualBERT is for example pre-trained using a masked multimodal language modeling objective, in which the model learns to utilize also the visual context to predict masked tokens. In CLIP however, a contrastive loss is used to align representations of the text and image into a joint vector space.

Chapter 3

Summary of Included Papers

3.1 Paper I

In this work, we train a large auto-regressive language model in the Swedish language, using data from the online forum Flashback. We use this paper to exemplify a typical parametric language model.

Flashback data is used for the purpose of obtaining a *conversational* system, that can respond to questions and take part in discussion. The resulting model is called Flashback-GPT.

We perform a human evaluation to assess the general quality of the generated output. The evaluation tries to asses the general conversational abilities of the model, in which the model often has to demonstrate factual as well as linguistic knowledge obtained from its pre-training. For this purpose, we specifically seek to two investigate two dimensions, that we call *humanlikeness* and *informativeness*. These are similar to sensibleness and specificity, as evaluated by Adiwardana et al. (2020). In humanlikeness, we want to understand to what extent a generated response is indistinguishable from a human response. Anything in a response that signals a human is unlikely to have written it, be it a commonsense error, lack of coherence etc. would flag a response as not humanlike.

Responses can be humanlike but less interesting. Often we would like a conversational system to produce more exhaustive responses than for example "Yes", "No" or "I don't know". We therefore also evaluate informativeness, where we ask whether a generated response *adds information* to the discussion. This is more directly targeting the model's ability to recall and output relevant knowledge.

As for our results, responses were deemed humanlike in 68% of the cases, and should be compared to human-written responses of 95%. 48% of generated responses were deemed informative compared to 83% of human responses. While the human evaluation is small scale, the results suggest a non-trivial

ability to converse informatively in Swedish across a wide range of topics.

In the context of our research questions, informative responses by the model suggest a capability to memorize, recall and express for example factual knowledge from the pre-training. However, as only 48% of the responses were deemed humanlike and informative, there is a large room for improvement. As has been shown in for example LaMDA (Cohen et al., 2022), fine-tuning the model to assess itself on sensibleness and specificity can be an effective method to improve performance further. While it was not investigated in this paper, they also show that adding a contextual augmentation mechanism improves groundedness, a measure of attributability that is also interesting from the perspective of our research questions.

Contributions TN came up with the idea of training a conversational language model on Flashback. He also performed the data collection, model training, designed the human evaluation and wrote the main parts of the paper.

AS contributed with ethical perspectives on the potential use and misuse of large language models, and wrote that corresponding section of the paper.

3.2 Paper II

In Paper I, we developed a parametric auto-regressive language model and evaluated some qualitative properties of its generated output. In this work, we target both RQ 1 and RQ 2, and propose a novel knowledge-centric language benchmark, that allows us to systematically investigate cross-modal transfer.

Most studies on the ability of language models to recall factual knowledge focus on learning from a textual modality only. But what if this information is more accessible in a different data modality than text? In this work, we ask whether we can improve factual associations of a language model – using images instead of text. Specifically, we investigate the extent to which a multi-modal language model can *textually express knowledge originating from visual data*. We denote this as the language model is "transferring" knowledge from a visual to textual modality.

Toward this end, we create a text-only cloze-style knowledge probing task that we denote *Memory Colors*. In Memory Colors, the task is to predict the prototypical color of objects such as bananas, snow or coal, without any image in context. The language model is asked to predict the masked word in constructed sentences like "The typical color of a banana is [MASK]". We only include objects for which there is a strong human consensus on their prototypical color. We chose to probe for memory colors as it is a type of knowledge that a visual modality is likely to embed, which means it constitutes a good testbed for our purpose.

We train a multi-modal auto-encoded language model that we denote CLIP-BERT. CLIP-BERT is fundamentally a BERT model but trained on a large parallel corpus of images and captions. The images are encoded using the image encoder of CLIP (Radford et al., 2021) and are concatenated with the token embeddings to form the input to the BERT model.

We evaluate CLIP-BERT in two variants. In the first variant, we evaluate to which extent the model has learned to associate each object with its prototypical color *implicitly* through the multi-modal pre-training. We can think of this variant as the model parametrically memorizes this information through pre-training, similar to how other factual associations are formed through textual pre-training.

In the second variant, we encode the masked sentence through the CLIP text encoder, and concatenate it with the input to the model, in the same way as representations of paired images are concatenated during training. This can be seen as a contextual augmentation, as we *explicitly* add additional information to the BERT model based on the input. Since CLIP is trained to align the representational spaces of the text and image encoder, this can be thought of as a representation of an "imagined" matching image. To be able to attribute correct predictions to images rather than text, careful filtering of the multi-modal training data is performed to remove examples where the memory color knowledge is expressed in text. For the model to achieve high performance, it needs to have implicitly memorized the prototypical colors from the training images. We find that for this type of visual knowledge, the contextual augmentation worked much better than the parametric variant, with

performance close to perfect on Memory Colors.

To summarize, in this work we investigate both RQ 1 and RQ 2:

1. **RQ 1**: We augment the language model with representations from a separately trained model (CLIP text encoder). Since CLIP's representational space clearly discriminates memory colors, the encoder effectively acts as a memory from which this information can be retrieved. This work shows that the text encoder of CLIP can work as an external memory module to a language model, and that using it can improve performance on Memory Colors substantially.

2. **RQ 2**: The standard BERT model performed very badly on Memory Colors, suggesting this type of knowledge is not well represented in the textual corpuses BERT is trained on. While this type of knowledge of course can be learned from an appropriate textual corpus as well, this work showcases that multi-modal data can positively complement the knowledge of a textually trained language model.

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

LH 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.

RJ provided supervision on the work and writing for the paper.

3.3 Paper III

In Paper II, we investigated methods for transferring knowledge from vision to language. In this work, we are also interested in cross-modal transfer, but now in a classification setting. We focus on the task of protest detection, which is a type of socio-political event detection. Protests can be detected in both text and images. In text, we are typically interested in reports or mentions of specific protests, and in images, protests are typically depicted by large crowds on streets with placards.

We treat protest detection as a binary classification problem, using either a textual sentence or an image, and formulate two research questions:

1. To what extent can the performance of a unimodal protest detection model transfer from one modality to another?

2. Can unimodal detection of protests be improved by using a multi-modal protest detection model?

To investigate this, we pool two existing protest detection datasets, for textual and visual detection respectively. We also use representations from the visual-and-language alignment model CLIP, and train a linear classification layer on top of the respective encoders. If we train the classifier on protest text data, we can obtain a visual protest detection model by swapping the CLIP text encoder for the image encoder. This is because CLIP's representational space is trained to be aligned for text and images. This also works the other way around, e.g. train on protest images, and swap the image encoder for the text encoder. We can use this method to evaluate the ability to transfer unimodal protest detection performance from one modality to another.

To investigate whether unimodal protest detection can be improved by using multi-modal data, we train the same model on both datasets simultaneously. In this setting, a data point is processed through its respective CLIP encoder and classified using the same joint layer.

Our results show protest detection performance can be transferred across modalities with high performance, in particular from text to image. On the contrary, the multi-modal training does not yield better results than the unimodal baseline, suggesting little complementary effects between the modalities.

Contributions This work was an extension of KA and RR's Master's Thesis project, which TN, RL, and RJ co-supervised. KA and RR implemented the code and ran the experiments, and made major contributions to the paper. The main ideas underlying this work were formed by TN, in conversation with AL. TN also made major contributions to the paper, specifically the abstract, introduction, and ethical statement. RJ also contributed to the paper by reviewing and providing feedback.

3.4 Paper IV

In paper II, we investigated contextual augmentation of an auto-encoded language model and showed its viability on a knowledge probing task. Recently, it has been shown that augmenting auto-regressive language models with retrieval can be effective to improve language modeling performance (Khandelwal et al., 2020; Yogatama et al., 2021; Borgeaud et al., 2022; Pan et al., 2023).

In all previous works, retrieval has been shown to reduce the perplexity (or similar metrics) of language models. This is an interesting finding and spurs questions to what degree language modeling is just a matter of memorizing training data (Tänzer et al., 2022). If a big portion of the parametric capacity of language models is used for verbatim memorization of training data, there is a potential for retrieval-augmented models to be downsized with kept perplexity. Retrieval augmentation is also intriguing from a memory management perspective as the retrieval database can trivially support addition, deletion and modifying operations. Furthermore, there is also potential to reduce hallucination if the output is more grounded and attributable to the retrieved context.

With these potential benefits relating to reduced footprint, grounding and updatability, an important question is what effect retrieval has on their generalization. The "soft" nature of parametric memorization makes language models very powerful in this regard. In contrast, with a retrieval augmentation, recall of information is "harder" in that only a fixed set of documents are retrieved to condition the generation.

In this work we take a focused look at generalization in one of the most recent retrieval augmented autoregressive language models, RETRO (Borgeaud et al., 2022). RETRO, with 7.5B parameters and a 2T token database for retrieval, achieves perplexity matching a 175B parameter GPT-3 model on the Pile. Specifically, we seek to better understand how this can be, and what tokens benefit from retrieval.

RETRO is a conditional auto-regressive language model, in which retrieval takes place at regular token intervals during generation. Using *chunked cross-attention*, the decoder can attend and incorporate information from retrieved context in the encoder. This way, if retrieval is removed, the model is reduced to a standard decoder-only autoregressive language model.

By comparing the token-wise perplexity with and without retrieval, we can better understand in what contexts retrieval is most helpful. We find that perplexity is reduced predominantely for tokens that overlap verbatim with retrieved context. The model effectively learns to exploit this by learning to copy in such cases. While verbatim overlap was found to be a core contributor to reduced perplexity in Borgeaud et al. (2022), their results were suggestive of some non-trivial generalization as well. In our analysis however, perplexity reduction can almost exclusively be attributed to overlap, which would suggest less of such non-trivial or semantic generalization. This is in line with the result of a similar analysis of the kNN-LM model (Drozdov et al., 2022).

This insight is important to guide further research within retrieval augmented language modeling, and highlights the need for careful leakage analysis to

assess generalization capabilities.

Contributions TN implementated the RETRO model, and was a major contributor to ideating on and executing the experiments. He also wrote the first draft of the paper.

ED contributed by preparing the data used for training and retrieval, as well as to ideation and discussions throughout the project.

RJ and MK provided supervision and guidance throughout the project, and made significant contributions to the general formalization and storytelling of the final paper.

Chapter 4

Discussion and Future Work

We will now summarize and discuss our research questions with regards to the learnings from the accompanying papers.

RQ 1: How should we represent and store knowledge in NLP systems?

In the accompanying works, we are only scratching the surface on the topic of knowledge representation and memory in language modeling. Large parametric language models (like Flashback-GPT in Paper I), have shown a high degree of generalization, and can produce novel texts of high quality both lexically as well as semantically. They do suffer from problems such as lack of attributability in the output, updatability and low resource efficiency. If we contextually augment the model through e.g. retrieval, we have the potential to improve on all three. However, as we argue in Paper IV, this might also have consequences for the generalization.

One can imagine having both a large parametric as well as an external memory. Can we get the strong generalization from a large parametric language model, together with the attributability and updatability benefits of non-parametric external memory? One way to realize this idea would be to start from a large pre-trained parametric language model, and post-hoc augment it using e.g. retrieval, similar to LaMDA (Cohen et al., 2022), WebGPT (Nakano et al., 2021) and GopherCite (Menick et al., 2022). In this case, however, we would still inherit the high costs of training and running the parametric memory.

Independently of the generalization, we show context augmentation can be more effective for memorizing and recalling visual information, compared to parametric memorization. We use memory colors to demonstrate a case where a textually trained language model's knowledge is poor, and where the information can be found in visual data.

As future work, it would be interesting to further explore how large language models trade off parametric knowledge and augmented knowledge. As an

example, if a model parametrically associates the text "The current president of the US is", with "Donald Trump", how does the model handle a conflicting contextual prompt stating the current president is Joe Biden? Can we develop methods to explicitly trade contextual vs parametric knowledge? This would be very useful for improved grounding and attributable output of language models.

RQ 2: From what data can we acquire knowledge?

The knowledge language models possess originates from data. Text is a very information-dense modality, which is also available in vast amounts. It is however likely that certain types of information are better represented in other modalities. In this thesis, we have investigated this for a visual modality in addition to text. We observe memory colors to be a knowledge gap in BERT, and show that we can enrich the language model with this knowledge, using visual data. We also investigate whether visual data can complement textual (and vice versa) during training of a protest detection classifier, but find this to not be the case in our experiment. We do however observe some degree of substitutability between the modalities, indicating there at least is some information overlap.

It has been argued true natural language understanding cannot be passively acquired from only text (Bisk et al., 2020; Bender and Koller, 2020). Language and words are grounded in their use and contexts. Inspired by this, multi-modal language modeling has been encouraged, where the contextual distribution is widened and words can be tied to e.g. visual perception. In practice, however, large quantities of high-quality multi-modal data is more difficult to collect than unimodal, which is a limiting factor in this direction. Recently, significant progress has been observed in several different vision-and-language problem formulations including image-to-text (Alayrac et al., 2022), text-to-image (Wang et al., 2018; Ramesh et al., 2022) and text-image-alignment (Radford et al., 2021; Jia et al., 2021). Our formulation is different from these, in that we primarily focus on whether multi-modal training is helpful in a text-only setting. We show this to be the case in the specific case of memory colors together with a contextual augmentation, but deem this to be relevant more broadly as well, for example in question-answering tasks such as WebQA (Chang et al., 2022) and MultimodalQA (Talmor et al., 2021).

As our experiments are all small-scale, it is interesting to consider the effect of increasing the scale to more data and modalities. One such example is GATO (Reed et al., 2022), in which a single model is trained on a multitude of tasks involving different modalities, including text and images but also interactive tasks such as playing games. It would be interesting to better understand the extent to which transfer is occurring across these modalities, and in particular transfer into the textual modality.

Bibliography

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

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.

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.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal QA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16495–16504.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vincent Y. Zhao, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. 2022. Lamda: Language models for dialog applications. In *arXiv*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Andrew Drozdov, Shufan Wang, Razieh Rahimi, Andrew McCallum, Hamed Zamani, and Mohit Iyyer. 2022. You can't pick your neighbors, or can you? when and how to rely on retrieval in the kNN-LM. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2997–3007, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

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)*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Teuvo Kohonen. 1972. Correlation matrix memories. *IEEE Transactions on Computers*, C-21(4):353–359.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:1908.03557*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2023. Knowledge-in-context: Towards knowledgeable semi-parametric language models. In *The Eleventh International Conference on Learning Representations*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Zellig Sabbettai Harris. 1954. Distributional structure.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal QA: Complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

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

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373.