



UMEÅ UNIVERSITY

Grounding in a World Full of Meanings – DRAFT

Adam Dahlgren Lindström

DOCTORAL THESIS, 2023
DEPARTMENT OF COMPUTING SCIENCE
UMEÅ UNIVERSITY
SWEDEN

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

dali@cs.umu.se

Copyright © 2023 by authors
Except Paper I, © TODO
Paper II, © TODO

ISBN TODO
ISSN TODO
UMINF TODO

Cover illustration by TODO
Printed by TODO 2023

Nothing is more usual and more natural for those, who pretend to discover any thing new to the world in philosophy and sciences, than to insinuate the praises of their own systems, by decrying all those, which have been advanced before them. A treatise of human nature (Hume, 1739)

Popular Science Abstract

Sammanfattning

Abstract

Preface

Acknowledgements

Contents

1	Introduction	1
1.1	Historical problems – tentative title	2
1.2	Difference between language model and world model	10
1.3	Multimodality	10
1.4	Neuro-symbolic AI	12
1.5	Human-centered AI	14
1.6	Compositionality	16
1.7	What Hill are we Climbing?	16
1.8	Identity of thesis	16
1.9	Research questions	17
1.10	Contributions	17
1.11	Structure	17
2	Literature overview - tentative title	21
2.1	Representations of Meaning	21
2.2	Brief history of language modelling	22
2.3	Needs major revision - Multimodal machine learning and language grounding	22
2.4	Neuro-symbolic machine learning	24
2.5	Knowledge graphs	27
2.6	Compositionality	28
2.7	REMOVE AND KEEP PARTS OF CONTENT - Properties for compositionality and grounding	32
2.8	Challenges and characteristics	34
3	Probing multimodal language models	35
3.1	What can probing tell us?	36
3.2	Probing Multimodal Embeddings for Linguistic Properties	40
3.3	NOT REWRITTEN AT ALL - Bridging Perception, Memory, and Inference through Semantic Relations	53
3.4	Probing to understand compositionality	59
3.5	Challenges and characteristics	60

4	Studying the behaviour of multimodal language models on reasoning and compositionality	61
4.1	NOT REWRITTEN AT ALL - Perception, Memory, and Inference	61
4.2	Using synthetic data	65
4.3	DeepProbLog and compositionality	67
4.4	Multimodal Word Math Problems	69
4.5	NOT REWRITTEN = CLEVR-Math	71
4.6	Challenges and characteristics	80
5	Ontology of Characteristics for multimodal language modelling	87
5.1	Ontology of Characteristics	87
5.2	Chomsky-like hierarchy for grounding and compositionality	92
5.3	Framework	92
5.4	Ideas	93
6	Learning Hierarchical Compositions	95
6.1	Compositional generalisation benchmark using hierarchical pseudoword concepts in CLEVR	100
7	Conclusions	107

List of Figures

1.1	Example of failures of large language models.	2
1.2	ChatGPT given the same prompt as ELIZA.	5
1.3	Illustration of the Peircean model of symbol-referent-thought	7
1.4	Tentative illustration of language model vs. world model.	8
2.1	Example taken from (Thrush et al., 2022)	32
3.1	Image-caption pairs (left) and how vectors representing the words ‘bat’, ‘club’, and ‘bird’ may be affected by the image information (above)	41
3.2	In task <i>SemanticCongruence</i> , the objective is to recognise semantically implausible captions.	44
3.3	In this work we focus on recovering synonyms, hypernyms, hyponyms, and meronyms from natural language models via probing to understand the prerequisites of integration with knowledge bases.	55
3.4	Image from (Sikarwar et al., 2022) todo ask for permission.	60
4.1	The triad of capabilities and components advocated.	64
4.2	Example image generated via Blender using CLEVR. One question generated to this image using the one hop template is “ <i>What is the color of the metal cylinder that is behind the cyan matte thing?</i> ” to which the answer is <i>red</i>	66
4.3	Confusion matrix for DeepProbLog on MNIST	68
4.4	Confusion matrix for DeepProbLog on ColorMNIST using the same color maps for training and testing.	68
4.5	Confusion matrix for DeepProbLog on ColorMNIST using different color maps for training and testing.	69
4.6	Selection of questions generated from this image: (i) <i>Remove all gray spheres. How many spheres are there? (3)</i> ,, (ii) <i>Take away 3 cubes. How many objects are there? (7)</i> , (iii) <i>How many blocks must be removed to get 1 block? (2)</i>	70
4.7	CLEVR-Math example question <i>Take away 2 matte cylinders. How many objects are left?</i> with corresponding mathematical equation $X = 9 - 2$	73

4.8	Example image-question pairs from CLEVR-Math, 4.8a showcase addition and subtraction, and 4.8b shows multihop reasoning. Answers in parenthesis.	74
4.9	The attributes are used evenly throughout the dataset, whereas the answers are biased towards the smaller numbers. The numbers are aggregated over all splits.	81
4.10	Samples from the test set of CLEVR-Math.	82
4.11	Examples of when CLIP and NS-VQA fails on multihop questions.	83
4.12	Sampling of incorrect answers by CLIP on multihop.	84
4.13	Sampling of incorrect answers by NS-VQA on multihop.	85
4.14	Confusion matrix for CLIP trained on 20 000 samples.	86
6.1	Example of card from the Wug Test (Berko, 1958), showing a task of applying morphological rules to a novel (pseudo)word.	96
6.2	TODO figure where GPT is asked to compute additions.	99
6.3	A simple example of data generated in CLEVR, where we see two pseudoconcepts; a) a <i>blargh</i> – two small cubes next to each other, and b) a <i>perde</i> – a large cyan sphere.	101
6.4	Example of a pseudoconcept hierarchy over the CLEVR vocabulary.	103
6.5	Illustration of how a curriculum can be used to investigate the effects of training on compositional generalisation.	104

List of Tables

3.1	Probing tasks for semantic embeddings, organized along three broader probing categories as investigated in Conneau et al. . . .	37
3.2	Overview of the investigated embeddings. The total size of the model, including models used to extract precomputed image features, is given in parenthesis.	46
3.3	Probing accuracies using a MLP with embeddings as input. The bottom three show for each model the difference between the best unimodal and the best merged embedding. All results are averaged over 5 runs and have variance ≤ 0.01	48
3.4	Accuracy per label of the tested models A more detailed account of the accuracy of the tested models for the task <i>NumObjects</i> . The class labels correspond to the number of objects annotated in the image..	50
3.5	Instances of the relations synonymy, hypernymy, and meronymy extracted from WordNet.	56
3.6	The probing accuracy on the semantic relations, with variance given in parentheses. The accuracy of a “largest class” strategy is shown next to each relation. All transformers give embeddings of 768 dimensions, with word2vec and GloVe using 300 dimension. Each relation contain 1712, 306, 2740, and 1630 samples, respectively.	58
4.1	Accuracy on evaluation data for both	68
4.2	An overview of the different templates implemented by CLEVR-Math. $\langle Z \rangle$, $\langle C \rangle$, $\langle M \rangle$, $\langle S \rangle$ are instantiated to size, color, material, and shape during the question generation.	75
4.3	Distribution of templates in each data split.	76
4.4	Huggingface dataset card for CLEVR-Math.	77
4.5	Accuracy on the CLEVR-Math dataset, shown for each template group and aggregated over all templates.	78
4.6	Accuracy over all templates for different dataset sizes.	78

Todo list

Chapter 1

Introduction

Example of how to use quotes at
the beginning of chapters

dali

One reason why learning language in visual settings leads to better compositional generalisation is that we speak about the exact same objects using different words in the word/concept hierarchy. When referring to an image of a cat, a speaker might use *norwegian forest cat*, *cat*, and *animal*, all in the same sentence structure depending on what level of resolution the context requires. Hence, using vision, a model has a clearer signal that these words all relate to each other more so than in the way they can all be used in the same way in a sentence to mean the same thing. However, this does not necessarily mean that it is clear whether a *cat* is an *animal*, or whether an *animal* is a *cat*.

TODO too straight on We have spent centuries systematically trying to understand and characterise human intelligence, how we understand our surroundings and make sense of each other. It is no surprise that Natural Language Processing (NLP) is at the core of many applications of artificial intelligence (AI). We have seen NLP playing a central role in many of the famous contributions to the field in general, such as ELIZA (Weizenbaum, 1966) and TODO. The last decade of AI research has been characterised by the domination of deep learning in everything from language and vision, to robot control and playing games (TODO CITE). In 2012, researchers showed great improvements on the seminal ImageNet challenge using deep learning as a combination of extending old ideas and the availability of the necessary hardware (Krizhevsky et al., 2017). Since then, deep learning has been used to better address many challenges in artificial intelligence, while also being subject to critique on topics ranging from ethical and discriminatory contexts (TODO CITE), to robustness on critical tasks such as autonomous driving (TODO CITE). This debate goes back to the 1980s, where J. A. Fodor and Pylyshyn argue that *connectionism* cannot address the real challenges of artificial intelligence but rather is a step back-

Figure 1.1: Example of failures of large language models.

wards. We will expand on the historical context more in detail throughout this chapter. Since the 80s, the research landscape has gone through an AI winter and changed tremendously, but many of the critiques are still applicable today. While deep learning systems might achieve high performance on benchmarks, there are as many examples of how they fail spectacularly. For language models, this includes racial bias or catastrophic failures leading to nonsensical repeating (TODO CITE). There are entire research fields trying to dissect and understand these failures and general characteristics. Naively, we can summaries the issue as the difference between deep learning and deep understanding, and that many benchmarks test model behaviour rather than verify specific characteristics central to intelligence. This thesis will focus on the effects of vision on learning language, shortcomings of visual language models, how reasoning tasks can inform us in building better language systems, and what role compositional generalisation plays in addressing the shortcomings we see.

TODO show some contemporary examples of deep learning failing.

1.1 Historical problems – tentative title

This section will outline four historically important topics/themes/problems; Chinese Room, Symbol Grounding, ELIZA, and Fodor on concepts.

Many neuro-symbolic methods take a practical approach to the frame problem by working on predefined sets of symbols, thus scoping what is considered “interesting” to a particular system. The symbol set can be viewed as the frame. In recent years, there are many examples of large language models passing the Turing test. The test itself has been critiqued throughout the years, with more recent examples emphasising how LLMs can pass one instance while still performing catastrophically on a related task. ChatGPT is one such example.

ELIZA

We can attribute some of the success of ChatGPT to the underlying deep learning techniques and progress on processing large amounts of data. However, the human-centered influence on the design, i.e. building an interactive model with a low threshold, is arguably the key to its widespread success. In many ways, this design is not that different from ELIZA (Weizenbaum, 1966), that impressed researchers and the general public almost 60 years ago. ELIZA convinced some users of its intelligence, with people reporting long meaningful conversations. The following is an exchange between ELIZA and one such user:

Men are all alike

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

It's true. I am unhappy.

DO YOU THINK COMING HERE WILL HELP YOU NOT TO
BE UNHAPPY

I need some help, that much seems certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY

My mother takes care of me.

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

My father.

YOUR FATHER

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice
that.

WHAT MAKES YOU THINK THAT I AM NOT VERY AGGRES-
SIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR
FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT
YOUR BOYFRIEND MADE YOU COME HERE

This exchange also highlights that gender bias issues is not novel to large language models, where, e.g., ELIZA becomes gendered based on its user (Dillon, 2020). Weizenbaum built ELIZA on principles from Rogerian psychotherapy (C. R. Rogers & Carmichael, 1942), where one core idea is that the therapist reflects the word of the patient in a positive light. With simple rules for how to

reflect the input of the user, ELIZA managed to elicit attribution of human-like feelings from the users. Douglas Hofstadter defines this as the Eliza effect (Hofstadter, 1995) in *The Ineradicable Eliza Effect and Its Dangers*, a preface of *Fluid Concepts and Creative Analogies: Computer Models and the Fundamental Mechanisms of Thought*;

a host of implications follow in the minds of many if not most readers, such as these: [...] computers understand the physical world; computers make analogies; computers reason abstractly; computers make scientific discoveries; computers are insightful cohabiters of the world with us.

This type of illusion is generally known as the “Eliza effect”, which could be defined as the susceptibility of people to read far more understanding than is warranted into strings of symbols – especially words – strung together by computers. (p. 157)

Hofstadter goes on to say that the Eliza effect is a part of AI researchers hyping achievements. We seem to have returned to the ideas that ELIZA was built on, equipped with much stronger building material and tools. Given our own flawed ability to determine intelligence from interaction, we should be critical of impressive results with interactive AI systems in general. This position motivates the work on probing conducted throughout this thesis, as a way to look beyond behaviour and into the box.

Generalisation is not a necessity for weak AI, and there is a vast range of applications where strong AI is not needed. For instance, monitoring and controlling the components of an aircraft can benefit from AI tools, but does not require a generally intelligent agent to work. With a human-centered perspective, this is a key argument — we want AI tools to support humans, not replace them. However, strong AI with general intelligence could replace.

Chinese Room

In his paper *Minds, Brains, and Programs*, John Searle introduces the Chinese Room thought experiment (or *Gedankenexperiment* as he puts it) as an argument for why strong AI cannot be achieved without “just by designing programs but would have to duplicate the causal powers of the human brain” (Searle, 1980). Specifically, Searle take issue with the following *strong AI* claims; 1) “*that the machine can literally be said to understand the story and provide the answers to questions*”, and 2) “*that what the machine and its program do explains the human ability to understand the story and answer questions about it*”. We can restate the Chinese Room thought experiment as follows;

Suppose that a person is locked in a room with a large amount of Chinese writing, and no knowledge of the language, and a set of rules for how to produce Chinese symbols from Chinese text input. The person then uses the Chinese writing and rules to interact with a native Chinese speaker through a text-only interface. If the rules

 Men are all alike

 It is not accurate to make a sweeping generalization that all men are alike. Men, like women, come from diverse backgrounds, cultures, and upbringings, and possess unique personalities, beliefs, and behaviors.

While there may be some similarities among men due to socialization, genetics, and cultural expectations, it is important to recognize and respect the individuality of each person, regardless of their gender.

It is unfair and inaccurate to judge or stereotype an entire group of people based on the actions or behaviors of a few individuals. It is important to approach each person as an individual, with their own unique qualities and experiences.

 They're always bugging us about something or other.

 It is not fair to make a sweeping generalization that men are always bugging women about something or other. While it is true that some individuals, regardless of gender, may be more persistent in their communication or requests, it is important to address such behavior on a case-by-case basis.

Effective communication is key in any relationship, and it is important to communicate your needs and boundaries clearly to the other person. If you feel like someone is consistently bugging you about something, it is important to have an open and honest conversation with them about how their behavior is affecting you.

Additionally, it is important to recognize that men, like women, have their own needs and concerns. It is important to approach any relationship with respect and understanding for the other person's perspective. By fostering mutual respect and understanding, we can build stronger and healthier relationships, regardless of gender.

Figure 1.2: ChatGPT given the same prompt as ELIZA.

are exhaustive enough, nobody can tell from the answers alone that the person in the room don't speak a word of Chinese.

Searle goes on to point out that the person locked in the room "*[..] produce the answers by manipulating uninterpreted formal symbols.*" and is simply behaving like a computer where the rules are analogous to a computer program. The main claim is that the person does not *understand* Chinese, but only simulates that ability, and that understanding can only be achieved with *intentionality*. He later summarised this position as "*Syntax by itself is neither constitutive of nor sufficient for semantics.*" in Searle.

While the Chinese Room thought experiment illustrates a core issue in the strong vs. weak AI debate, it is also simplified. We will defer the *strong* vs. *weak* AI debate to some other time, there have been many replies to Searles' position taking issue with both the formulation and the premises. TODO Perhaps address which criticism this thesis subscribes to? One interesting critique is that while the non-Chinese speaker cannot assign any meaning to the sentences it manipulates, it inherits meaning from the Chinese speaker who is interacting with the world. This argument is similar to how people ascribe meaning to systems such as ELIZA, that meaning is in the eye of the beholder. Today, the default assumption of deep learning research is to assume that the person *does* know Chinese only from applying rules. In that context, it is important to point to the contemporary debate and critique of large language models with the Octopus Test introduced by Bender and Koller. However, and use the Chinese Room argument to illustrate why it is important to not only examine the behaviour of a system. In Chapter 3 of this thesis, probing is used in a way to pry open the proverbial door to the Chinese room and look inside. Some of the work investigates how strongly a symbol is grounded to an image, and the implications of this for multimodal methods. This leads us to our next topic, symbol grounding.

The Symbol Grounding

Following Searles' phrase "*Syntax by itself is neither constitutive of nor sufficient for semantics*", others also emphasise the impossibility of learning Chinese from dictionary alone. Stevan Harnad proposes that intelligent agents must solve the *symbol grounding problem* in order to give meaning to the language, by connecting it to physical systems and subjective experiences (Harnad, 1990). Harnad distinguishes between intrinsic and extrinsic meaning; the output from the Chinese room only carries extrinsic meaning imposed by the Chinese speaker. Importantly, he argues that meaning cannot come from symbols alone, but must be built on top of the intrinsic meaning of our nonsymbolic interpretations of our sensory input. In a way, Harnad extends on Fodor in saying that it is not enough that "*that the meaning of the symbols comes from connecting the symbol system to the world "in the right way".*". Instead, giving words intrinsic meaning is intertwined with the cognitive processing of sensory input. Motivated by this division into intrinsic and extrinsic meaning, Harnad proposes a

The semiotic triad

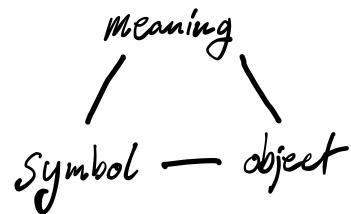


Figure 1.3: Illustration of the Peircean model of symbol-referent-thought

hybrid system that combines symbolic and non-symbolic elements. Moreover, he provides the following motivation for connectionism, in a description of what we now would call a neuro-symbolic system.

Connectionism is one natural candidate for the mechanism that learns the invariant features underlying categorical representations, thereby connecting names to the proximal projections of the distal objects they stand for. In this way connectionism can be seen as a complementary component in a hybrid nonsymbolic/symbolic model of the mind, rather than a rival to purely symbolic modeling. Such a hybrid model would not have an autonomous symbolic “module,” however; the symbolic functions would emerge as an intrinsically “dedicated” symbol system as a consequence of the bottom-up grounding of categories’ names in their sensory representations.

Harnad later describes the *frame problem* (McCarthy, 1960) in relation to the symbol grounding problem as the problem of connecting iconic and categorical representations to symbolic representations that are required by a given context. For instance, we do not name every concept that we possibly can while tasked with answering what color the cat is in an image. I.e., humans perceive and process all our sensory input but we only connect them to symbols when those symbols are useful to us.

Humans ground language to our experience of the world, meaning that through our interaction with it we attach meaning to language. The alignment of our internal model and our use of language means that while our world view

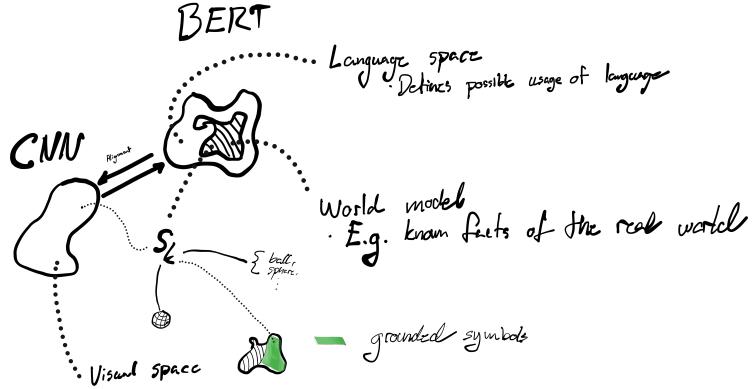


Figure 1.4: Tentative illustration of language model vs. world model.

is slightly influenced by our language, the concepts we learn are more universal.

Many of the issues with large language models are related to the fact that they are also world models, and that we have no direct way to observe or intervene on that internal world model. For instance, bias or errors in "understanding" are related to the model of the world. The process of externalising this internal model is grounding it in other modalities. Our language reflects our own internal world views, and similarly we could do multi-agent simulations in order to derive internal world models from agent interactions. This can be a continuation of Machine Theory of Mind (Rabinowitz et al., 2018) where agents are instantiated with different sets of beliefs and behaviours. The task is for agents to use interactions to construct a model of the other agents' internal parameters. An extension would be to do this but with interaction via language. Along these lines, Mahowald et al. distinguish between *formal* and *functional* use of language. Formal competence concerns *knowledge of linguistic rules and patterns* whereas functional competence concerns *understanding and using language in the world* (Mahowald et al., 2023). In their work, they show that large language models are good at the former, but fail the latter. One of their conclusions is that *[..] a model that excels at real-life language use would need to be an AGI, and argue that a human-like AGI cannot be reached simply by getting really good at predicting upcoming words*. This supports many of the points made here.

This thesis takes the position that bridging the gap between visual input and abstract concepts consists of three parts;

1. instantiation,
2. compositional learning, and
3. grounding.

Grounding is the process of establishing a mapping between visual element(s)

and abstract concept(s). A system can use this mapping in resolving which of the abstract concepts are seen in a specific image, thus instantiating e.g. a logical expression with visual variables. Compositional learning is the mechanism of building new concepts using previous knowledge as the building blocks. These three processes allow for generalisation beyond the initial domain defined.

We can take an example of the expression $p(x) \wedge q(x) \wedge z(y)$. In the previous definition, grounding translates into understanding that $z(y)$ maps to y being a cylinder. Instantiation could be mapping $p(x)$ to a specific object in a given image. Compositional learning can be learning the higher-order relation $R(x, y) \rightarrow p(x) \wedge q(x) \wedge z(y)$ so that the system can instantiate the relation R in images going forward. This incremental learning The difference to e.g. answer-set programming is that these concepts are not derived by the system to exhausting, even if there is also room for that within these three processes. Rather, compositional learning constitutes having a teacher naming the relation R , and the system connecting the previous knowledge in composing this new piece of knowledge, rather than learning it as an atomic concept.

TODO Fodor - What are concepts?

Fodor claims that *Connectionists are committed, willy-nilly, to all mental representations being primitive; hence their well-known problems with systematicity, productivity, and the like.* (J. A. Fodor, 1998). Instead of constituent structures of mental representations, the claim is that neural networks only have primitive, or atomic, representations of concepts. However, this is easily refuted by looking at works such as word2vec where adding or removing properties of a given concept results in another concept representing this change. While there is certainly merit to Fodors claim, the many subscribe to the idea of foundation models and that neural networks are capable of mental representations other than those primitive. It is important to clarify that this does not entail that neural networks get this for free, but rather that it is possible to train them to attain such representations. As one example, work by Lovering and Pavlick; Pavlick shows that neural networks indeed can exhibit the systematicity of compositionality. This is by no means clear without detailed inspection of the internal workings of neural network models.

- Representational theory of mind
- *It's a general truth that if you know what an X is, then you should also know what it is to have an X.*
- Unclear theory of what concepts are, could be argued for to not build systems around such flawed theories
- *Maybe having the concept X comes to something like 'being reliably able to recognize Xs and/or being reliably able to draw sound inferences about Xness'.* p. 3
- *[a] theory of meaning must answer 'What is it to understand a language?'*

- There is the Idea DOG. It is satisfied by all and only dogs, and it has associative-cum-causal relations to, for example, the Idea CAT.
- RTM tolerates thought without language
- Hume suggests mental representations are images
- *Connectionists are committed, willy-nilly, to all mental representations being primitive; hence their well-known problems with systematicity, productivity, and the like.*
- Mental representations have constituent (part/whole) structures

1.2 Difference between language model and world model

Many of the problems of large language models are rarely that they did not produce human-like language, but that the language did not match our own world model. For instance, telling you to saw a door in half instead of opening the door is not wrong language-wise, but might be socially awkward. There is also an increasing body of work on using language models as knowledge bases (Petroni et al., 2019a), erasing the separation. A mechanism for aligning world views, and adjusting the inner representation of a large language model to adhere can be a useful tool going forward. It is also important to distinguish between adhering to the world model, and generating human-like language that does not. For example, a langauge model could be prompted to converse about an alternate reality or generate a made-up story. In this scenario, some aspects might be coherent with our world models, but someone completely unthinkable (e.g. long-since dead) might be the president of a country. One alternative is the intergration with knowledge graphs, as this could provide a "fixedness" in separating out a world model from the language model. Knowledge graphs are also clearly compatible with reasoning frameworks.

1.3 Multimodality

Many aspects of intelligence need to work together over many modalities to achieve general intelligence. Deep learning needs to be complemented with deep reasoning and understanding. There are certain properties that can be identified, that can be controlled for. These properties guides us towards building better systems. Deep learning in combination with modules for, e.g., logical reasoning and different modalities can help us build systems with controlled for properties. Integrating these modules is difficult, and a joint semantics that bridges them can be key to utilising the strength of each module. It is important that all modules can act independently, so that the whole is greater than the sum, but the parts can still be used in a unimodal setting.

This section should give an overview of multimodal machine learning, outlining the different kinds of approaches to the problem. Also, characterise language-for-vision vs. vision-for-language vs. language-and-vision (Frank et al., 2021).

Our interaction with the world consists of multiple senses. We smell, touch, hear, and see things to make sense of the world. Each sense represents a different perspective of an observation or an event. These channels of information are known as different modalities.

Liang et al. gives an overview of the principles, challenges, and open questions of multimodal machine learning. (Uppal et al., 2022) gives a similar survey for language and vision. According to the authors, “we are now closer than ever to achieving intelligent agents that can integrate and learn from many sensory modalities”. They identify 6 key technical challenges; *Representation*, *Alignment*, *Reasoning*, *Generation*, *Transference*, and *Quantification*. This thesis covers topics of Representation, Alignment, partially Reasoning, and Quantification. The authors define multimodal as *./.. the computational study of heterogeneous and interconnected modalities*. These two core principles, heterogeneity and interconnectedness are further split up into several dimensions.

The dimensions of heterogeneity are *element representation*, *distribution*, *structure*, *information*, *noise*, and *relevance*. Information covers the fact that different modalities have different levels of information density under different circumstances. For instance, dark footage contains less information than bright, but certain objects scene might carry heavy weight in relation to the uninformative speech heard in a recording. This thesis is mainly concerned with element representation, structure, and information.

The second principle, interconnectedness, can be split into modality connections and modality interactions. The connections can be either statistical or semantic. The statistical connections can be of associative or dependent nature as, e.g., correlations found by deep learning methods, or temporal or causal dependencies. The semantic connections concern correspondance such as those of explicit grounding, or relationships between higher-level concepts such as hypernyms.

Multimodality has been a loosely defined term used to describe research on methods for heterogeneous data. In (Parcalabescu et al., 2021), the authors argue from this insight that an explicit definition is needed. They outline the difference between human-centric and machine-centric definitions previously used, and how they both have shortcomings. The authors instead consider multimodality defined relative the task itself, whether the model or the data representation is heterogeneous, and the complementary aspects of the data used.

- VALSE, task-independent benchmark for vision and language centered on linguistic phenomena (Parcalabescu et al., 2022)
- Neural natural language generation survey (Erdem et al., 2022)
- Visually grounded transformers, such as Vilbert (Lu et al., 2019).

- Learning visually grounded sentence representations (Kiela et al., 2018), perhaps move to ??
- Arguments for why imagination is a useful concept to model for multi-modality (Elliott & Kádár, 2017).

1.4 Neuro-symbolic AI

In the definition of the symbol grounding problem, Harnad described what we today would call neuro-symbolic systems. According to A. Garcez et al., neuro-symbolic AI aims to *combine two most fundamental cognitive abilities: the ability to learn from experience, and the ability to reason from what has been learned.* Many such systems aim to provide scalable learning *and* reasoning, where neural methods are better at the former and symbolic methods at the latter. Neuro-symbolic AI has a long tradition, surveyed by A. S. d. Garcez et al. over twenty years ago and Sun and Bookman describing the field emerging during the 1990s. The idea predates artificial intelligence as a field by decades, where McCulloch and Pitts describes a *logical calculus of the ideas immanent in nervous activity* and is often attributed with first describing neuro-symbolic systems. It is only recently that the field has gathered mainstream interest, where in the last few years publications in top conferences has grown steadily (Hamilton et al., 2022; Sarker et al., 2021).

Neuro-symbolic AI is motivated by cognitive science, and the observation that neural and symbolic methods have complementing strengths and weaknesses (Hitzler et al., 2022). In particular, since the widespread success of deep learning, neuro-symbolic methods are motivated by the failures and shortcomings of neural network-based approaches that have been made apparent in recent years. In cognitive science, with the most famous motivator being Daniel Kahneman and Amos Tversky's work on dual process theory later published as *Thinking Fast and Slow* by Daniel Kahneman (Kahneman, 2011).

(Sarker et al., 2021)
 (Besold et al., 2017)
 (A. Garcez et al., 2019)
 (A. d. Garcez et al., 2022)

TODO write section

1.4.1 Kuhnian perspective

A body of research can positioned in a broader context in many ways. Famous philosopher Thomas Kuhn introduced the term *paradigm shift* in this book *The Structure of Scientific Revolutions* (Kuhn & Hawkins, 1963). Kuhn studied the history of science and the progress of scientific knowledge, and Kuhn and Hawkins describes a paradigm shift as a scientific change through 5 phases; 1) pre-paradigm, 2) normal science, 3) crisis, 4) scientific revolution, and 5) post-revolution. A crisis arises when the reigning theories of a paradigm cannot explain oddities that seem to require considerable efforts outside of the current

scope of theories to address. This leads to a paradigm shift where the underlying assumptions of the old paradigm are questioned to produce a new paradigm. Once a new paradigm is established, the field can return to the practice of normal science. Kuhn argues that science alternates between normal science and revolutions through these phases.

- Phase 1 – It exists only once and is the pre-paradigm phase, in which there is no consensus on any particular theory. This phase is characterized by several incompatible and incomplete theories. Consequently, most scientific inquiry takes the form of lengthy books, as there is no common body of facts that may be taken for granted. If the actors in the pre-paradigm community eventually gravitate to one of these conceptual frameworks and ultimately to a widespread consensus on the appropriate choice of methods, terminology and on the kinds of experiment that are likely to contribute to increased insights.[13]
- Phase 2 – Normal science begins, in which puzzles are solved within the context of the dominant paradigm. As long as there is consensus within the discipline, normal science continues. Over time, progress in normal science may reveal anomalies, facts that are difficult to explain within the context of the existing paradigm.[14] While usually these anomalies are resolved, in some cases they may accumulate to the point where normal science becomes difficult and where weaknesses in the old paradigm are revealed.[15]
- Phase 3 – If the paradigm proves chronically unable to account for anomalies, the community enters a crisis period. Crises are often resolved within the context of normal science. However, after significant efforts of normal science within a paradigm fail, science may enter the next phase.[16]
- Phase 4 – Paradigm shift, or scientific revolution, is the phase in which the underlying assumptions of the field are reexamined and a new paradigm is established.[17]
- Phase 5 – Post-Revolution, the new paradigm’s dominance is established and so scientists return to normal science, solving puzzles within the new paradigm.[18]

What does this have to do with learning language with machines? I argue that AI research is possibly in the middle of phase 3, after the revolution of deep learning in the early 2010s. While the impact and amount of impressive results of current deep learning models (current paradigm) is undeniable, there is an every-growing body of work showing how such models fail considerably and might not be fixable (TODO CITE). As a result, there is a strong wave of research on neuro-symbolic methods (new paradigm) designed to address the faults of deep learning. From a scientific philosophy perspective, it is a great paradigm shift as it clearly combines and encompasses previous results while contributing to something greater than the sum of the two. Kuhn notes that it

is a good thing for science that a paradigm shift do not occur often or easily. Hence, time will still tell whether the neuro-symbolic paradigm is strong enough to reign.

1.5 Human-centered AI

One key question persistent through most of the discussions was about shared models between humans and machines. A shared model can be used to align the understanding of different parties, ground new concepts, and make, e.g., reasoning explicit within the shared model. This does not mean that the shared model, such as a knowledge graph, is the same thing as the internal model of, e.g., a machine. Much like humans, we only make things explicit in shared models when a situation requires us to do so, e.g., in communication of a decision. One key property of such a shared model is how to handle the open vs. closed world assumption (Reiter, 1981). There are multiple candidate methods which tries to fill such a gap, such as argumentation, narrative networks, and knowledge graphs. One question for this thesis could be to evaluate the properties we require of such shared models and how the existing frameworks cover those properties. One outcome could be to construct a hybrid/extension/new framework that is more comprehensive.

Key concepts in human-centric AI are communication and collaboration, transparency and explanation, control and adaptivity, and the social dimension. Shared semantic representations are central in communication and collaboration, as the common representation of intents, what concepts are taken into account, and what goals to achieve. Transparency is a natural consequence of these shared semantic representations, as they are structured in a way that gives insight into the inner workings of a model. However, just as with humans, it is possible to give one explicit motivation behind an answer or action while having a completely different internal (malicious) motivation. Shared representations give control as a human user can inspect and intervene on a machine representation such as a knowledge graph. It is also possible to quickly adapt to new knowledge, rather than requiring resource intense re-training. Shared semantic representations can also allow for personalisation, as e.g. the knowledge graphs for two users can differ according to experience or preferences, or contextual frame of reference.

Human learning is supervised in the sense that a teacher is guiding a student in the learning process. Supervised learning does not reflect this process, as the 'teacher' does not nudge the 'student' in the right direction or towards more interesting questions or pieces of knowledge. Kakas and Michael argues that learning should be guided by reasoning and explanations, and that argumentation can incrementally revise a world model through reduction or resolution of conflicting knowledge presented in a set of training examples. As a part of the foundational theory of HCAI, guided learning can be useful as its own definition and general mechanism. Guided learning is a synergistic process where the learner interacts with the teacher around its learned world model and knowledge

therewithin. The task of the teacher is to point out gaps that needs attention, or outright correct the student.

Human-centric AI (HCAI) is not about immitating how humans work, although it can be used as inspiration. Neural networks are famously inspired by, and etymologically related to the neurons in our brains. Neuro-symbolic methods are inspired by cognitive theories such as dichotomy presented by Kahneman, and can be used as mental models in designing new architectures (L. d. Raedt et al., 2020). Strong and simple mental models, such as the two previously mentioned, are powerful tools in laying the foundation for new research fields. However, simple models often hide the complexity of the problem. This leads to entire genres of models suffering from similar issues. For neural networks, this can be characterised by their opaqueness or sensitivity to learning data. It is important to consider the characteristics we expect from a HCAI system, so that we can carefully design new architectures that have a more holistic foundation to build on, rather than purely simple mental models.

Shared semantic representations facilitate interaction between humans and machines, allowing for communication grounded in a joint understanding of the world and collaboration around concepts that can be explicitly adapted over time. With symbolic semantic representations, powerful reasoning tools become available, giving systems capabilities that are otherwise hard to achieve. These properties are important in human-centric AI, as collaborative situations rely on a common understanding of premises, intentions, and goals. This makes natrual language processing in human-centric AI different from traditional NLP, where document summarisation is a completely valid task and valuable undertaking but does not have the same requirements that comes with human-computer interaction. Some of the key components of general intelligence, such as compositionality, also benefit from incorporating symbolic semantic representations.

1.5.1 Explainability

Explainability is needed both in evaluating systems, and as an affordance towards users. Neural networks are notoriously difficult to explain, whereas traditional methods like decision trees are much more straightforward. Combining neural networks with e.g. logic opens up for explainability by design, contrasted with the ad hoc explanations of, e.g., LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and DeepLIFT (Shrikumar et al., 2017). These methods are themselves vulnerable to adversarial attacks, as exemplified in (Slack et al., 2019). However, when humans explain decisions they (rarely) refer to brain activity and specific neurons but rather give justifications on higher levels. Most of the time, intuition rather than explicit reasoning is the underlying mechanism of our decisions and actions, as argued by, e.g., Mercier and Sperber (2009). This could support the post hoc explanations, but on the other hand, as McCarthy said, "*Artificial intelligence is not, by definition, simulation of human intelligence*" (Maker, 2006).

1.6 Compositionality

TODO Unclear if this is should be here

1.7 What Hill are we Climbing?

The field of artificial intelligence is sometimes referred to as a landscape with hills and valleys. In this analogy, valleys represent dead ends, and the highest hills, or peaks, represent something like our notion of artificial general intelligence. The goal then is to find the highest hill and climb it. One problem with climbing hills is that only moving upwards does not guarantee that you find the highest peak. Following this algorithm in my home town leaves me at 48 meters above sea level, hardly Mt. Everest. Therefore, it is important to know how to identify the landscape of high hills, and know when we are in a position to start climbing such a hill (G. Marcus & Davis, 2019).

Dileep George - Addressing what might be fundamental differences between current models and human-like intelligence

- Data-efficiency and causality
- Learned world models compatible with reasoning
- Grounding language with mental simulation
- Utilizing emergent insights from scaled-up models
- Utilizing insights from cognitive science and neuro-science

Current large language models are not compatible with reasoning frameworks, and expecting them to arrive to stable reasoning that can be explicitly explained is a long shot.

1.8 Identity of thesis

This thesis takes the position that

- Understanding language is a multimodal endeavour
- Concepts are compositional in nature, this should be reflected in methods for learning
- Neural networks alone are not robust reasoners
- Neuro-symbolic methods fulfill many criteria for more transparent and robust machine learning
- Discrete symbols alone are limiting the generalisability of systems

One problem with deep learning is the data and the metrics used, not the method itself. It is not clear whether neuro-symbolic approaches are clearly better than deep learning ones, or that general intelligence cannot be deep learning based. One important observation is that if we want our models to be capable of reasoning, this must be reflected in how we train them. A part of this means identifying the properties we want and that we can control for in datasets and with metrics. These can be used to construct learning situations that give meaningful models. That is where this thesis comes in.

1.9 Research questions

The aim of this thesis is to contribute to language technology that is capable of robust reasoning in multimodal contexts, while moving towards more trustworthy and transparent AI. These are (tentative) research questions that are addressed through the thesis:

- (R1) What are the contributions of introducing vision to language modeling?
- (R2) What are advantages and disadvantages of neuro-symbolic methods in multimodal language processing?
- (R3) Does utilizing hierarchical structures improve compositional generalisation in language grounding?
- (R4) What are the challenges and opportunities for language-centric learning on multimodal data, and what future research directions are there?

We use probing in Chapter 3 to answer (R1), and revisit the question in Chapter 4. Chapter 4 and ?? address (R2), with Chapter ?? mainly addressing (R3). Each chapter summarises the challenges pertaining to the topic covered, and we summarise the challenges and outlook in Chapter 7.

1.10 Contributions

The thesis contributes to our understanding of language modelling and the implications of learning with multimodal data. A common problem with language models is that specific properties, such as compositionality, are not explicitly tested for, but rather that focus lies on behaviour (e.g. appearing human).

1.11 Structure

- Chapter 1, Defining the (historical) context, scope, and position of thesis, how thesis relates to classical problems such as symbol grounding, Chinese room, et c.
- Chapter 2, Literature review of what is multimodality, modelling language in machine learning and generic characteristics as mini-ontology/review/taxonomy

- Chapter 3, Analysing internal structures by probing multimodal models
- Chapter 4, Analysing behaviour of multimodal models on reasoning and compositionality
- Chapter 5, Learning hierarchical representations of concepts through curriculum learning
- Chapter 6, Looking forward

More fleshed out - used for planning

1. Define the context and world view of thesis
2. Overview of the intersection of LLMs, multimodal machine learning, and grounding
3. Probing multimodal language models
 - Important to test both behaviour and internal structure
 - Probing as a way to understand the internal structures,
 - Metrics does not tell a complete story
 - COLING, EMNLP papers for probing
 - CLEVR-Math for reasoning aspects and shortcomings of NeSy in NLP
 - TFaS-HCAI paper for lack of HCAI perspective
4. Behaviour of multimodal language models/Grounding and compositionality
 - Concepts are floating point vectors, hard to reason over -> neuro-symbolic
 - Current LLMs not Gricean speakers (i.e. truthful, brief, informative)
 - Data splits to test behaviour, e.g. CLEVR-Math compositionality tests
 - NeSy models tie too strongly to inductive bias, DeepProbLog and CLEVR-Math experiments
 - Reasoning in multimodal data, CLEVR-Xs
 - LearnAut paper and continuation to give a more theoretical perspective
5. Characteristics
6. Learning hierarchical representations of concepts through curriculum learning
 - Neuro-symbolic

- Compositional structures
- CLEVR-Math continuation

Work to include

- COLING probing-paper
- EMNLP probing-paper
- CLEVR-Math paper
- Continuation of CLEVR-Math
- LearnAut paper
- Stable Confusion work with Hannah
- DeepProbLog experiments with confounding information
- Thinking Fast and Slow-paper x HCAI short paper

Chapter 2

Literature overview - tentative title

Example of how to use quotes at
the beginning of chapters

dali

TODO Not sure how to start this, or how to divide content between intro and this chapter.

2.1 Representations of Meaning

There are many approaches to model meaning, both implicitly and explicitly. Continuous representations, such as word or sentence embeddings, approximate meaning by closeness in a continuous vector space. Two words that are close indicate that they have some shared semantics. There are also representations of meaning based on the embedding of graph structures, encoding knowledge into a vector space with a similar semantic connotation. Discrete representations of meaning take many different forms. There are graph-based approaches, such as Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and more recently BabelNet Meaning Representation (BMR) (Navigli et al., 2022).

Discrete vs. continuous vs. probabilistic but discrete concepts? Easy with discrete for toy problems (e.g. CLEVR) to build vector representations where one bit represents the presence of a feature. However, to scale things, continuous representations have been key. Most of the time, features or properties are not discrete but rather on a scale. Even for toy problems this is the case, e.g. as the color blue is not one single color or not covering the entire object. Probabilistic based discrete representations with an intensity associated with each feature is more flexible in describing the real world.

Representing meaning is a multifaceted problem where not only do we need

to model the concept behind a word, e.g. the action connected to it, but also model the intentions of other agents and the state space in which we act. A model that only does the first will be inadequate.

2.2 Brief history of language modelling

Traditional methods, Word2vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), BERT, GPT-X

2.2.1 Foundation Models

In the comprehensive overview of neural network based models, or *foundation models* as the authors name them, (Bommasani et al., 2021) outlines many of the properties and challenges of foundation models. In this section, we summarise the ones related to models in multimodal and/or reasoning contexts. Finally, we argue why *foundation models* is a poor name and why we should focus on foundational properties instead.

Dupoux gives an overview on how results from cognitive science in infant language learners can be used to build better language systems. TODO elaborate on the properties we find from this

2.2.2 Mechanism for introducing new symbols

Many neuro-symbolic methods rely on a given set of symbols to manipulate, such as the logic program language of NS-VQA (Yi et al., 2018) that they learn the meaning of. However, generalising to learning new symbols is less emphasised. A generic mechanism for introducing new symbols via human-computer interaction would equip a wide range of methods with this capability. Thus, such a mechanism could bridge the gap between neuro-symbolic and neural approaches in their abilities to learn new concepts.

TODO read and cite (Bennett, 2022).

The Foucault effect; there is no meaning without discourse (Foucault, 1991). The discourse is a socially constructed limitation on what is sayable, what should be conserved, remembered, reactivated, or appropriated. Foucault argues that in the governing of social systems, we cannot talk about meaning detached from the discourse. A perfectly valid action in the 1800s can be totally unacceptable today. Similarly, language that carried great meaning before, might have lost that meaning today.

2.3 Needs major revision - Multimodal machine learning and language grounding

In transformer-based language models working only on text, words are grounded by the company they keep. Summarise arguments by Bisk (Bisk et al., 2020)

Visual Question Answering (VQA) and Visual Reasoning - Datasets

One of the first VQA datasets proposed was the DAQUAR dataset (Malinowski & Fritz, 2014) based on real images of indoor scenes. VQA is another widely used dataset (Antol et al., 2015) with images from MS-COCO dataset (T.-Y. Lin et al., 2014). Questions are manually created and answering these require commonsense knowledge and reasoning. The CLEVR dataset (Johnson et al., 2017) is based on automatically generated scenes and questions, giving great control over the distribution of instances. With CLEVR, one can decide to generate a training set with images having only a specific combination of objects (red cubes and blue cylinders), and a test set with a different combination of objects (red cylinders and blue cubes), as done in, e.g., CLEVR-Hans (Stammer et al., 2021). This control allows us to study various aspects like compositional generalisation of systems.

Closely related is the CLEVRER (Collision Events for Video Representation and Reasoning) dataset ([yi2019clevrer](#)) and CLEVR-Hyp dataset ([sampat2021clevr_hyp](#)). The questions on videos in CLEVRER requires reasoning about the state of objects after an video event, instead of after actions in text as in CLEVR-Math. CLEVR-Hyp focus on VQA where reasoning about effects of actions, and CLEVR-Math introduces an additional mathematical reasoning dimension to the problem. GQA is another relevant dataset, where real world images are annotated with rich scene graphs and a large set of relations and attributes, and focuses on compositionality in visual reasoning (Hudson & Manning, 2019a). Graph learning is a heavily studied area, with applications in multimodal domains such as robotics (J. Ji et al., 2020; Wald et al., 2020; Xia et al., 2021; Yu et al., 2021).

Experiments with Kandinsky patterns (Holzinger et al., 2019) show that neural networks are easily confounded by visual reasoning tasks with shapes, colors, and patterns that can be difficult to distinguish but follow clear rules. The Winoground dataset (Thrush et al., 2022) shows similar results, where no state-of-the-art visual reasoning method is able to distinguish between two confounding captions and images.

Existing Approaches to VQA

Most of the earlier approaches in VQA were based on purely neural models that first encoded the two inputs - the image and the accompanying question into embeddings using networks like Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) networks and then the two embeddings were forwarded to a classifier that would then predict the answer to the question ((Ben-Younes et al., 2017), (Fukui et al., 2016a)). Another category of approaches are the attention mechanism-based approaches that identified the regions in the image that were relevant to answering the associated question ((P. Wang et al., 2017), (Shih et al., 2016)). Graph neural networks (Narasimhan et al., 2018) have also been applied in VQA where both text and the image are represented as graphs and a multi-modal vectorial representation is learned that captures

the alignment of nodes in the two graphs (Radford et al., 2021) introduced the CLIP models where a representation of the image is learned with natural language supervision by leveraging the already available huge datasets for image captioning. More recently, neuro-symbolic approaches have been used in addressing the task of VQA like Neuro Symbolic Concept Learner (NSCL) (Mao et al., 2019a) and Neuro-Symbolic Visual Question Answering (NS-VQA) (Yi et al., 2018). These approaches convert the input image and text into an intermediate semantic representation and then employ a quasi-symbolic program executor to derive an answer from these semantic forms. We use CLIP and NS-VQA as baselines as they are state-of-the-art on multimodal language modelling and on the CLEVR dataset, respectively.

2.4 Neuro-symbolic machine learning

DeepProbLog (Manhaeve, Dumančić, et al., 2018), DeepStochLog (Winters et al., 2021), ConceptLearner (Mao et al., 2019b), RNNLogic (Qu et al., 2021b), Logic Tensor Networks (Badreddine et al., 2021), OWL2Vec (J. Chen et al., 2021), Neural Logic Machines (Dong et al., 2019), Extracting rules from Neural Networks (Duch et al., 1998), Gefner model free (Geffner, 2018), neural prediction systems (Goyal et al., 2021), neural state machine (Hudson & Manning, 2019b), predicate learning in ASP (Law et al., 2021), augmenting neural networks with FOL (T. Li & Srikumar, 2020), SATNet (P.-W. Wang et al., 2019a), Neuro-symbolic = neural + logical + prob (De Raedt et al., 2019), Target languages vs inductive biases for learning to act and plan (Geffner, 2021)

Visual grounding is addressed in neuro-symbolic machine learning in various ways. Inductive Logic Programming (ILP, (Muggleton & De Raedt, 1994)) is a form of symbolic learning and reasoning with first-order clauses for learning relations in data. Background knowledge is used together with examples to induce a hypothesis as a logic program, describing positive and negative examples in given data. ILP has neuro-symbolic approaches to machine vision problems (Dai et al., 2015; Varghese et al., 2021).

Probabilistic Logic Programming (PLP (Dantsin, 1992; Ng & Subrahmanian, 1992)), provides reasoning under uncertainty, with methods such as (De Raedt et al., 2007). (L. D. Raedt & Kersting, 2008) gives an overview of the combination of PLP and ILP, Probabilistic ILP. Neuro-symbolic PLP, such as DeepProbLog, can realise predicates as trainable neural networks, as a way to ground and reason about visual concepts (Manhaeve, Dumancic, Kimmig, De meester, & De Raedt, 2018; Weber et al., 2019; Winters et al., 2021). In the Neuro-Symbolic Concept Learner by (Mao et al., 2019b), programs for visual question answering are learned by combining modules for neural perception, semantic parsing of programs from language, and program execution. Other work builds on similar ideas (B. Zhang et al., 2021). The idea of Neural State Machines by (Hudson & Manning, 2019c) builds on probabilistic graphs to learn and reason about concepts for visual question answering. Logic Tensor Networks (Badreddine et al., 2022; Serafini & Garcez, 2016) are used for semantic

image interpretation as a partial grounding problem (Donadello et al., 2017). (A. Garcez et al., 2019) gives an overview of integrating learning and reasoning with logical representations.

Answer Set Programming (ASP) is another approach to knowledge representation and reasoning based on searching for stable models, answer sets, as solutions to given logic programs (Lifschitz, 2002; Marek & Truszczyński, 1999; Niemelä, 1999). Neuro-symbolic ASP has been used in reasoning about objects in video (Suchan et al., 2018).

Finally, grounding in language and vision with neural networks is extensively researched for tasks such as Visual Question Answering (VQA) (Antol et al., 2015; Chabot et al., 2018; Fukui et al., 2016b). Neuro-symbolic approaches to VQA include (Yi et al., 2018).

The above domains have different perspectives on learning and grounding, with overlaps such as between ILP and PLP (L. D. Raedt & Kersting, 2008).

Relationship between learning and reasoning

Kakas and Michael argues for a synergistic relationship between learning and reasoning (Kakas & Michael, 2020). Learning provides the elements/knowledge used in reasoning, while reasoning provides inferences that can be used as inductive bias when learning or extend the base of knowledge. One of their questions is; how do we exploit the reasoning process to enhance the learning process? Similarly, Luc Steels argues that AI suffers from a paradox called the “... hermeneutic circle: *To understand the whole we need to understand the parts but to understand the parts we need to understand the whole (Gadamer, 1975)*” and uses this as an argument against the linear progression of information in common data-drive AI methods (STEELS et al., 2022). Luc Steels et al. define the process of understanding as “*We frame the process of understanding in terms of a process of generating questions, reducing questions, and finding answers to questions.*” and propose to facilitate this process via narrative networks (STEELS et al., 2022). This method of measuring understanding could be extended to other semantic representations, such as argumentation, to check its validity and find commonalities.

MOVE EARLIER - Absoluteness of generalisation

Kakas and Michael (Kakas & Michael, 2020) point out that generalisation cannot be absolute, referring to the problem of induction (L. Henderson, 2020), and pointing to Humes’ *A Treatise of Human Nature* (Hume, 1739) with “... *inductive generalisation that is universal and absolute runs into logical difficulties as we cannot be sure that a future case will not contradict the generalization*”. This is used to motivate that argumentation is a flexible framework wherein such logical difficulties are possible to resolve naturally, as an argument only holds until evidence proves otherwise. One example the authors give is the difference between *all beans from this bag are white* and *all normal beans from this bag are white*, given that all beans drawn from a bag during training are white. The

second leaves room for the possibility that there are beans of other colors than white.

How does this relate to what we saw with CLEVR-Math and the experimental results showing that multihop generalisation was difficult? And similarly, how do you fit confounding information into this picture? When color is fixed for MNIST in training but not during testing, the generalisations during training that color is the main feature, or the shape, are both valid all else equal. How can you model both feature dimensions independently to be able to instruct a model to focus on shape, not color, in a particular test, independent on the training procedure?

2.4.1 Datasets for reasoning

- CLEVR (Johnson et al., 2017)/CLEVRER (Yi et al., 2020)
- CLUTRR (Sinha et al., 2019)
- GQA (Hudson & Manning, 2019a)
- FEVER/FEVEROUS (Thorne et al., 2018)
- Visual Genome (Krishna et al., 2017a)
- MSCOCO
- WebQA (Chang et al., 2021)
- VisualSem (Alberts et al., 2021)
- Winoground (Thrush et al., 2022)

2.4.2 Thinking Fast and Slow, and neuro-symbolic AI

Daniel Kahneman and Amos Tversky categorises the human mind into system 1 and system 2, responsible for fast and slow thinking respectively (Kahneman, 2011). Thinking Fast and Slow, and other dual-process theories, have had a big influence on the discourse in AI and development of new systems and methods. For example, the field of neuro-symbolic AI is heavily influenced by the mental model of fast and slow components, such as artificial neural networks in combination with logic programming. L. d. Raedt et al. claims that “Kahneman, 2011 has put the quest for neural symbolic computation A. d. Garcez et al., 2015; A. S. d. Garcez et al., 2012; Hammer and Hitzler, 2007 high on the research agenda” However, in many cases, these architectures do not reflect the systems described in many dual-process theories and would not fall under those definitions. Simply building a dual-component system is not sufficient in order to realise the full potential of cognitive dual-process systems. We argue that for Human-Centric AI (HCAI), the misalignment of dual-process theory and its instantiation in AI systems together form a foundation that is more than shaky. For example, neuro-symbolic methods with neural networks used as input to

symbolic components are still sensitive to the same type of biases that the AI community in general is tackling. Bias exists in humans' fast thinking, but is countered by slow thinking and adjusted over time accordingly. The relationship between neural and symbolic components cannot be unidirectional. Similarly, not all cognitive theories on this topic suggests two components, but three or a plethora of components. Neuro-symbolic AI would benefit from research in this direction, acknowledging that there is existing work on this TODO CITE.

It is important to remember that dual-process theory contain other works (e.g. J. Evans, 1996; Sloman, 1996; Stanovich and West, 2000, and that they also provide insights on their own. E.g. Slomans tripartite Reflective, Algorithmic, and Autonomous minds Stanovich, 2009.

A definition of fast and slow thinking given by J. S. B. T. Evans and Stanovich gives the following definition of fast and slow thinking;

Our preferred theoretical approach is one in which rapid autonomous processes (Type 1) are assumed to yield default responses unless intervened on by distinctive higher order reasoning processes (Type 2). What defines the difference is that Type 2 processing supports hypothetical thinking and load heavily on working memory.

However, neuro-symbolic methods rarely adhere to this definition.

- L. d. Raedt et al.: "Kahneman, 2011 has put the quest for neural symbolic computation A. d. Garcez et al., 2015; A. S. d. Garcez et al., 2012; Hammer and Hitzler, 2007 high on the research agenda"
- Opportunity for Human-Centric AI to build systems based on these ideas with a focus on
 - establishing a clearer computer science foundation based in dual-process theory and other cognitive theories,
 - building systems that better reflect our intelligence,
 - human-centric use cases where these capabilities are required.

Critics:

- A good story, but poor evidence. Many specialised subsystems that are not as easily categorised is more correct Keren and Schul, 2009.
- Dual-process theory fails because the features of fast and slow are not aligned Kruglanski and Gigerenzer, 2018
- Processing types are continuum, not clearly separable Osman, 2004

2.5 Knowledge graphs

Perhaps move this somewhere else initially and have a section here how they are used in learning and reasoning contexts.

2.5.1 Metrics in learning graphs

These notes can be used to support the need for the properties framework.

There is work (Berrendorf et al., 2020; Kadlec et al., 2017; Pezeshkpour et al., 2020; Rossi & Matinata, 2020; Y. Wang et al., 2019) criticising how knowledge base completion is evaluated. In (Rossi & Matinata, 2020), the authors make the observation that less than 15% of entities cover more than 80% of the facts in many of the datasets. By only predicting facts of this small set of entities, a model can achieve good performance. Thus, a model can memorise the explicit facts of a certain entities without learning anything about the general relationships they describe. There is a clear parallel to how large language models work, where the training does not hold out information such that the testing actually tests for the generalisability of the model. The contributions of (Rossi & Matinata, 2020) are the definition of a set of properties useful in capturing the relations rather than the small set of entities.

They use the inverse relationship train-test leakage examples from FB15K and WN18, showing that father-of and child-of are inverse relations. I.e., learning that one is the inverse of the other means that you can explicitly learn one fact in training and reproduce the inverse in testing.

The authors argue that global metrics such as mean rank et c. does not show strengths and weaknesses of different models making it difficult to compare them. Entity frequency will have a too big influence. They also point to other work that highlight that the metrics only measure positive test facts, but not that false or nonsensical facts get low scores.

Based on these results we can ask ourselves "*link prediction for what purpose?*". Is it to better model the world with graph completion? Unclear how the metric reflects on the performance in an application. Usage for a recommender system gives a different set of desirable properties than if used in a database setting, or medical applications. The application matters, hence the expected properties should be understood and possible to control for.

2.6 Compositionality

Compositionality is an important component of generalised intelligence, the alternative being infinite memorisation. Compositionality is one of the reasons why distributional semantic representations such as `word2vec` saw such success. Szabó gives the following definition of the *principle of compositionality*:

The meaning of a complex expression is determined by its structure and the meanings of its constituents.

adapted from

The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined.

As a simple example, the sentence “*A blue sphere.*” refers to multiple concepts (color, shape) that we use to understand the sentence. Without a compositional understanding of these concepts, we might consider *blue sphere* as a single concept, not recognising the likeness between red spheres and blue spheres. However, the above definition vaguely defines meaning. For instance, a blue sphere might refer to Planet Earth in some contexts, and a billiard ball in others. We can extend this definition to

For every complex expression e in \mathbf{L} and every context c , the occasion meaning of e in \mathbf{L} at c is determined by the structure of e in \mathbf{L} and the occasion meanings of the constituents of e in \mathbf{L} at c .

for a given language \mathbf{L} . G. F. Marcus gives the following requirements for compositionality:

- Stable encodings of individual elements
- An operation that concatenates pieces of trees together
 - or disassembles wholes into parts
- Iterative process for (de)constructing larger structures
- Representational formats for trees (or something very similar).

There is work on neural networks to achieve this, such as G. Hinton. However, neural networks have long been criticised for lacking such mechanisms for compositionality J. A. Fodor and Pylyshyn, 1988. Hupkes et al. decomposes testing of compositionality into *systematicity*, *productivity*, *substitutivity*, *localism*, and *overgeneralisation*. One of the strengths behind human cognition is often attributed to its capabilities for compositional generalisation (Werning et al., 2012). Similar to our artificial neural effigies, our ability for continuous computing that allows for arbitrarily small changes to concepts used in reasoning. This is known as the continuity principle and underlies the advancement of deep learning (G. E. Hinton et al., 1986). Smolensky denotes the trade-off between the continuity and the compositionality principles as the *Central Paradox of Cognition*. In their work on neurocomputational computing, Smolensky et al. attributes the achievements of convolutional neural networks (CNNs) for vision and transformers for language to their compositionally structured processing of input (J. Henderson, 2020). For CNNs, the convolutional layers imposes spatial structures on the computing. For transformers, it is possible to show that they are equivalent to graphs with weighted links between symbols, thus imposing a compositional structure of concepts (Dwivedi & Bresson, 2020). Smolensky et al. argues that architectures that respect both the continuity and compositionality principles will address some of the issues with current AI systems such as their lack of reasoning capabilities. In this, the authors define Tensor-Product Representations (TPRs) and Neurally-Encoded Compositionally-Structured Tensor (NECST) computing as a theoretical framework to frame this. Recent approaches, such as the NECSTtransformer, build on these concepts (Schlag et al., 2019).

Measuring compositionality (Andreas et al., 2019; Chaabouni et al., 2020; S. Xie et al., 2022). IDEA: Check all major works (CLIP, Stable Diffusion, DALL-E 2 et c.) for keywords related to compositionality (e.g. disentangled et c.). In discrete representations, compositionality is a property that falls out by definition, but in continuous representations this is much more opaque.

The binding problem in artificial neural networks (Greff et al., 2020)

N. Kim and Linzen introduce COGS as a compositional generalisation challenge based on semantic interpretation. COGS is rule-based such that there are systemic gaps that can only be filled if a model has learnt concepts compositionally. For each sentence, there is a corresponding logical form giving complete control over how the data is generated. The test set then contains novel combinations of concepts that were not seen during training. For example, the training data can contain the following two sentences: *The cat loves the girl*, and *The hedgehog sees the cat*. To test whether a model can use these concepts to compose and understand novel sentences, we can give it the sentence *The boy loves the hedgehog*.

Work by IBM on compositional generalisation and neuro-symbolic methods (Basu et al., 2021; Ito et al., 2022; Klinger et al., 2020; Riegel et al., 2020; Sen et al., 2022). J. D. Fodor et al. critiques the “unreality” of semantic representations.

2.6.1 Benchmarks for Compositional Generalisation

TODO Sort out division of material between this and previous section

Lexical generalisation means using a known word in a new context, whereas structural generalisation means creating a new combination of familiar structures. When looking at dependency trees, lexical generalisation means exchanging the word at a leaf node with a word of equivalent word class. In structural generalisation, an entire sub-tree is exchanged with another subtree previously seen. What this dichotomy fails to address is when a completely new concept is learnt by building a compositional understanding of it using previous knowledge. We call this compositional grounding. N. Kim and Linzen show that neural networks can achieve great performance on lexical generalisation, i.e. using known words in new contexts, but that they fail completely on structural generalisation. Weißenhorn et al. and Qiu et al. both show that neural models that are made aware of structure can do structural generalisation. Qiu et al. identify that transformer models can be augmented with synthetic data that is generated from structured methods, in their case quasi-context free grammars. Weißenhorn et al. uses neural network components for dependency parsing and constructing a graph representation, hence building highly structured representations of sentences. It might be possible to see all useful structures given enough data, but humans clearly perform structural generalisation with far less data (Linzen, 2020).

Another important concept that can help accelerate a model’s ability to generalise compositionally is curriculum learning. By inducing a compositional bias early on, the rest of a training process can adhere to that structure much like

how human knowledge is built compositionally by learning the building blocks first as in addition before multiplication.

B. M. Lake introduces a test for systematic generalisation, defined as an algebraic capacity to combine novel sentences from known components. As an example, learning what *jump* means, a human can be asked to *jump twice* without being told what the combination entails given that *twice* is known. This is something neural networks have been shown to fail at, especially for recurrent neural networks (B. Lake & Baroni, 2018b). The dataset introduced by B. M. Lake uses pseudowords, such as *bicket* and abstract outputs in the form of colored dots, to test this property without dependency on a specific language. This could be extended to the CLEVR domain, by introducing pseudowords for abstract concepts such as three blue cubes being called a *blargh*.

In developmental psychology, assigning new meaning to a new word, rather than as a referent to something previously known, is called mutual exclusivity. It makes up the three components of lexical learning, together with the *taxonomy assumption* and *fast mapping* (Golinkoff et al., 1992). It is important to distinguish between mutual exclusivity and learning a word in a different language. For instance, consider a new word *blargh*. It could be a synonym for sphere, in which learning *blargh* is similar to learning a new language (i.e. anchoring known concepts in a new language space). However, if *blargh* means *blue sphere*, or *3 blue spheres*, then it is a compositional learning task. Building a new concept out of previously known ones like this is a mechanism that we study in the following chapters.

Compositional bias can be a structural property of an architecture, or an emergent property of the training data or procedure. This is one example of an important innate bias that exists in humans, allowing us to learn under vastly different circumstances.

‘Syntax is an algebra, semantics is an algebra, and meaning is a homomorphism between them’ (Janssen, 1986).

2.6.2 MOVE CONTENT - Winograd schema challenges

Levesque et al. (2012) introduced the Winograd schema challenge as an alternative to the Turing test.

The trophy doesn’t fit in the brown suitcase because it’s too big.
What is too big? A) the trophy, or B) the suitcase?

Kocijan et al. (2020) outline problems with the assumptions in the original Winograd schema challenge, and some of the approaches used to solve the challenge. They point out that neural language models such as BERT (Kenton & Toutanova, 2019) can solve the challenge, but still fail on tasks related to common sense reasoning. In the multimodal domain, Thrush et al. (2022) proposes a Winograd schema for visio-linguistic reasoning with images and text. Shown in Figure 2.1, the task consists of pairs of images and corresponding descriptions, where the descriptions are flipped versions of each other. Thus, they contain the same symbols, but represent different meanings. They show that a range of

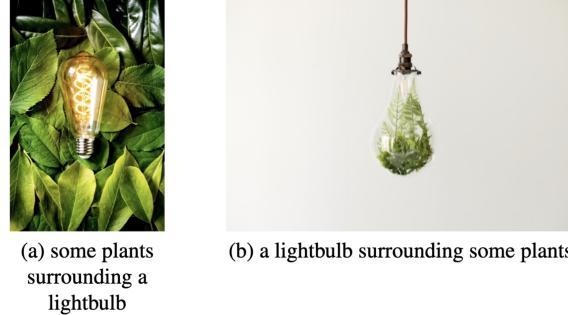


Figure 2.1: Example taken from (Thrush et al., 2022)

models in the state of the art of multimodal machine learning fails significantly on this task.

Winogradeqe challenges is an effort in moving away from evaluating models by their performance on large datasets where the task is to, e.g., predict words, and rather focus on a smaller set of more thought through examples targeting specific properties or behaviours. But as has been shown for the Winograd schema challenge, new methods can be devised to solve such tasks while still exhibiting flaws which the task was supposed to cover. It is worth discussing when the pattern of *devise new test → ML fails on test → new method solves it* is a pattern of self-fulfilling prophecies.

2.7 REMOVE AND KEEP PARTS OF CONTENT - Properties for compositionality and grounding

It is necessary to define what *learning language* means, and what *grounding constitutes*. In this section, we outline a couple of cases of grounding new language. Characterising these cases will help us understand where and how language models fail and how we can train them differently to our benefit. Many of the problems with current language models stem from confusing training goals with how we want to use them. While the community at large is debating over neural, symbolic, or neuro-symbolic methods, this thesis takes a neuro-symbolic stance. The thesis device methods for compositional grounding and outlines the dimensions for capturing language modelling.

In this endeavour, it is important to clarify why we want to model language, in what applications and scenarios we see them used. This will guide us in what design choices are necessary to achieve our goals, rather than building as generalist a model as possible. It is important to distinguish between learning language, and learning from language. Some methods do both, like transformers, whereas

some methods only learn from language, like knowledge base construction.

The hypothesis of this thesis is that an external continuously compositional representation of concepts and world state.

Explanations are only needed once there is a misalignment between my internal model/theory of mind of the other agent and its behaviour.

Where do we draw the line between training and inference?

There is a dichotomy between algebraic and inductive bias approaches, does it have to be?

Compositionality in syntax versus in semantics. Current large language models are good at producing output that shows a compositional understanding of the surface form (i.e., how can we formulate a correct sentence). However, they lack in semantic compositionality and there are few measures of such.

Similarly to the definition of a Gricean speaker by Merrill et al., we can define characteristics such as systematicity, productivity, and compositionality. Pavlick investigates how deep learning models exhibit semantic structures. In some of their work, the authors show how different properties of concepts are learnt at different stages throughout a network (Lovering & Pavlick, 2022). The authors show that concepts are learnt compositionally such that structural concepts with shapes and formations are not affected by the color seen in training and testing. This work aims at answering to which degree neural networks employ systematicity in constructing their understanding of its input data, also exemplified in (R. Patel & Pavlick, 2021).

One key point that Pavlick does, is that representations should be continuous and that compositionality does not imply discreteness. A discrete set of symbols does not imply compositionality if a new symbol is grounded for each new combination of properties without relating them back to already known components. Similiarly, there is good evidence that continuous representations, such as those of neural networks, exhibit a compositional structure to some degree. This implies that it is definitely possible to learn concepts compositionally with neural networks, but it is still unclear how effective this process is. This also implies that failing on a compositionality test does not imply that the underlying system is not compositional.

J. A. Fodor and Pylyshyn defines compositionality as "*The ability to produce/understand some sentence is intrinsically connected to the ability to produce/understand certain others... [they] must be made of the same parts.* In contrast, Kamp and Partee gives a more lenient definition of *the meaning of a sentence is a function of the meaning of its words and the way in which they are combined*. The later definition does not say anything about how understanding parts translates between inputs, but only that meaning is derived from the words themselves in combination with their ordering. However, for a compositional system, understanding that constituent parts can be reused across different inputs is a key property of building an efficient system. The hypothesis is that this is a criteria for achieving compositional generalisation.

2.8 Challenges and characteristics

What is the purpose of the representation

- Transparency
- Allow explicit reasoning
- Shared representation in communication between humans and machines
- Allow for representing different world views/states.

What are the properties of the representation?

- Continuous concepts should be possible to represent
- Structured representation to allow for explicit reasoning
- Possible to construct new concepts compositionally from previously known
- Possible to represent 'possible world states'
- Possible to represent multiple modalities
- As little inductive bias as possible
- Possible to do multiagent grounding (mapping between 'world views')

2.8.1 Challenges in neuro-symbolic reasoning

2.8.2 Challenges in Compositionality

- Difficult to construct difficult datasets

2.8.3 Challenges in Multimodal Reasoning

Summarise challenges.

- Generalise outside given program structures
- Scalability
- How to introduce and ground new symbols, e.g. with clevr: Introduce arbitrary relations/concepts such as four red objects is called a gallub.

Chapter 3

Probing multimodal language models

Example of how to use quotes at the beginning of chapters

dali

One major challenge with neural network-based methods is to understand their inner workings, which we commonly refer to as “looking into the black box”. Even if we have constructed a benchmark where good performance should tell us how a model behaves, it is difficult to know if the way a system produces an output follows a procedure that is actually desirable. In Chapter ??, we talked about this as an implication by the Chinese Room (Searle, 1980) thought experiment, where a computer system applies rules over Chinese characters to pass the Turing test. Searle questions whether this system *understands* Chinese, stipulating that it merely simulates it by hijacking meaning as imposed by the user. This perfectly illustrates a big issue with machine learning in general, and deep learning specifically, in that it is difficult to draw strong conclusions about the capabilities of a model strictly from observing its behaviour. With this in mind, we now turn to *probing* as a method to open up these black boxes.

Probing is one approach to reveal what information an embedding actually encodes (Conneau et al., 2018; Hupkes et al., 2020a; A. Rogers et al., 2018; Yaghoobzadeh et al., 2019). An embedding is a vector representation of anything from a word to an entire image that we obtain by feeding input to a neural network and extracting the activations from a layer in the network. For some methods, like word2vec, the embedding is the final output that is used on its own in a downstream task. Conversely, we can extract an embedding from any layer of a neural network to observe what information is encoded at certain stages throughout the network. In general, we say that such embeddings capture some semantic information distilled from the syntactical information present in the input. Two well-known semantic embeddings are word2vec (Mikolov et al., 2013a)

and BERT (Devlin, Chang, et al., 2019). In (Mikolov et al., 2013a) the authors demonstrate the additive compositionality characteristic of word embeddings in word2vec. In particular, if words or phrases occur frequently together in the same context, the sum of the vector representations of two single words is close to the vector representation of a phrase that combines those single words. For example, the sum of the vector representations $\text{vec}(\text{Russia}) + \text{vec}(\text{river})$ is close to the vector representation $\text{vec}(\text{Volga river})$ (see (Mikolov et al., 2013a)). BERT (Bidirectional Encoder Representations from Transformers) is introduced in (Devlin, Chang, et al., 2019) and is a bidirectional language model. Unlike word2vec, BERT considers the left and right context surrounding a word and pre-trains deep bidirectional language representations in an unsupervised way. BERT can then be fine-tuned to another task as an instance of transfer learning.

In this chapter we look at two sets of probing experiments. Section 3.2 shows probing for visual information in multimodal embeddings, giving us insight into how visual information is encoded in multimodal embeddings. For Section 3.3, we test how well language embeddings capture information about semantic relations such as synonyms and hyponyms.

3.1 What can probing tell us?

Conneau et al. define a (linguistic) probing task to be a classification task that categorizes sentences according to specific linguistic properties, such as sentence length. Commonly, probing tasks are agnostic to the specifics of encoder architectures and can therefore be used to compare across different methods. In (Conneau et al., 2018) linguistic probing tasks are systematized building on work in (X. Shi et al., 2016) and (Adi et al., 2016). In particular, Conneau et al. define a probing task to be a classification task that categorizes sentences according to specific linguistic properties. For example, given an encoder such as an LSTM, pre-trained on some downstream task, the sentence embeddings that it produces can be used to train a grammatical classifier for the probing task that asks to determine the number (singular or plural) of the subject of the main verb. If the sentence embedding captures that information in an accessible way, this may be revealed by the classifier performing well, whereas the classifier will certainly not be able to perform well if the information is not appropriately captured. Thus, the main idea of probing tasks is to draw conclusions from the classifier performance about the probed embedding; if the classifier succeeds it means that the semantic embedding captures interpretable information regarding the aspect under consideration. Hewitt and Liang argue that the performance of a probe alone is not sufficient, and introduce so-called *control tasks* to improve interpretability of probing tasks. A control task reveals whether high accuracy of a probing task really indicates that semantic representations encode a linguistic property, or whether the probing task itself learns this property. In particular, a probing task is complemented with a control task that associates random outputs to the properties under consideration (for example, POS tags).

Type of information probed for:		
Surface	Syntactic	Semantic
Sentence length	Bigram shift	Tense
Word content	Tree depth	Number of subjects
	Top constituent	Number of objects
		Semantic incongruence
		Coordination inversion

Table 3.1: Probing tasks for semantic embeddings, organized along three broader probing categories as investigated in Conneau et al.

Thus, a control task with low accuracy indicates that a corresponding probing task with high accuracy does indeed encode the probed property. Simply put, a probing task is a classification task in which a network is trained on a given embedding. The task is chosen so that the performance of the trained classifier provides insight into the nature of the information captured by the embedding. Finally, ignorance about what is actually captured in an automatically learned semantic representation may lead to serious consequences of various kinds such as propagating discrimination bias (Bolukbasi et al., 2016; Brunet et al., 2019; Caliskan et al., 2017), or causing safety hazards in robotics by inducing unexpected robotic actions that put humans at risk (Orseau & Armstrong, 2016; Wachter et al., 2017).

The probing tasks proposed by Conneau et al. probe sentence embeddings and are categorized according to the type of linguistic properties they capture: *surface,- syntactic,- and semantic information*. We now give a brief account of these categories, which are outlined in Table 3.1. Surface information comprises probing for sentence length (number of words) and the word content (whether the sentence contains a given word). The probing tasks in the syntactic category ask to detect bigram shift, tree depth and top constituent, revealing whether the embedding makes certain syntactic information accessible. Bigram shift tries to predict whether two adjacent words have been swapped (that is, encoding the syntactic order of words). Tree depth asks to determine the depth of the syntactic tree of the sentence, and the top constituent task asks the classifier to determine the sequence of the top constituents directly below the sentence (S) node. The probing tasks that probe for semantic properties are tense, subject and object number, Semantic Odd Man Out, and coordination inversion. The tense task consists in finding the tense of the main verb, whereas the subject and object number tasks ask to predict the grammatical number of subjects and objects of the main verb, respectively. The task Semantic Odd Man Out is about predicting whether a sentence has been modified or not (i.e., a random noun or verb was replaced with another noun or verb). Coordination inversion probes for the information whether two coordinate clauses in a sentence have been switched. For example, “They might be only memories, but I can feel each one” and “I can still feel each one, but they might be only memories” (Conneau

et al., 2018).

These probing tasks were defined for unimodal embeddings of natural language. Machine learning that utilize multimodal embeddings is a lively field (Adi et al., 2017; Felix et al., 2018; Socher et al., 2013), but little is known about what properties these multimodal embeddings actually capture. Work, such as by H. Wu et al., aiming to analyze embeddings according to the composition of their encoded concepts is rare.

The establishment of probing tasks is one way to gain systematic knowledge about what embeddings actually capture. Another complementary way is to build taxonomies of multimodal machine learning techniques and multimodal embeddings. Such taxonomies are proposed, for example, by Baltrusaitis et al. and Beinborn et al. Both groups of authors categorize embeddings according to different but partly overlapping criteria. The taxonomy by Baltrusaitis et al. classifies approaches according to five categories of criteria: (a) *representation* – how complementary and redundant information is represented, (b) *translation* – how data is mapped between modalities, (c) *alignment* – whether and how elements in the different modalities are aligned, (d) *fusion* – how information coming from different modalities is integrated, and (e) *co-learning* – in which ways the learning exploits multimodality.

The taxonomy by Beinborn et al. for (learning) multimodal representations distinguishes between (f) *concept representations* – embeddings that use low-level representations of concepts, (g) *projections* – embeddings that represent concepts using only one of the modalities, and (h) *compositional representations* – approaches that fuse or jointly embed the different modalities.

Multimodal probing tasks can support the location of a given method in a taxonomy without requiring intimate knowledge of its inner workings: probing which information is accessible by a network trained on the resulting embeddings provides insight into what information is present and how it is represented. Some major difficulties of multimodal processing tasks and representations are discussed (from the perspective of multimodal grounding) by Beinborn et al. Their discussion illustrates the usefulness of multimodal probing in general, and of visual-semantic probing in particular:¹

Combining complementary pieces of information Different modalities contribute to the information content of multimodal input in complementary ways. For example, highly relevant visual properties, like the fact that birds have wings and violins are brown, are not usually mentioned in text as they are the default. Conversely, taxonomic and functional relations between concepts are poorly represented in images. Probing tasks that check whether, e.g., the word *brown* relates to images of violins would allow to draw conclusions about how successfully these dimensions are combined in the embedding.

Representation of abstract concepts Multimodal grounding of verbs is difficult in comparison to grounding nouns and adjectives. This should not come as a surprise because verbs denote more abstract concepts than many nouns and

¹We extract two aspects from the four challenges discussed by Beinborn et al., basically combining challenges 2–4, as our focus is not on grounding.

adjectives do. Abstract concepts like *together*, *theory*, and states of mind give rise to similar difficulties. Probing tasks that evaluate how well such concepts are represented in multimodal embeddings would thus be highly useful.

Combining complementary information: Different modalities contribute different qualitative data. For example, highly relevant visual properties (e.g. birds have wings, violins are brown) are not represented in text, whereas taxonomic and functional relations between concepts are poorly represented in images. An open question is to which extent image information contributes when combined with text (most research investigating along these lines, focus on nouns and adjectives only). The authors argue for multimodal approaches that go beyond concept similarity.

Multimodal grounding of verbs: research investigating multimodal aspects for verbs are extremely rare. Existing work suggest that the performance for verbs is significantly worse (compared to nouns, adjective). The authors in (Beinborn et al., 2018) compare different representations of combining visual and textual verb pairs.

Imageability of abstract words: Concrete words can be visualized and visually represented much better than abstract words (e.g. together, theory). This also holds for verbs with a high degree of embodiment (e.g. fall, dive) compared to verbs with a lower degree of embodiment (e.g. know, decide).

Selective multimodal grounding: we lack an understanding of how to combine concept representations and this is a difficult tasks since, for instance, image collections are much more diverse for concrete concepts (e.g. ladder, car) than for abstract concepts (e.g. happiness, intention). Thus, approaches that perform selective multimodal grounding constitute a more plausible approach to sentence processing according to the authors. An open question is how to visually represent coordinating expressions (e.g. but, nor, and, or, so, yet). Casper et al. provides a comprehensive overview of probing methods.

TODO Fit somewhere else In *multimodal* semantic analysis, the syntactic domain is a Cartesian product of two or more domains, such as an image with a caption. The syntactic domain is sparse, where words are represented as one-hot vectors and images as their pixel values. The dense embedding, on the other hand, captures semantic information that a complex model can interpret. They allow for semantic analysis beyond what is possible with more syntactic or handcrafted representations. One example of a simple example is how it is possible to define distance between embeddings that carry semantic information. There are many models based on machine learning techniques that jointly process the input modalities (Shen et al., 2019; H. Wu et al., 2019). Multimodal learning models such as DeViSE (Frome et al., 2013) demonstrate in particular that *zero-shot* learning can be significantly improved by engaging multiple modalities.

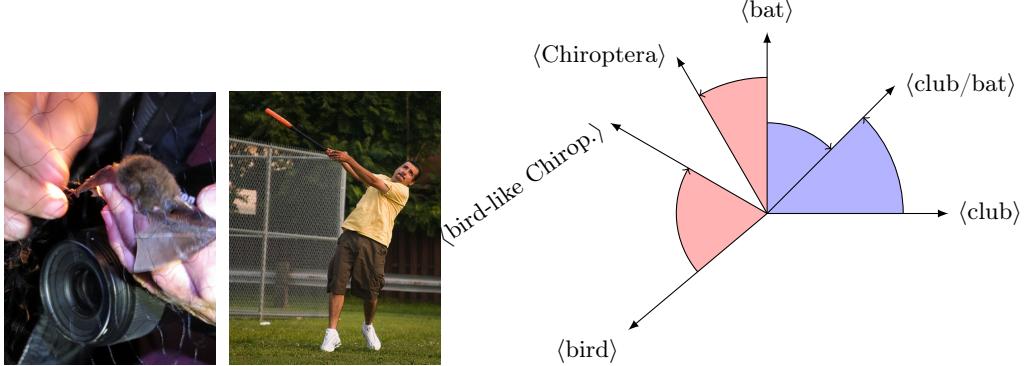
3.2 Probing Multimodal Embeddings for Linguistic Properties

Semantic analysis aims to infer meaning from data, relating objects in a syntactic domain to objects in a semantic domain. In natural language processing, semantic embeddings from methods such as word2vec, BERT, and GPT-3, revolutionised semantic analysis of text. The embeddings map words to real-valued vectors which reveal semantic aspects, for example, if words are related in meaning or belong to the same topic. Creating such an embedding means to enrich as well as filter out information. Unavoidably, some (usually surface and syntactic) information will be lost in the process of projecting words and their contexts onto a representation that focuses on meaning. For example, when the word ‘bat’ is seen in the context “the bat hits the baseball”, its embedding vector is quite different from what it would have been if the context was “a small bat captured by a zoologist”. Hence, there is a trade-off: the better we capture the semantics, the more surface and syntactic information becomes blurred. It depends on the downstream task what is the right balance between abstraction and detail. What complicates matters is that embeddings are automatically learned rather than crafted by hand, and thus it is not clear precisely which aspects such an embedding actually represents.

The combination of modalities that has hitherto received the greatest interest is the pairing of images and text, with their embeddings commonly called *visual-semantic embeddings*. Throughout the rest of this paper, we focus on visual-semantic embeddings, and base the empirical part of our work on the dataset *Common Objects in Context* (MS-COCO), which consists of images with captions (T. Lin et al., 2014).

When semantic analysis is applied to the text component of an image-caption pair, the visual information can resolve semantic uncertainties such as in the phrase “a man with a bat in his hands”. Figure 3.1 shows two MS-COCO images,² each image accompanied by two of its associated captions in the dataset. While humans would probably glean the correct interpretation of ‘bat’ and ‘club’ from the text alone (but not of ‘bird’), the visual-semantic information is much less ambiguous. The vector diagram to the right of the images illustrates how one might imagine the vectors of a word embedding such as word2vec to be affected by moving to the multimodal embedding, including visual information. Imagine the vectors $\langle \text{bat} \rangle$, $\langle \text{bird} \rangle$, and $\langle \text{club} \rangle$ to be those of the pure word embedding. That is, for simplicity we assume that the words are embedded as in the original word2vec embedding, without taking the context provided by the sentence into account. In particular, the two occurrences of ‘bat’ in the captions are represented by one and the same vector $\langle \text{bat} \rangle$, and similarly for the two occurrences of ‘club’. Incorporating the information in the corresponding

²The authors gratefully acknowledge the MS-COCO dataset (T. Lin et al., 2014) as the source of the two photographs, licensed under <https://creativecommons.org/licenses/by-nc-nd/2.0/> CC BY-NC-ND 2.0 and <https://creativecommons.org/licenses/by-nc/2.0/> CC BY-NC 2.0, respectively.



A tiny bat is held by someone with a camera.
A man in shorts is swinging a bat.
A man gently attempts to feed a baby bird.
A man is swinging a club with both hands.

Figure 3.1: Image-caption pairs (left) and how vectors representing the words ‘bat’, ‘club’, and ‘bird’ may be affected by the image information (above)

image may affect the vectors. The principle is shown in red for the combination of ‘bat’ and ‘bird’ with the left image, and in blue for the combination of ‘bat’ and ‘club’ with the right image. On the left, $\langle \text{bat} \rangle$ becomes $\langle \text{Chiroptera} \rangle$ (i.e., the vector now represents the mammal of the order Chiroptera) and $\langle \text{bird} \rangle$ becomes $\langle \text{bird-like Chiroptera} \rangle$, intuitively representing a hybrid between birds and Chiropteras. In the right, both $\langle \text{bat} \rangle$ and $\langle \text{club} \rangle$ are turned into a vector $\langle \text{club/bat} \rangle$ representing bats in the sense of clubs. While the information becomes semantically more accurate, other aspects are lost, e.g. whether the word ‘bat’ or ‘club’ was used, and probably also the fact that the second caption on the left actually mentioned a bird.

We propose probing tasks for visual-semantic embeddings (in other words, images with captions). In particular, we are interested in tasks that shed light on whether and how a given embedding makes use of the image information in relation to linguistic phenomena such as synonyms and polysemy.

Section ?? motivates our approach and relates it to existing work. Section 3.2.1 provides a systematic discussion and formalisation of probing tasks for visual-semantic embeddings. With this, we hope to map out which properties probing tasks of various types can be used to investigate. Section 3.2.2 introduces three concrete probing tasks that illustrate our approach, and which are used in our actual experiments reported on in Section 3.2.3. The code is publicly available.³ The conclusion in Section 3.2.5 summarizes our findings and lists future challenges for multimodal probing.

³<https://github.com/dali-does/vse-probing>

3.2.1 Systematic Probing for Properties of Visual-Semantic Embeddings

In this section we develop a general view of visual-semantic probing tasks, and lift the ideas of Conneau et al. to the multimodal realm. Consider a property Π that a given embedding E may or may not have. In the visual-semantic case, such a property may be “the embedding associates visual properties with the nouns in the text component” or “the embedding encodes the number of objects in the image”. A *probing task* is defined to be a machine learning task – usually a classification task – that is designed in such a way that a model can be trained on E , and the achieved performance allows to draw conclusions regarding the extent to which E possesses property Π .

We are specifically interested in developing probing tasks for visual-semantic embeddings E , where Π is a property that reflects aspects of the multimodal nature of E . Ultimately, the goal is to come up with tasks that probe how the embedding maps the individual modalities into a common space. While we are not quite there yet, below we provide a general discussion of what to look for, and how such tasks may be categorized. Probing tasks that meet the following requirements seem to be especially valuable:

1. The task is a well-defined classification problem on combined (i.e., joint or coordinated) embeddings of two or more modalities.
2. The task gives insight into whether and how the multimodal embedding integrates the modalities.
3. The task has a simple and well-defined structure, so that the results are straightforward to interpret.
4. The task can be evaluated on standard data sets, or on datasets that can be created from such.

We propose that the probing tasks are organized according to how they make use of the information in the sample data to map out embedding characteristics. For the visual-semantic case, at an abstract level, each probing task either probes the embedding of the original text-image pair (T, I) , or it is based on turning (T, I) into (T', I') in a well-specified manner, such that by comparing the performance of a classifier on $\text{emb}(T', I')$ and $\text{emb}(T, I)$, one can draw conclusions about the embedding. Depending on how T' and I' are obtained, different types of probing tasks arise.

Direct Probing

Probing tasks based on $\text{emb}(T, I)$, that is, without inflicting changes on either part, are easy to implement, but have limited potential to reveal information about the specifically multimodal characteristics of the embedding. Nevertheless, some of the probing tasks by Conneau et al. have meaningful counterparts in this context. Here, we mention only the number of concepts, which is similar

to sentence length and translates into *complexity*: given $\text{emb}(T, I)$, the task is to determine $|T|$, $|I|$, and $|(T, I)|$, where $|T|$ is the number of concepts mentioned in T (objects and properties of objects, say), $|I|$ is the number of concepts in I (i.e., the number of segments and their properties), and $|(T, I)|$ is the number of concepts in (T, I) . In the latter, an image segment and its counterpart in T would be counted only once. Note that an embedding may be expected to be ideal for determining $|T|$ and $|I|$ if it keeps the two modalities entirely separate, while good performance on the task of determining $|(T, I)|$ indicates a tighter integration.

Creation of Inconsistencies

By considering $\text{emb}(T - x + y, I)$ or $\text{emb}(T, I - x' + y')$ where y and y' do not align with x and x' , respectively, the effect of inconsistencies can be studied. For example, nouns in T aligned to objects in I may be replaced with other nouns, and similarly for adjectives referring to attributes of objects in I such as position, color, size, form, and number. Variants may rely only on injecting inconsistent information, that is, $\text{emb}(T + x, I + y)$, where x and y form an inconsistent pair such as $x = \text{ball}$ and $y = \text{cube}$. However, depending on the nature of the embedding this may require to make sure that $T + x$ is actually a reasonably well-formed sentence.

The Challenge of Interpreting Probing Results

We end this section with an urge for caution in the interpretation of probing task results, especially in the multimodal setting, and even more so when the results are “negative”.

Consider the task of determining the length of the caption of a text-image pair. If classifiers trained on this task perform well, this indicates that the embedding is not well integrated. The reason is that a well-integrated embedding would blur the distinction between the image and the caption, presumably associating a high sentence length even if a complex image is provided with a short caption. Unfortunately, the converse is not true: if classifiers perform badly, the reason may equally well be that the textual part of the embedding simply does not capture sentence length, or that the chosen classifier was unsuitable for the task. It may thus be easier to interpret a probing task that asks for the number of *objects* present in the text-image pair (see Section 3.2.2). Even in this case, poor performance does not necessarily say much about the nature of the embedding, because also a highly integrated embedding can be unsuitable for the counting task. However, despite these difficulties, this type of probing task may yield important insights if one is aware of the interpretation pitfalls.

3.2.2 Concrete Probing Tasks

This section illustrates the abstract principles introduced in Section 3.2.1 through a set of concrete probing tasks. These tasks will be experimentally tested in Sec-



- 1.1 A *child* holding a flowered umbrella and petting a yak.
- 1.2 A *checker* holding a flowered umbrella and petting a yak.
- 2.1 A young *man* holding an umbrella next to a herd of cattle.
- 2.2 A young *mime* holding an umbrella next to a herd of cattle.
- 3.1 a young *boy* holding an umbrella touching the horn of a cow.
- 3.2 a young *wad* holding an umbrella touching the horn of a cow.
- 4.1 A young *boy* with an umbrella who is touching the horn of a cow.
- 4.2 A young *bear* with an umbrella who is touching the horn of a cow.
- 5.1 A *boy* holding an umbrella while standing next to livestock.
- 5.2 A *fry* holding an umbrella while standing next to livestock.

Figure 3.2: In task *SemanticCongruence*, the objective is to recognise semantically implausible captions.

tion 3.2.3 and will, in future work, be extended with tasks of the types proposed in Section 3.2.1 to highlight complementary aspects of the semantic embeddings.

Direct Probing

Our first proposed probing tasks are instances of direct probing, as discussed in Section 3.2.1: *ObjectCategories* and *NumObjects*. In *ObjectCategories*, the task is to determine which of the 80 MS-COCO object categories are present in a given image. To turn the task into a simple classification task, we restrict the dataset to image-caption pairs in which only one of the 80 object categories is present (possibly multiple times). The second direct probing task, *NumObjects*, asks to estimate the number of object instances in the image. For this task, we bin the object instances present in an image into 6 bins (5 equidistant bins for the interval 0–29, and one bin for ≥ 30 objects).

Semantic Congruence

Detection of semantic incongruity is an example of a probing task that arises from the creation of inconsistencies (see Section 3.2.1). It reveals whether the information propagated by $emb(T, I)$ is sufficient to recognize that a caption has been modified, and to what extent this information stems from the visual part I . The associated probing task *SemanticCongruence* is the classification task that asks whether a caption has been modified. Later, we will perform this task on both $emb(T, \emptyset)$, and $emb(T, I)$. Without the image information, the decision must be based on purely linguistic features such as syntactic form, relative word frequencies, semantic consistency, and so forth. When the image is present, the model can also exploit incongruities between the modalities to detect modifications.

The characteristics of this probing tasks are largely determined by how the captions are modified, something that can be accomplished in numerous ways. FOIL-COCO by Shekhar et al. consists of modified MS-COCO pairs obtained

by choosing, from each caption, a name of an object category and replacing it by another noun taken from the same MS-COCO super category. The replaced nouns occur in more than one caption, but their substitutes are salient in that they are not among the objects annotated in the image. To create plausible captions, the authors over-generate captions and use an LSTM trained on the original dataset to keep only the highest ranking ones.

To explore a range of linguistic features broader than nouns, which are the focus of FOIL-COCO, we compile a corpus of modified captions in which the linguistic head of each caption has been replaced. The procedure for modifying a caption works as follows. First, we run the Stanford dependency parser (Qi et al., 2020) on the caption to pick out the head. The parser also provides us with a part-of-speech tag for the head, which we use as input to the classical disambiguation algorithm by Lesk. The algorithm returns the most likely synset (synset) and the abstract category assigned to the word by Wordnet. The replacement word is picked from a synset that is in the same Wordnet category. For example, if the head is ‘walk’ in the abstract category *verb.motion* then we might choose ‘fly’ from the same category. For simplicity, we avoid proper nouns. When the head is a verb, we prefer replacement words sharing the same set of frames, i.e., that can fill the same functions. Finally, we inflect the replacement word to match the inflection form of the head, and also mimic capitalization. To obtain a challenging data set, we generate $N = 10$ modified sentences for each caption and then use BERT (Devlin, Chang, et al., 2019) as a language model to select the best scoring alternative. This yields sentence pairs such as that of Figure 3.2.

3.2.3 Experiments

This section describes our experiments with direct probing (see Section 3.2.2) and semantic congruence probing (see Section 3.2.2).

Experimental Setup

Dataset We use the Microsoft Common Objects in Context (MS-COCO) dataset curated by T. Lin et al. It consists of approximately 123 000 images, each with at least five human-written captions. The object categories of the manually annotated image segments comprise 80 object categories, grouped into 11 supercategories. We use the splits provided by Karpathy and Fei-Fei, consisting of 82 783 train, 5000 validation, and 5000 training images, respectively. For testing, 5000 image-caption pairs over 1000 images are used of the test data, limited by what precomputed values are used by the investigated models. This split is originally used in training all the multimodal embeddings. We use image features precomputed by VGG19 (S. Liu & Deng, 2015) and ResNet-152 (He et al., 2016), as detailed in Table 3.2.

Models The visual-semantic models used for our probing tasks are VSE++ (Faghri et al., 2018), VSE-C (H. Shi et al., 2018), and HAL (F. Liu et al., 2020). In ad-

Model	Precomputed features	Emb. Size	Parameters
VSE++	VGG19	1024	15.5(159.2)M
VSE-C	ResNet-152	1024	13.8(74.1)M
HAL	ResNet-152	1024	11.3(71.6)M
GPT-2		768	117M
BERT		768	110M

Table 3.2: Overview of the investigated embeddings. The total size of the model, including models used to extract precomputed image features, is given in parenthesis.

dition, we use the well-known unimodal language models BERT (Devlin, Chang, et al., 2019) and GPT-2 (Radford et al., 2019a).

Following the taxonomy by Beinborn et al., VSE++, VSE-C, and HAL are cross-modal transfer models trained via joint learning on the MS-COCO dataset. The implementations of VSE-C and HAL are both based on the open source code for VSE++. We use pretrained versions of these models, as provided with the respective papers. VSE++ learns visual-semantic embeddings by incorporating hard negatives into the loss function and using a similarity function that scores higher for the correct image-caption pairs than for the semantically incorrect ones (that is, for the negative samples). VSE-C learns instead by manipulating the original captions in the MS-COCO dataset so that they constitute contrasting image-caption pairs. HAL uses the same architecture as VSE++, but tries to avoid the so-called hubness problem where the results are skewed by frequently occurring vectors, by making the loss function aware of such structural properties of the data.

As all $X \in \{\text{VSE}++, \text{VSE-C}, \text{HAL}\}$ embed the two modalities individually (though trained on the actual multimodal data), each results in two separate models X_{text} and X_{image} . We use these models in our experiments, in addition to “true” multimodal models X_{avg} and X_{conc} obtained by averaging and concatenating (resp.), the corresponding vectors in X_{text} and X_{image} .

BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional language model introduced by Devlin, Chang, et al. It considers the left and right context surrounding a word, and relies on unsupervised learning to pre-train deep bidirectional language representations. We use an existing BERT model trained on the BookCorpus with 800 million words and on the English Wikipedia pages with 2 500 million words (Devlin, Chang, et al., 2019). The last model, GPT-2 (Generative Pre-Training, second generation), is a transformer-based unidirectional language model trained on 40GB of lightly curated Internet text (Radford et al., 2019a). We use the Transformers library for these models (Wolf et al., 2019a).

Probing We perform the three classification tasks *ObjectCategories*, *NumObjects*, and *SemanticCongruence*. For *NumObjects*, the label distribution between

the 6 bins (see Section 3.2.2) is 47 443, 17 580, 9 626, 4 549, 2 061, 1 524 during training and 3 025, 1 060, 470, 240, 130, 75 during testing. Our baseline is naively guessing the largest class. The *ObjectCategories* task is based on 9 629 and 1 145 samples in the training and test data, respectively. For the *SemanticCongruence* task, a modified caption is chosen with probability 0.5, and remains the same for all models tested for fair comparison. We note that more complex models could yield higher accuracies, but following the results of (Hewitt & Liang, 2019) on probe model selectivity, this improvement does not necessarily reflect the availability of the information probed for. Therefore, we use two classifiers for probing; one multilayer perceptron (MLP) with one hidden layer of 256 nodes and sigmoid activation, and one linear classifier with softmax activation. Both models use a dropout of 0.2, similar to (Conneau et al., 2018). The probing models are trained on the MS-COCO data for 30 epochs using the cross-entropy loss function. In all cases the models start to converge within the last 10 epochs. The results reported are for the test split.

Embeddings For each probing task, the input to the classifier is either the image embedding from one of our used models ($\text{VSE}++_{\text{image}}$, $\text{VSE-C}_{\text{image}}$, $\text{HAL}_{\text{image}}$) or the text embedding from one of our models ($\text{VSE}++_{\text{text}}$, $\text{VSE-C}_{\text{text}}$, HAL_{text} , BERT, GPT-2). For the size of the embeddings we refer to Table 3.2. For *SemanticCongruence* the X_{text} input consists also of modified captions. In order to contrast the probing results obtained with those for embeddings containing the full visual-semantic information, we also consider X_{avg} and X_{conc} , for $X \in \{\text{VSE}++, \text{VSE-C}, \text{HAL}\}$. All weights of each model are frozen, meaning that no weights are updated for the embedding models during the probing.

Experiment Details The implementation is written in Pytorch 1.4.0 and trained on a NVIDIA Tesla V100 32GB GPU using CUDA 10 with Tensorflow 2.1. The models are all trained for 30 epochs, where each epoch times in at 100 seconds on average, and the experiments are conducted using the Adam optimizer with learning rate 1.0×10^{-4} for *ObjectCategories* and *NumObjects*, and 1.0×10^{-3} for the *SemanticCon* probing task. During initial experiments SGD was also considered but Adam showed better performance.

The implementations of $\text{VSE}++$ ⁴, VSE-C ⁵, and HAL ⁶ the open sourced Github repositories with the best corresponding pretrained models are used. For BERT and GPT-2, the Python library Transformers⁷ is used to access pretrained models. In both cases the base model is used, since initial experiments showed no significant difference when using larger models and in interest of keeping the comparison fair given that the larger models are substantially larger than the visual semantic embedding models. All pretrained models are outlined in Table ???. Random numbers generated with `Numpy` uses a fixed seed of 1974, to make the experiments reproducible.

⁴<https://github.com/fartashf/vsepp>

⁵<https://github.com/vacancy/VSE-C>

⁶<https://github.com/hardyqr/HAL>

⁷<https://github.com/huggingface/transformers/>

Embedding	ObjectCat.		NumObjects		SemanticCon.	
	MLP	lin	MLP	lin	MLP	lin
<i>Baseline</i>	-		0.605		0.502	
<i>Image</i>						
VSE++ _{image}	0.753	0.768	0.646	0.613	0.502	0.506
VSE-C _{image}	0.754	0.675	0.654	0.629	0.503	0.504
HAL _{image}	0.799	0.730	0.674	0.633	0.533	0.510
<i>Text</i>						
VSE++ _{text}	0.862	0.863	0.627	0.610	0.739	0.710
VSE-C _{text}	0.838	0.805	0.629	0.617	0.763	0.756
HAL _{text}	0.826	0.648	0.625	0.611	0.730	0.737
BERT	0.878	0.365	0.622	0.599	0.816	0.768 ⁸
GPT-2	0.811	0.137	0.617	0.585	0.792	0.718
<i>Merged</i>						
VSE++ _{avg}	0.862	0.876	0.658	0.638	0.707	0.662
VSE++ _{conc}	0.911	0.901	0.661	0.641	0.743	0.713
VSE-C _{avg}	0.831	0.783	0.665	0.636	0.735	0.713
VSE-C _{conc}	0.896	0.879	0.666	0.652	0.776	0.758
HAL _{avg}	0.847	0.820	0.667	0.642	0.712	0.702
HAL _{conc}	0.903	0.849	0.683	0.648	0.730	0.730
<i>Improvement by merging</i>						
VSE++	0.049	0.038	0.015	0.028	0.040	0.003
VSE-C	0.058	0.074	0.012	0.023	0.013	0.002
HAL	0.077	0.119	0.009	0.015	0.000	-0.007

Table 3.3: Probing accuracies using a MLP with embeddings as input. The bottom three show for each model the difference between the best unimodal and the best merged embedding. All results are averaged over 5 runs and have variance ≤ 0.01 .

3.2.4 Results and Analysis

Table 3.3 shows the results for the *ObjectCategories*, *NumObjects* and *SemanticCongruence* tasks, using both a MLP and a linear probe. While the two types of probes perform differently, the relative behavior across embeddings is very similar. A notable deviation from this general rule is the performance of BERT and GPT-2 on the *ObjectCategories* task (see below).

ObjectCategories We note that the **text-only** embedding for all three visual-semantic models yields better performance on the *ObjectCategories* task than the corresponding text-only embedding, with the exception of the linear probe for **HAL**. Worth noting is that **out of the multimodal embeddings, HAL performs well on image-only but worst on text-only for both probes**. Further, there is a large gap between the performances of MLP and linear probes on the BERT

⁸Since BERT is used during the generation of congruencies, this result is somewhat self-referential.

and GPT-2 embeddings. This supports the conclusion of Hewitt and Liang that MLPs, rather than acting as probes, may simply learn the task itself if provided with sufficiently rich embeddings as input, and that, therefore, linear probes may be a more appropriate choice.

Note also that BERT performs **best for both probes in the text-only case, while GPT-2 scores the lowest**. All merged embeddings significantly outperform their corresponding unimodal embeddings, with **concatenated VSE++ scoring the highest for both probes**. Merging the embeddings shows an improved accuracy of **3.8–11.9%** across both probe types, which suggests that the visual-semantic models combines the multimodal data in a useful way to capture which objects are present in a scene. Overall **VSE++** seems to best capture and combine information about the object categories, beating BERT and GPT-2 by a large margin for both probes.

NumObjects The results for the *NumObjects* task show that the text embeddings consistently encode the probed information in a less accessible manner than the corresponding image embeddings which are, in turn, outperformed by their merged counterparts. Using MLP probing, HAL reaches the highest accuracy on both image-only and its merged embeddings, whereas VSE-C appears to be on par with HAL on merged embeddings under a linear probe, the precise result depending on the merging strategy. It is worth noting that the improvements from merging the embeddings are small, but are larger when using a linear probe. Once again, this supports the conclusion of Hewitt and Liang as it indicates that the weaker probes exhibit a better sensitivity.

It is worth noting that the best result for the *NumObjects* task is only about 8% better than the baseline. This seems to indicate that the task could be improved. Most of the images contain fewer than 10 object instances, thus falling into classes 1 and 2.⁹ Table 3.4 display the per-class accuracy, showing that the accuracy for most embeddings and models is above 90% for class 1, and between 30–50% for classes 2,3, and 6. Classes 4 and 5 (i.e., 18–23 and 24–29 object instances) yield accuracies of approximately 4–18% and 3–15%, respectively. Further, the per-class accuracies show that the linear probes show performance comparable to the MLP probe on the first three classes, but never learn the 24–29 object class, and very few of the 18–23 and ≥ 30 samples.

Image scenes containing 0–5 object instances can exhaustively be described with words, mentioning numbers and listing distinct objects explicitly (“a cup and a fork”), whereas scenes containing 18–29 objects are harder to explicitly describe. The high accuracy for scenes with more than 29 objects may be due to the fact that the large number of object instances is a “property of the image” and might therefore be described with words such as “crowd”. A more balanced distribution could amplify the differences. Table 3.4 show that for class 1 (i.e. 0–5 object instances) the performance of the text-only embeddings is slightly better than for the multimodal embedding, which in turn performs slightly better than the image-only embedding. The results for class 1 is in average 0.9 for

⁹Remember that we have 6 output labels representing the number of object instances.

all three embeddings. For class 2 and 3 (i.e. 6-11 and 12-17 object instances) the performance for all three embeddings drops significantly (to an average of 0.2) and text-only has the lowest performance whereas the performance of multimodal embeddings is higher than for image-only embeddings. For class 4 and 5 (i.e. 18-23 and 24-29 instances) the performance continues to drop for all three embeddings with text lowest performance and slightly better or worse performance for image-only and multi-modal (depending on the considered model and embedding). For class 6 (30+) the performance for all embeddings and models increase again (and is similar to that in class 2, 6-11 objects). 30+ text lowest again where the multimodal embedding are better for VSE++ avg (not conc) and for HAL mm does not add anything.

Model	0–5		6–11		12–17		18–23		24–29		≥ 30	
	MLP	lin	MLP	lin								
<i>Image</i>												
VSE++ _{image}	0.918	0.946	0.322	0.227	0.252	0.085	0.192	0.004	0.120	0.000	0.400	0.184
VSE-C _{image}	0.923	0.928	0.397	0.338	0.280	0.254	0.062	0.073	0.000	0.000	0.480	0.533
HAL _{image}	0.909	0.966	0.414	0.209	0.330	0.021	0.154	0.000	0.040	0.000	0.533	0.000
<i>Text</i>												
VSE++ _{text}	0.927	0.926	0.287	0.256	0.153	0.090	0.039	0.038	0.000	0.000	0.240	0.133
VSE-C _{text}	0.920	0.962	0.390	0.218	0.287	0.049	0.105	0.000	0.002	0.000	0.560	0.187
HAL _{text}	0.937	0.958	0.264	0.148	0.124	0.000	0.085	0.000	0.024	0.000	0.227	0.000
BERT	0.961	0.992	0.162	0.040	0.132	0.000	0.000	0.000	0.032	0.000	0.200	0.000
GPT-2	0.921	1.000	0.216	0.000	0.230	0.000	0.038	0.000	0.054	0.000	0.240	0.000
<i>Merged</i>												
VSE++ _{avg}	0.909	0.944	0.380	0.311	0.301	0.136	0.115	0.038	0.056	0.000	0.427	0.240
VSE++ _{conc}	0.930	0.940	0.375	0.315	0.310	0.165	0.181	0.054	0.152	0.000	0.320	0.240
VSE-C _{avg}	0.928	0.963	0.337	0.281	0.254	0.101	0.077	0.020	0.000	0.000	0.533	0.253
VSE-C _{conc}	0.927	0.951	0.294	0.336	0.166	0.139	0.040	0.069	0.008	0.000	0.251	0.280
HAL _{avg}	0.925	0.967	0.402	0.265	0.254	0.163	0.127	0.015	0.080	0.000	0.507	0.077
HAL _{conc}	0.920	0.947	0.430	0.329	0.303	0.223	0.173	0.000	0.112	0.000	0.520	0.000

Table 3.4: Accuracy per label of the tested models A more detailed account of the accuracy of the tested models for the task *NumObjects*. The class labels correspond to the number of objects annotated in the image..

SemanticCongruence The results obtained from the *SemanticCongruence* probing suggest that the additional information provided by the multimodal component does not make up for the relative loss of linguistic information. This becomes particularly clear when using linear probing. VSE-C_{text} outperforms VSE++_{text} and HAL_{text}, but is in turn clearly outpaced by the unimodal embeddings BERT and GPT-2. If we add visual information (to VSE++, VSE-C, and HAL), the performance generally does not increase, and even decreases in one instance. Our interpretation is that the alternative captions can be recognized from linguistic patterns such as verb-preposition agreement and other

contextual information solely from having a good language understanding. Going back to Figure 3.2, we recognize that a well-formed sentence can still be highly unlikely given an understanding of language, just as Chomsky’s famous example “Colorless green ideas sleep furiously” (Chomsky, 1975). Further, although the visual information could provide additional clues, it also adds noise and makes the relative proportion of linguistic data smaller. If this interpretation is correct, an improved linguistic quality of the alternative sentences should make the visual information more valuable for the task. Since this visual information cannot encode whether the caption was modified, $\text{HAL}_{\text{image}}$ aligns with the results of Hewitt and Liang, suggesting that this MLP probe learns something other than the probing task. Finally, we note the good performance of BERT despite the fact that BERT was the embedding used to select the most convincing alternative captions, which should make them particularly apt at confusing BERT.

TODO Fix figures and description This section gives a sample of images from MS-COCO, together with original captions ($x.y.1$), where $x \in \mathbb{N}$ identifies the image and $y \in \{1, \dots, 5\}$ identifies the original caption, and two series of modified versions, one series ($x.y.2$) of lower quality, and one series ($x.y.3$) of higher quality, both modifying the caption $x.y$. We use the higher quality series in the semantic congruence task. The modified versions illustrate some of the challenges of automatically generating syntactically valid alternatives. The most common reason for poor captions is that lexical disambiguation has failed, or that we do not control for verb-preposition coherence.



- 1.1.1 This is a *case* full of yellow bananas.
- 1.1.2 This is a *giant* full of yellow bananas.
- 1.1.3 This is a *squirt* full of yellow bananas.
- 1.2.1 *Bananas* a tightly packed and boxed for delivery to the market.
- 1.2.2 *Ivy* a tightly packed and boxed for delivery to the market.
- 1.2.3 *Bunt* a tightly packed and boxed for delivery to the market.



- 2.1.1 That *looks* like a wall mural in the background of this photo ...
- 2.1.2 That *occupies* like a wall mural in the background of this photo ...
- 2.1.3 That *runs* like a wall mural in the background of this photo ...
- 2.2.1 A huge heard of sheep are all *scattered* together.
- 2.2.2 A huge heard of sheep are all *pumped* together.
- 2.2.3 A huge heard of sheep are all *resurfaced* together.

Summary We see that the multimodal embeddings in the *merged* section of Table 3.3 outperform their **image- and text-only** embeddings on the tasks *ObjectCategories* and *NumObjects*. This indicates that the text- and image-only embeddings complement each other in what information they encode, and that merging them can utilize this fact. The concatenated embeddings yield consistently better performance than the averaged ones, probably because the complementary information is fully retained. It is not clear how well the text- and image-only embeddings project to the same space, which together with the introduction of noise from the respective modality can cause averaging to drown out important information. Still, averaging gives better performance than unimodal approaches **except for VSE-C on ObjectCategories**. The first two tasks are highly visual, which makes it only reasonable that the image embeddings encode more information of concern in these problems. It is also **suggested** from the results that the state-of-the-art unimodal text embeddings have a better semantic language understanding. It seems that there is a trade-off between language modeling versus understanding visual concepts, and that the training of the multimodal models has favored the latter. This idea also aligns with the fact that these models are built for image-to-text and text-to-image retrieval, a task for which the unimodal embeddings are insufficient. Interestingly, HAL seems to be more focused on visual information as seen in the results on *ObjectCategories* and *NumObjects*. This could help explain why HAL outperforms VSE++ and VSE-C on text-to-image and image-to-text retrieval. We also note, importantly, that the language models are larger by factors 1.45 up to 10, excluding/including the network used to precompute image features, respectively. This can help explain why the multimodal models are not as capable in distilling the probed information in the text-only *ObjectCategory* task as BERT. To conclude, the results show that the image and text embeddings complement each other in understanding visual concepts, but that this does not extend to the understanding of language itself, as shown in the results on *SemanticCongruence*. Therefore, we conjecture that there is significant room for improvement on the multimodal embeddings for understanding scene semantics.

3.2.5 Conclusions

Probing semantic embeddings with neural-network based classifiers is like looking into a black box with a lens that is itself a black box. Valuable information can still be derived, but experiments that take this approach must be made with care, and the results analysed with caution. One approach to mitigate such opacity is proposed by Hewitt and Liang, namely that the probing task is complemented with a control task to alleviate a possible misinterpretation of what semantic representations actually encode.

In the multimodal setting, it is helpful to use probing tasks (as well as complementing control tasks) that are simple, well-defined, and easily implemented on standard data sets. The importance of a task being well-defined is illustrated, albeit in a negative way, by the *NumObjects* task: Since there are countless equally valid ways to semantically decompose an image, it can simultaneously

be true that an image shows dozens of sheep and that it shows a single herd. The flaw is arguably not as much in the task itself, as in the combination of task and data set. We may, for example, expect that the *NumObjects* task comes to its right in situations where logical units of counting are understood in advance, e.g., in the case of camera footage tracking traffic congestion, where a natural unit would be the number of vehicles. An interesting finding from our initial experiments was the importance of linguistic compared to visual information for complexity estimation and semantic incongruity detection.

3.3 NOT REWRITTEN AT ALL - Bridging Perception, Memory, and Inference through Semantic Relations

In Chapter ??, we saw how Bender et al.; Bender and Koller postulate that it is impossible to learn meaning from surface form alone, and express concerns about what is perceived as an over-reliance on large-scale pretrained neural networks. This line of thought supports the interest in hybrid systems that amalgamate elements from complementary learning paradigms (see, e.g., (Hohenecker & Lukasiewicz, 2020a; Pearl, 2019; van Bekkum et al., 2021; P.-W. Wang et al., 2019b)). In (Dahlgren et al., 2021), we argue that this calls for an explicit distinction to be made between the faculties of perception, memory, and inference. We therefore promote the development of systems that consist of subsystems with responsibilities corresponding to the three faculties. Such future systems would thus consist of a perception component realised by a neural network, a component that provides explicit memory in the form of a knowledge base, and a third one performing symbolic inference, that is, rule-based reasoning.

We suggest to study how the subsystems can be aligned so for a seamless information flow between them. We view it as particularly important that (i) the network and the knowledge base together yield a consistent treatment of semantic relations and (ii) training takes the knowledge base into account, so that the resulting embeddings are consistent with established facts. Our conceptual discussion is complemented by a preliminary empirical evaluation of six popular English language models, which we subject to linear probes to test their abilities to capture central semantic relations.

After a brief discussion of related work in Section 3.3.1, Section 3.3.2 discusses the role of semantic relations in the context of our envisioned triad system while Section 3.3.3 and 3.3.4 of this paper complement our conceptual discussion with a preliminary empirical evaluation of the chances to achieve (i) by probing six popular language models with respect to a semantic relation learning task.

3.3.1 Related work

There is a rapidly growing literature on relation extraction and hybrid systems. Petroni et al. observe that language models such as BERT (Devlin, Chang, et al., 2019) and GPT-3 (T. B. Brown et al., 2020) are imprinted with large amounts of common sense and factual knowledge during training. If this information can be reliably extracted then, they argue, word embeddings could find a new use as knowledge bases. To test the practicality of this approach, they consider a knowledge extraction task where a language model is given a sentence containing a subject word x and a relation \mathcal{R} , but where the object word y has been removed, and the model should guess the missing y (i.e., rank the vocabulary words) based on the fact that x and y are in the relation \mathcal{R} . The sentences are generated based on manually constructed templates, one per relation. For example, to the relation *birth-place*, they use the template “⟨subject⟩ was born in ⟨blank⟩” and instantiate it to “Dante was born in ⟨blank⟩”. The most important baselines are two variations of the relation extraction model by Sorokin and Gurevych. Key findings are that language models appear to be better at learning one-to-one relations, whereas the relation extraction models are better at picking out many-to-many relations. Petroni et al. also find that the choice of template has an impact on the performance of the language models, and point this out as an item for future work.

Bouraoui et al. pick up this thread and propose a method for extracting good template sentences from BERT, and using these to fine-tune BERT so as to improve its performance on relation extraction. For a target binary relation \mathcal{R} (represented as a set of ordered pairs) and a sample of pairs $R \subseteq \mathcal{R}$, they filter the training data for sentences expressing that x and y , with $(x, y) \in R$, have the relation \mathcal{R} , and which would still be natural if x and y were simultaneously replaced by some other $(x', y') \in R$. Finally, they fine-tune a language model to predict, from an instantiation of one of the remaining sentences with a pair (x'', y'') , whether $(x'', y'') \in \mathcal{R}$. The most relevant aspect of this work for the present effort is the evaluation of the Bigger Analogy Test Set (also known as BATS) which contains 40 relations with 50 instances per relation (Gladkova et al., 2016). Bouraoui et al. report a mixed performance on the type of semantic relations considered here, namely hypernyms and hyponyms.

Additional methods for choosing template sentences are proposed by Jiang et al. who, similar to Bouraoui et al., mine the training data for suitable sentences. A dependency analysis on candidate sentences makes it possible to extract a larger variety of phrases that express the desired relationship than Bouraoui et al. can. The authors also generate candidate sentences by paraphrasing. In short, they find that both mined and paraphrasing have their usages, and that combinations of template types, e.g., manually constructed and mined, often perform well.

Poerner et al. question the conclusion by Petroni et al. that BERT contains factual knowledge derived from the training data. The authors believe that in may cases, BERT simply exploits superficial similarities and general patterns to guess what is most likely. For example, from the fact that a person has a

typically French surname, BERT could guess that that person is actually French without having learned the nationality of the particular person. To expose this weakness, (Poerner et al., 2019) remove what they believe are easily guessed pairs of subjects and objects from the data set of (Petroni et al., 2019b). They also provide a modified version of BERT, E-BERT, in which the embeddings of entities mentioned in Wikipedia have been replaced by a symbolic entity embedding. They find that E-BERT outperforms both BERT and ERNIE on the trimmed data set, but also that a combination E-BERT and BERT (taking the average of or concatenating the embeddings) give higher accuracy than either on its own.

Rosenbloom model different types of declarative and procedural memory with what is essentially weighted hypergraphs, in which nodes correspond to actions and conditions, and edges to activation functions. Procedural and declarative memory are distinguished based on the direction in which values are propagated through the hypergraph. The analogy to human cognition is that procedural memory contains information about how to do something, whereas declarative memory concerns facts and events.

3.3.2 The role of semantic relations

As the brief account given in the previous section shows, there is a solid body of work on the extraction of relations from language models (see Section 3.3.1), to derive facts such as that the birth place of Olga Tokarczuk is Sulechów, Poland, and that the capital of Bolivia is La Paz. Looking to knowledge bases, it is natural to view them as graphs, where nodes represent objects and properties, and edges represent semantic relations. Finally, for logical inference, basic semantic relations such as synonymy, hyponymy, hypernymy, and meronymy play a central role. We recall that words are synonyms if they have (nearly) the same meaning; that a hypernym of a concept is a generalisation of that concept (e.g., ‘bird’ is a hypernym of ‘sparrow’), while a hyponym is an instance of the concept (e.g., ‘spider’ is a hyponym of ‘arachnid’), and that a meronym of a concept is a part of the whole (e.g., ‘branch’ is a meronym of ‘tree’); see Table 3.5 for examples found in WordNet (Miller, 1992).

For logical inference, we can infer that starfish are not fish from knowing

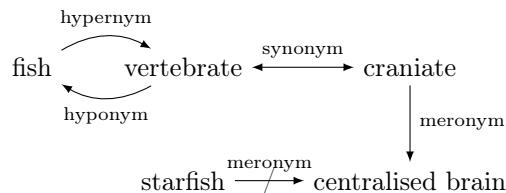


Figure 3.3: In this work we focus on recovering synonyms, hypernyms, hyponyms, and meronyms from natural language models via probing to understand the prerequisites of integration with knowledge bases.

Synonymy		Hyponymy		Meronymy	
band	set circle ring	assumption	theory miracle audacity	house	library attic porch
office	agency bureau authority	copper	metal penny policeman	road	bend crossing turnout
origin	root source blood	correction	improvement therapy punishment	song	words language chorus

Table 3.5: Instances of the relations synonymy, hypernymy, and meronymy extracted from WordNet.

that ‘heart’ is a meronym of ‘craniate’ but not of ‘starfish’ (all craniates have hearts whereas starfish do not), ‘vertebrate’ is a hypernym of ‘fish’ (fish are vertebrates), and ‘craniate’ is a synonym of ‘vertebrate’. See Figure 3.3 and Table 3.5 for further examples.

To achieve a seamless integration of a neural network with a knowledge base of relations and an inference engine, we propose to devise methods for (i) enabling the network to utilise the knowledge base, but fall back on the less certain information in the embedding when necessary and (ii) taking the relations in the knowledge base into account during network training, so that the trained network reflects the contents of the knowledge base. In this endeavour, we believe that particular emphasis should be placed on the treatment of lexico-semantic relations such as meronymy, hyponymy, and synonymy because of their central role in logical deduction and lexical semantics.

3.3.3 Empirical study: method

To gain some initial insight into how well state-of-the-art pretrained contextual embeddings handle lexico-semantic relations, we conducted experiments on word embeddings generated by ALBERT (Z. Lan et al., 2020), ROBERTa (Y. Liu et al., 2019), BERT (Wolf et al., 2019b), and GPT-2 (Radford et al., 2019b). We also included Word2Vec (Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) models in our experiments, for comparison. These are all self-supervised learning algorithms, based on neural networks and built to translate words into vector representations. BERT and GPT-2 are transformer models, each having 12 encoder layers. ROBERTa is a retraining of BERT on a larger data set, while ALBERT is an extension of BERT that has a higher data throughput with 10x fewer parameters, and thus scales better.

In contrast to the works discussed in Section 3.3.1, we do not extract relations from the embeddings by means of linguistic templates. Rather, we view sentence

extraction as an instance of *probing* (Conneau et al., 2018; Hupkes et al., 2020a; A. Rogers et al., 2018; Yaghoobzadeh et al., 2019), a diagnostic method to reveal what aspects of the input the embedding actually encodes. Probing tasks should ideally be agnostic as to the underlying encoder architecture, so that results are transferable between embeddings (Dahlgren et al., 2021; Hewitt & Liang, 2019). Random control tasks (Hewitt & Liang, 2019) are implemented, see discussion in Section 3.3.4. In our experiments, we considered the following probing task: Given a pair of word vectors, we ask whether the encoded words are in relation \mathcal{R} . This avoids the optimisation problem linked to the choice of template seen in (Petroni et al., 2019b).

All experiments are on the English language, and the data set used in our experiments was obtained from WordNet as follows. We first built a vocabulary V by taking the 5 000 most common nouns in the Brown corpus (Kucera & Francis, 1967) and removing those not found in WordNet (Miller, 1992). This resulted in a vocabulary of 3497 words. For each word w in the vocabulary V and target relation $\mathcal{R} \in \{\text{hypernym}, \text{meronym}, \text{synonym}\}$ we then picked words v and v' in V such that $(w, v) \in \mathcal{R}$ and $(w, v') \notin \mathcal{R}$, and stored these as triples (w, v, v') .

We formulate a classification task for each relation \mathcal{R} , and probe each of the investigated models for their ability to capture each relation in their respective embeddings. Each classification task is based on 1 712, 306, 2 740, 1 630 samples for synonyms, meronyms, hypernyms, and hyponyms respectively. We use a linear classifier probe as these better reflect the availability of the information probed for, as shown in (Dahlgren et al., 2021; Hewitt & Liang, 2019). From (w, v, v') , positive (w, v) and negative (w, v') examples are drawn with equal probability, labeled either 0 or 1, to represent if the tuple represents a negative or a positive pair. The binary labels are given together with either (w, v) or (w, v') as input to the probe by concatenating both word embeddings. We train the probe for 10 epochs using 5-fold cross validation, using softmax activation, dropout of 0.2 to prevent memorising samples, and cross-entropy loss with the Adam optimizer using a $lr = 0.001$. We average the results over 5 runs. The experiment is implemented with Pytorch for CPU and uses the Huggingface (Wolf et al., 2019b) library for all pretrained transformers, and the Gensim (Rehurek & Sojka, 2011) library for word2vec and GloVe. The experiments completed within 1 hour on an Intel i7-based Linux laptop with 32GB RAM. The code is available on Github¹⁰.

3.3.4 Results and discussion

Table 3.6 displays the numerical results, with the header row showing, for each relation \mathcal{R} , the size of the larger of the two classes. This number coincides with the control tasks implemented to measure selectivity, which are omitted to limit redundancy. The table shows linear probe classification accuracy for each language model, with the variance written out within parentheses. As can

¹⁰<https://github.com/dali-does/semprof>

Embedding	Synonyms (50.1)	Meronyms (54.2)	Hypernyms (51.0)	Hyponyms (50.7)
Word2Vec	61.5 (1.8)	68.8 (5.0)	69.1 (1.5)	54.1 (1.7)
GloVe	63.2 (2.3)	73.3 (6.0)	68.7 (2.0)	55.7 (1.7)
ALBERT	51.9 (2.6)	48.7 (2.2)	51.2 (1.8)	51.7 (2.9)
ROBERTa	61.7 (1.9)	62.7 (5.9)	64.1 (1.2)	58.2 (2.8)
BERT	56.7 (1.2)	57.2 (3.6)	64.2 (1.6)	51.1 (0.3)
GPT-2	58.0 (1.2)	61.8 (5.3)	65.0 (1.3)	52.4 (2.5)

Table 3.6: The probing accuracy on the semantic relations, with variance given in parentheses. The accuracy of a “largest class” strategy is shown next to each relation. All transformers give embeddings of 768 dimensions, with word2vec and GloVe using 300 dimension. Each relation contain 1712, 306, 2740, and 1630 samples, respectively.

be expected, the variance is highest for meronyms where there is least data. Various observations can be made by comparing the results for the individual embeddings. Particularly worthwhile noting is the fact that GloVe and word2vec performs on par or better than the contextual embeddings, except for the case of hyponyms. This behaviour was seen with 5 and 20 training epochs as well.

The relatively strong performance of the pre-transformer solutions may not be surprising as far as synonyms are concerned, since their construction builds around aligning words found in the same context. However, we would not have expected similar results for hypernyms and even lesser so for meronyms. We note that ALBERT does not accessibly encode any of the relations, resulting in random guesses. This could be because ALBERT is trained using tenfold fewer parameters to produce much smaller embeddings, and might have less room for this type of information. Since ALBERT is comparable in performance to, e.g., BERT on many data sets and other metrics, this needs further investigation to see to what extent these relations are present in the data sets. The complexity of the probe could also be the culprit, as an embedding with lower dimensionality poses a more difficult task for a probe with limited capabilities of separating intertwined concepts. These results do not mirror those of Z. Lan et al., which indicates that the relations studied here could receive more attention in future evaluations of language embeddings. ROBERTa seems to generally outperform the other transformers, especially on the hyponyms, taking into account that not all results are statistically significant. Hypo-/hypernym relations usually follows a tree hierarchy, with hypernyms directed towards the root. This gives a decreasing number of hypernyms, for example, `fish` has six hypernyms but 39 hyponyms in WordNet, and it is likely that less common words will be chosen as a positive example for hyponyms. Weighting the words according to frequency could show different results, but filtering words based on the data the models are trained on is counterproductive to the purpose of these probes. ROBERTa is better able to capture synonyms, which could be an effect of the much larger dataset used in training compared to the other BERT-models leading to more

of the less common examples of hyponyms being seen more. One hypothesis on why GPT-2 also shows poor performance is that Wikipedia is removed from the training data. The proposition is that many Wikipedia articles explicitly outlines hyponym relations, e.g. in “*The cat is a [domestic species of small carnivorous] mammal*”¹¹.

Summarising the results, the fact remains that according to our probes no model covers the relations reliably. If this observation is confirmed by further experiments, it supports the case for a combination of neural networks, traditional relational knowledge bases, and inference engines. With this architecture, established facts could be retrieved from the knowledge base and complemented by less certain facts deduced by the network to cover up for missing information without causing inconsistencies. The results also indicate that a significant threshold should be applied for transferring relational knowledge derived from an embedding to a knowledge base, if this should be done at all, to avoid large error propagation. This is especially important if the “facts” in the knowledge base are considered to be absolute truths rather than tentative findings.

In conclusion, the reliability of the probe could improve with evaluation sets from relations found in knowledge bases, and a correlational study between probing accuracy and downstream NLP tasks could further support the usefulness of studying these relations.

3.4 Probing to understand compositionality

Recent work to write about

- (Weiss et al., 2021)
- (Keysers et al., 2020)
- (Belinkov, 2022)
- (Kobayashi et al., 2020)

Sikarwar et al. use ReaSCAN to gain insight into how multimodal transformers learn to compose. They propose an extended set of data splits targeting compositional generalisation, and show how multimodal transformers are capable to generalise systematically to some extent. They complement the performance over these splits with a linear probe using the self-attention weights, Figure 3.4 shows the probing results. This indicates that the learned model does good job but does not manage to disentangle perfectly.

3.4.1 Compositionality of CLIP on CLEVR attributes

TODO PROBING IS A METHOD TO MEASURE COMPOSITIONALITY, WHICH IS EXACTLY WHAT IS DONE IN OUR EXPERIMENTS

¹¹<https://en.wikipedia.org/wiki/Cat>

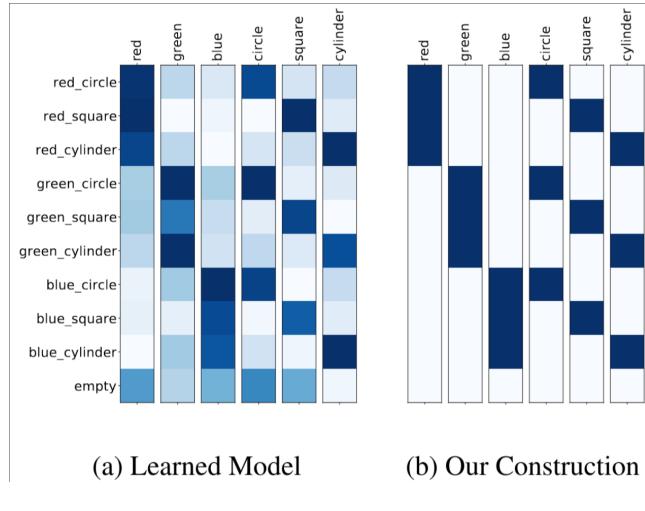


Figure 3.4: Image from (Sikarwar et al., 2022) todo ask for permission.

3.5 Challenges and characteristics

Chapter 4

Studying the behaviour of multimodal language models on reasoning and compositionality

Example of how to use quotes at
the beginning of chapters

dali

We established in Chapter 3 that multimodal language models represent visual concepts more distinctly. In this chapter, we will study how multimodal language models behave on reasoning tasks over visual scenes. As it turns out, compositional generalisation is central to achieve good performance on the mathematical problems we devise. It also turns out that the available benchmarks are not complex enough, or does not cover multimodal data. Previous literature highlights how neuro-symbolic methods are much stronger on reasoning tasks. Hence, we are on a quest to construct multimodal benchmarks that compare neuro-symbolic and deep learning methods on reasoning and compositional generalisation.

4.1 NOT REWRITTEN AT ALL - Perception, Memory, and Inference

As an answer to recent contributions about the conjectured impossibility of learning meaning from surface form alone, and the dangers of large language models, we argue in this paper that an explicit distinction should be made

between (i) perception, (ii) (explicit) memory, and (iii) inference. We envision a triad of interacting subsystems with corresponding responsibilities (see Figure 4.1). *Perception* provides the interface between the system and its environment by learning and recognising patterns, which is typically realised in the form of a language model. *Explicit memory* is a structure of concepts and relations between the concepts, in other words, a knowledge base of facts. *Inference*, finally, corresponds to mathematical or rule-based reasoning and provides, for example, classical logic and arithmetic. We note here that such axiomatic systems cannot be deduced from data, only conjectured and tried against data. Our position is that natural language systems should thus combine continuously updated language and other perception models (e.g., computer vision) with one or more symbolic knowledge bases that relieve the models from learning concepts and their relations, and finally one or more inference engines to provide formal reasoning. It is said that deep learning is the next best hammer to every nail. However, one should not forget that the goal must be building houses rather than producing hammers. Good tools are required, but they are not ends in themselves. In a multimodal natural language context, deep learning is a powerful and versatile tool for processing text and images as far as perception goes. Bender and Koller conjecture that one cannot learn the relation between the surface form of language and its communicative intent from the former alone (Bender & Koller, 2020a), and Bender et al. (Bender et al., 2021a) argue that the race for bigger and bigger datasets and language models should be seriously questioned as it carries considerable environmental, social, and scientific risks. Training these models consumes huge amounts of energy, big datasets are skewed against minorities and the underprivileged, and the focus on size hampers progress towards more intelligent solutions. We think that these are valid points, but would like to qualify the by two comments:

First, learning from large datasets is not inherently problematic. A human who is given access to the type of datasets used for knowledge extraction can, over time, be expected to learn useful facts and be wiser for the experience. For example, a reader who has been taught how the shape of the earth can be derived from physical observations will not soon adopt the flat earth theory from reading about it in online media. Human rationality makes us more robust than neural networks with respect to how new data points affect us. Our beliefs, values, and reasoning abilities are the “knowledge base” that makes us, if not immune, then much less susceptible to integrating misleading information. In contrast, current deep learning does not have the capability to discard a data point as false and choose not to learn from it, but must take everything encountered at face value and adjust the network parameters accordingly. What makes large datasets problematic are therefore the specific conditions of the learning process that characterise deep learning. We believe that the addition of symbolic knowledge would make overly large models unnecessary: having access to a multiplication algorithm is more space efficient than memorising a huge multiplication table, which can in any case never be complete.

Second, learning surface form is a challenging research problem and the improvement of existing techniques can create substantial value if, as argued

earlier, language models are not viewed as complete world models, but rather as models of form that may be compatible with any number of worlds. In their work, Bender and Koller do not address the downstream systems that make use of pretrained language models. However, we would like to suggest that it is precisely at the point of application that the language model can be said to infer meaning, namely by bridging the gap between the perceived world and the system’s internal representation of the world.

4.1.1 Position statement

Let us stop and reflect on how humans learn. Comparing humankind to other species, our success is largely due to collective learning: we systematically codify knowledge so that we can store and transfer it in a compact form, relieving individuals from having to learn everything from scratch. If we look farther than our predecessors, it is because we are standing on their shoulders. We teach pupils simple algorithms to multiply numbers rather than expecting them to figure it out themselves, or to memorise each product of two numbers as a separate fact. We also teach them how to read, in order to update themselves with new facts from newspapers or the Internet. To mimic this to a certain degree, and thus overcome deficiencies such as those listed in (Bender et al., 2021a; Bender & Koller, 2020a), we propose that machine learning systems in general and language learning systems in particular should be equipped with similar abilities. In support of this proposal, we would like to point out:

1. Merely using more data seems to us an, if not outright impossible, then at least an inefficient, means of producing systems with human-like faculties of reasoning.
2. In contrast to, e.g., recent works that attempt to use word embeddings as knowledge bases (Bouraoui et al., 2020; Petroni et al., 2019b), we believe it is more effective to realise the faculties of perception, reasoning, and memory as separate computational unities. In practice, this means integrating neural networks with external knowledge bases and inference engines.
3. The ability to assess new data points in the light of accepted knowledge and, if appropriate, disregard them rather than incorporating them into the model, can make systems more robust against the imperfections of training data, and make the learned model more coherent.

4.1.2 Related work

There is a rapidly growing body of literature on hybrid machine-learning systems (see e.g. (Hohenecker & Lukasiewicz, 2020b; van Bekkum et al., 2021; P.-W. Wang et al., 2019b)). We discern two main lines of work. The first combines perception and inference, in other words, deep learning and rule-based or neuro-symbolical reasoning. A good example is AlphaGo (Silver et al., 2017) which

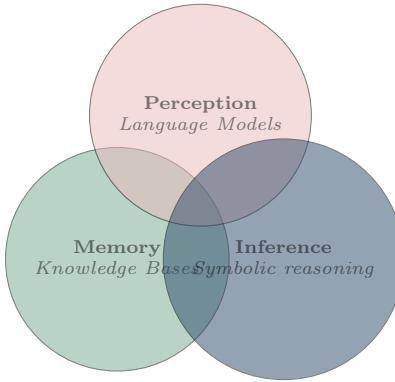


Figure 4.1: The triad of capabilities and components advocated.

augments Monte Carlo search trees with deep neural networks. An example from computer vision is DeepProbLog (Manhaeve, Dumancic, Kimmig, Demeester, & Raedt, 2018) which separates perception from inference. More precisely, it employs a simple convolutional neural network to identify digits from MNIST, supported by probabilistic logic programming for modelling and reasoning.

A survey of perception-inference hybrid systems is given by (L. D. Raedt et al., 2020) which address neuro-symbolical and statistical relational approaches to integrating learning and reasoning. The authors provide examples that leverage the strength of both methods, such as (Ellis et al., 2018). They also identify open challenges, e.g., leveraging the effectiveness of deep learning for symbolical representation learning.

The second line of work combines perception and explicit memory, that is, deep learning and knowledge bases. Lecue investigates the role knowledge graphs have in explainable AI (Lecue, 2019) and explains how knowledge graphs can be integrated with deep neural networks to aid explainability, to bootstrap natural language models, and to disambiguate between word senses when uncertainty arises. Knowledge graphs are also studied in the context of neural network architectures, see for example the recent review of graph neural networks (J. Zhou et al., 2020). Reasoning over knowledge bases is explored in e.g. (Hohenecker & Lukasiewicz, 2020b; Minervini et al., 2020; Qu et al., 2021a), and is reviewed in (X. Chen, Jia, et al., 2020).

Another set of writings treat knowledge bases in a multimodal framework. Multimodality generally means that information is drawn from a heterogenous source of data, the most studied combination being language and visual data. Here, knowledge bases help connect the modalities (Kannan et al., 2020; Krishna et al., 2017b; Pezeshkpour et al., 2018; Zhu et al., 2015). For purely visual data, it is known that knowledge graphs can aid machine vision tasks (Marino et al., 2017), and also that unimodal language models stand to benefit from the addition of knowledge bases (Petroni et al., 2019b). How other modalities and how the construction of data sets limits what a model can learn is outlined

in (Bisk et al., 2020).

Covariate shift (Sugiyama & Kawanabe, 2012; Sugiyama et al., 2007) and concept drift (Gama et al., 2014; Lu et al., 2018) are related to the problem of disregarding data points using existing knowledge. A difference between training and test distributions is a covariate shift, and (Schneider et al., 2020) improves model robustness using covariate shift adaptation. Shifting over, e.g., time is called concept drift, and (Webb et al., 2016) characterises such drifts.

4.1.3 Key challenges

Although a more diverse set of tools can be expected to have advantages as opposed to a total reliance on deep learning or neuro-symbolic methods, it also leads to new challenges:

- Which overall system architecture is needed to enable a seamless integration of perception, memory, and inference subsystems as indicated in Figure 4.1?
- Can we design learning strategies that validate new facts using the knowledge base and inference system, and integrate new data into the model only if it is consistent with accepted knowledge and values?
- Can the language model and the knowledge base evolve over time, and how do we keep them aligned as language changes and the meaning of words start to drift?
- Can we integrate the inference engine and the language model to enable heuristic search and inference that make use of the capabilities of the language model?

Given the richness and diversity of solutions that can be attained by fusing perception, explicit memory, and inference, we believe that the fields of machine learning and neuro-symbolic inference can draw on the strengths of each other to gain a truer understanding of meaning.

4.2 Using synthetic data

One issue with large language models of recent times is that they perform poorly on the tail end distribution of tasks, and rely on enormous amounts of data mined from the internet and other sources (Bender et al., 2021b). The mined data itself perpetuates bias, such as systemic racism, and relies on the law of large numbers to sufficiently cover the underlying distribution of language and knowledge. Recent examples include PaLM(Chowdhery et al., 2022), where the authors train on a highly gender-skewed dataset (masculine pronouns are mentioned 5.5 times more often than feminine). Synthetic datasets allow us to control this distribution and generate examples which highlights a specific edge case or behaviour such as compositional generalisation. The downside is that it

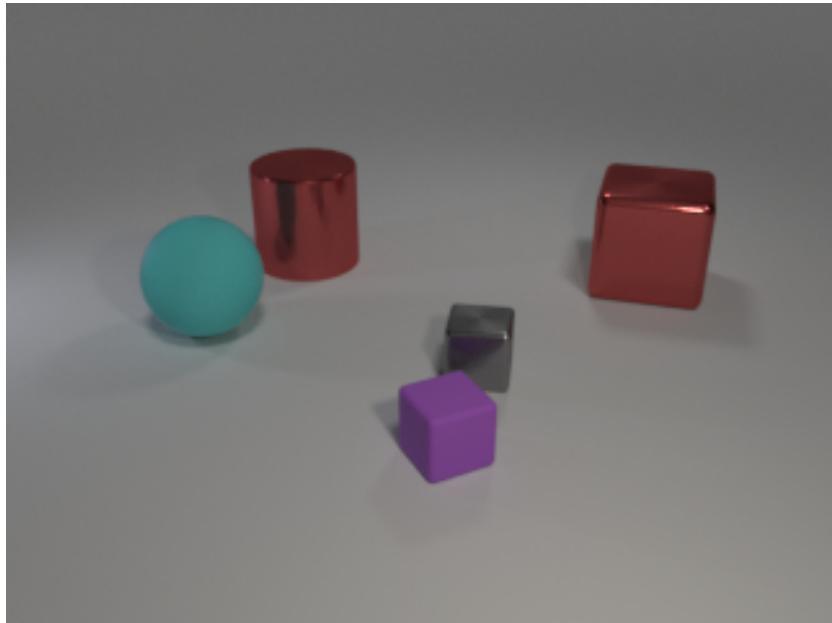


Figure 4.2: Example image generated via Blender using CLEVR. One question generated to this image using the one hop template is "*What is the color of the metal cylinder that is behind the cyan matte thing?*" to which the answer is *red*.

is difficult to generate the diversity that large internet based datasets capture. Therefore, synthetic datasets are a good complement to real world datasets to control for specific expected behaviour or properties. As such, it is suitable for our purposes of examining reasoning and compositional generalisation.

One system for generating such datasets is CLEVR (Johnson et al., 2017), which uses the 3D modelling software *Blender* to generate images of 3D scenes with objects of different size, shape, and material. From the internal representation of the 3D scene, questions on the contents of the images are generated from a set of templates and a predefined functional language. The questions are formed to challenge the reader to reason about the different shapes, materials, and colors of the objects, and their relation in space. Figure 4.2 shows an example of an image and a question-answer pair generated using CLEVR. One great strength with synthetic data is that it gives great control over the distribution of instances. Since the publication of CLEVR in 2017, several extensions, continuations, and modifications have been introduced (Arras et al., 2022; Kottur et al., 2019; Z. Li et al., 2022; R. Liu et al., 2019; Salewski et al., 2022; Sampat et al., 2021; Stammer et al., 2021). Stammer et al. introduce experiments with confounding information in CLEVR-Hans, where properties such as color is used to confound a learning algorithm. To illustrate the concept of confounding information, B. Kim et al.; Rieger et al. use ColorMNIST where

each digit is associated with a fixed color during training but randomised in testing. Without any inductive bias, it is uncertain how a learning algorithm should distinguish between the color and the digit, as they are always seen together. A learning algorithm can still use both information channels to define the concept compositionally like $digit_1 = \text{color}(blue, o) \text{ and } \text{shape}(1, o)$. However, given only a few examples in training where a digit is not associated with a fixed color, a learning algorithm should be able to separate these two features. There should be a metric for measuring the sensitivity to compositionality. Having an IID dataset as a requirement for learning is an impossible requirement for any real data. Therefore, low sensitivity to compositionality is an important component in reliable learning algorithms. An argument against only using deep learning methods is that they pick up on signals that are not interesting to the task. TODO horse example, wolf example. Conversely, neuro-symbolic reasoning methods are designed to handle these situations better than pure neural networks.

4.3 DeepProbLog and compositionality

We will now look at an example of how confounding information can fool a neuro-symbolic method. DeepProbLog (Manhaeve, Dumančić, et al., 2018) is a neuro-symbolic model for learning and reasoning with neural networks. We use confounding colors with ColorMNIST (B. Kim et al., 2019; Rieger et al., 2020) to test how DeepProbLog generalises compositionality. This way, a model that picks up on the color as important will fail on the test set. The hypothesis is that the neural component in NeSy models will have better guidance via the symbolic information that is put into the model. DeepProbLog has not been tested on such data to the best of our knowledge, only on the regular MNIST dataset (Manhaeve, Dumančić, et al., 2018). Here we compare how DeepProbLog performs on ColorMNIST versus the regular MNIST dataset, and outline the differences.

We train and evaluate DeepProbLog on MNIST under three setups. The original experiment on addition with MNIST digits reported on in (Manhaeve, Dumančić, et al., 2018) is used as a baseline. The same experiment but with colored digits from ColorMNIST is used in two settings; consistent colors between training and testing (ColorMNIST_{same}), and fixed colors during training and randomised colors during testing (ColorMNIST_{diff}). The experiments on colored images uses the NeSyXIL model to handle color. Both experiments are run with the same hyperparameters (Adam, learning rate 1e-03), for the same number of epochs. The experiments build on code from (Stammer et al., 2021)¹ and (Manhaeve, Dumančić, et al., 2018)². Table 4.1 shows the accuracy of learning to do addition using MNIST digits using DeepProbLog with and without confounding colors. Figure 4.3 shows the confusion matrix with grayscale MNIST digits. Figure 4.4 shows the confusion matrix when color is

¹<https://github.com/ml-research/NeSyXIL>

²<https://github.com/ML-KULeuven/deepproblog>

	MNIST	ColorMNIST _{same}	ColorMNIST _{diff}
DeepProbLog	0.7798	0.4498	0.1178
TODO Neural			

Table 4.1: Accuracy on evaluation data for both

Figure 4.3: Confusion matrix for DeepProbLog on MNIST

consistent between training and testing, suggesting that the extra dimensions of color leads to a more difficult problem. Training the model differently is necessary to rule out factors related to model convergence. Figure 4.5 shows

Figure 4.4: Confusion matrix for DeepProbLog on ColorMNIST using the same color maps for training and testing.

the confusion matrix when color is confounding between training and testing, leading to almost random performance.

These results suggests that DeepProbLog handle confounding colors in MNIST poorly. This contradicts the supposed strong compositionality of neuro-symbolic methods. Further experiments are necessary to identify why we see this behaviour, but this insight can be used to guide the design of future methods and datasets.

One philosophical issue is that for the model has no way of knowing whether we want it to learn the shapes or the colors of digits. During training, both color and shape directly map to the number, meaning there is no information available to the model to realise that the color is not important. In a sense, our biased expectations on the model are not aligned with what our data suggests. This shows that even for small toy datasets it is difficult to create learning tasks

	Action																		
	10	5	13	7	4	12	3	8	14	15	2	9	11	6	17	0	1	16	18
Predicted	10	47	18	1	52	0	16	0	74	0	8	12	2	0	36	0	0	0	0
	5	58	56	75	77	12	64	16	138	12	21	0	0	132	0	0	0	0	0
	13	3	0	0	0	0	0	0	0	0	0	0	0	0	30	1	0	0	0
	7	36	18	79	142	64	22	85	35	57	52	8	64	21	0	0	0	0	0
	4	41	0	0	0	17	40	50	0	27	68	92	1	45	46	54	0	0	77
	12	43	0	0	19	0	0	23	0	0	0	0	0	2	2	1	0	0	0
	3	89	16	29	42	0	19	0	28	1	0	0	97	0	86	0	0	0	0
	8	37	62	59	37	34	60	19	73	53	0	34	0	0	102	0	0	38	0
	14	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	71	21	42	44	3	0	2	31	22	0	122	0	58	0	0	0	0
	11	1	0	0	1	0	0	0	0	2	1	0	0	40	54	46	0	0	0
	6	44	48	15	0	76	64	97	64	41	0	52	0	66	55	0	53	63	0
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	5	5	0	0	0	8	0	0	11	7
	1	13	0	0	0	0	0	1	0	0	0	0	0	29	19	0	0	0	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Accuracy: 0.1178
Accuracy: 0.1178
Accuracy: 0.1178

Figure 4.5: Confusion matrix for DeepProbLog on ColorMNIST using different color maps for training and testing.

that are not ambiguous.

4.4 Multimodal Word Math Problems

Now that we have outlined how we can use synthetic data from CLEVR to build multimodal benchmarks, we will use word math problems as the basis for tasks. Word math problems are suitable because solving them implies reasoning over concepts and their relations in language. Consider the following word math problem,

Adam has three apples, and Eve has five. Eve gives Adam all her apples. How many apples does Adam have, if he eats one?

For a system to answer this question, it must reason in multiple steps, as well as translate verbs into mathematical operations. Small changes in the text will also lead to large semantic changes, e.g. changing *eats* to *finds*. An arbitrary number of sentences with actions also require compositional generalisation (X. Chen, Liang, et al., 2020; Keysers et al., 2020; N. Kim & Linzen, 2020; B. M. Lake, 2019; Saqr & Narasimhan, 2020; Shaw et al., 2020), the capability to "generate infinite use of finite means" (Chomsky, 2014). Word math problems are a great setting for benchmarking systems on their generalisability in the intersection of natural language, reasoning, and vision. Previous work have mostly explored word math problems in a text only setting, like the problem shown above, using neural networks (Robaidek et al., 2018; Sundaram et al., 2020; Sundaram & Khemani, 2015), and other methods (Mitra & Baral, 2016; Sundaram & Abraham, 2018). Math Word Problem Solving (MAWPS) (Koncel-Kedziorski et al., 2016) was one of the earlier datasets introduced in the domain and collected around 3320 single/multi equation word problems involving operators +, -, *, /. These word problems were annotated with equations involved and the answer (solution of the equation). More recently, larger datasets like Algebra Question Answering with Rationales (AQuA-RAT) (Ling et al., 2017) were introduced and it has around 100K multiple choice questions annotated with equations and a textual explanation for the rationale behind the equations. (A. Patel et

al., 2021) illustrated the deficiencies in MAWPS dataset by introducing another dataset named Simple Variations on Arithmetic Math word Problems (SVAMP). SVAMP is created by making minor variations to problems in MAWPS (A. Patel et al., 2021) showed that state-of-the-art neural solvers trained on MAWPS performs poorly on the SVAMP dataset. See (Huang et al., 2016) for an overview of how to construct word math problems.

With word math problems as the basis, we introduce a multimodal word math problem dataset with images and corresponding mathematical tasks. We generate 3D scenes using CLEVR (Johnson et al., 2017), creating CLEVR-math. As discussed earlier, previous work such as CLEVR-Hans (Stammer et al., 2021) uses CLEVR in similar fashions to generate 3D data to examine specific behaviour. Figure 4.6 shows an example from this dataset. One important

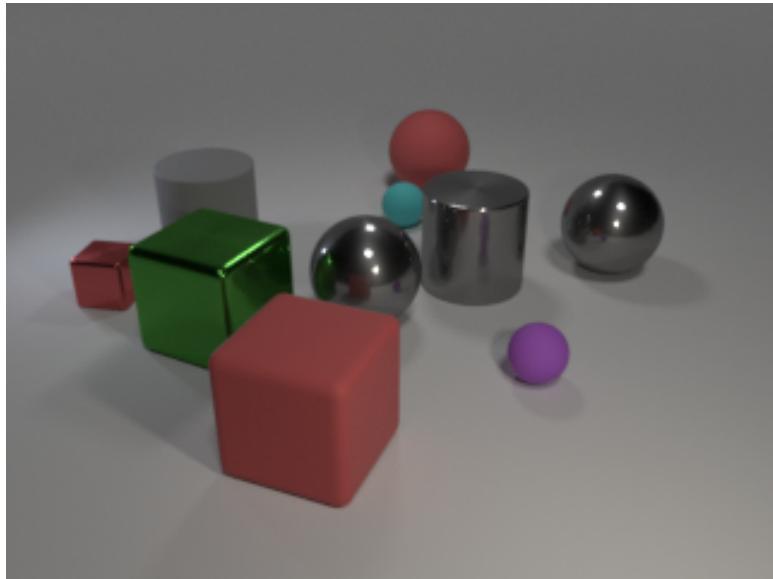


Figure 4.6: Selection of questions generated from this image: (i) *Remove all gray spheres. How many spheres are there? (3)*, (ii) *Take away 3 cubes. How many objects are there? (7)*, (iii) *How many blocks must be removed to get 1 block? (2)*

aspect of the human mind is that we envision changes without them physically manifest (TODO CITE). Imagining changes and inferring the consequences is a big part of how we reason. CLEVR-Math pose problems that tests a models ability to imagine such change, since the tasks involve chains of operations and queries over the resulting (internal) state. Mathematical operators are well suited to build compositional generalisation tasks since the operators work recursively over any novel combination of mathematical elements

Therefore, testing the effect of compositionality on the mathematical reasoning provides a well-defined but broad domain where compositionality is a core

principle. This could be tested by fixing the domain to a single color, shape, material, and size, and ask the same questions again. Additionally, including an increasing amount of confounding information is a related experiment. This could be done by always presenting blue cubes at training time, and randomising at testing. Important to measure is how the degree of confounding information affects the model. It should be enough to have only a couple of non-confounding examples (e.g. 5 out of 100) to realise that cubes are not always blue and should be disentangled as a concept. The hypothesis is that neural networks would need something close to equal distribution over the combinations and properties in order to achieve good performance.

When considering benchmarks for grounding, reasoning, or compositionality, it is important to not only measure behaviour but also internal structures of systems. Given enough capacity, a neural network would be able to memorise everything necessary to perform reasonably well. Similarly, a symbolic-driven system could introduce a new symbol for each combination of concepts or properties of objects. A *blue ball* can be encoded as a single concept, with *red ball* being encoded completely independently. Now, if we only test for behaviour, then this is difficult to uncover using the standard approaches in the deep learning community. One way of getting around this is by probing the internal structures to uncover desiderata. One such instance is by using confounding information, e.g. by always associating a shape with a specific color. There are multiple ways in which this can be investigated. For shapes and colors, it is possible to fix one shape-color pair while letting the other shapes and colors combine randomly. If a model is able to categorise shapes and colors to be part of the same concept categories, then it should be able to realise that spheres can be other colors than blue since the shape category does not have a fixed relationship to the color category. A slightly different experiment is to always associate shapes with specific colors, and measure the impact of introducing a few samples where this is not the case. For instance, if spheres are always blue, but one sample contains a red sphere, that might not be enough evidence that spheres are not synonymous with blue objects. However, once there have been 10 such samples, there is more reliable evidence that this is not the case. Given different systems, the ratio between fixed and randomised shapes and colors can reveal a lot about its inner workings. A desiderata is to have as small a supporting set as possible to break up such fixed relationships. The hypothesis is that neural networks perform poorly under such conditions.

On the topic of grounding, we can construct pseudoword instances of groupings of already known concepts and properties. This allows us to focus the evaluation on the underlying mechanism rather than whether a model might have seen something resembling the instance during training.

4.5 NOT REWRITTEN = CLEVR-Math

Solving mathematical word problems requires one to be able to map the natural language text to a mathematical expression, identifying the known and unknown

quantities and the operators to be used. Again, we consider the following math word problem,

Problem: *Adam has three apples, and Eve has five. Eve gives Adam all her apples. How many apples does Adam have, if he eats one?*

Equation: $X = 3 + 5 - 1$

Minor changes in the text may result in large semantic changes, e.g. changing just one word in the above problem - *eats* to *finds*, will change the equation to $X = 3 + 5 + 1$. Most of the recent efforts in automatic math word problem solving treat it as a translation task (from word problem to equation) and have employed sequence-to-sequence networks or sequence to tree (generating the expression tree of the equation) networks ((Luong et al., 2015), (Z. Xie & Sun, 2019),(J. Zhang et al., 2020)).

While text-based math word problems are a great setting for natural language understanding, it would also be interesting to consider word problems which are accompanied by a diagram, and the information required to derive the solution has to be captured from both its textual and visual representations. That is, part of the problem scenario description is expressed as text and the other part is represented in the form of an image. We introduce such a multi-modal math word problem dataset, CLEVR-Math (since it is based on CLEVR dataset (Johnson et al., 2017)), where each problem has a textual and a visual description (image). Based on the strengths of using synthetic data previously discussed, CLEVR-Math allows us to test the ability of systems to generalise to unseen combinations of actions in the word problem. For example, we can train on single mathematical operations, and test on chains of operations. Figure 4.7 shows a sample problem in CLEVR-Math.

While each instance in CLEVR dataset has an image and a natural language query about the scene depicted in the image, in CLEVR-Math, the natural language query may not be about the scene represented in the image, but about the state of the scene after/before a sequence of actions are applied on the scene. The actions in our case are addition/removal of specific type of objects to/from the original scene. We believe this is an interesting problem setting as the ability to envision changes without them being physically manifested is an important aspect of the human mind.

Our contributions are two-fold, we

- construct an open source multi-modal math word problem dataset, CLEVR-Math and
- analyse the performance of state-of-the-art neural and neuro-symbolic (NeSy) solutions for solving such multi-modal problems.

Our results and analysis shows how both neural and NeSy methods are unable to compositionally generalise to chains of operations.

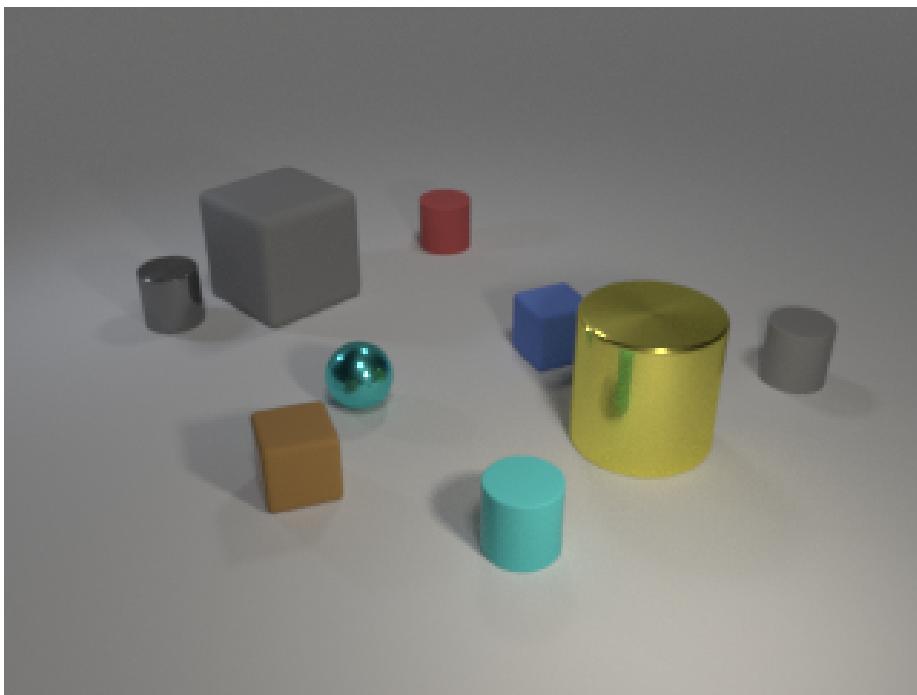


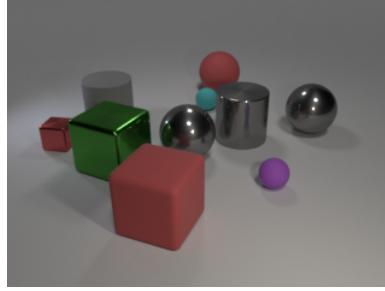
Figure 4.7: CLEVR-Math example question *Take away 2 matte cylinders. How many objects are left?* with corresponding mathematical equation $X = 9 - 2$.

4.5.1 Constructing CLEVR-Math

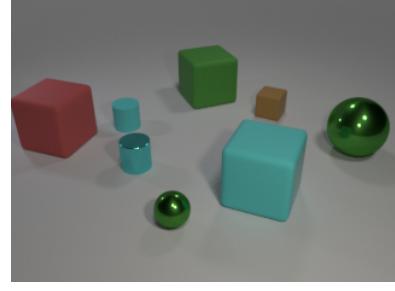
We construct the CLEVR-Math dataset as an extension of CLEVR by introducing three new functions and 13 templates. Using the codebase provided with CLEVR, we generate new questions based on the original scenes. We categorise the 13 templates into six types, all based on addition and subtraction. The domain is restricted to numbers between 0 – 10 to conform with CLEVR.

New CLEVR functions:

The three functions that we implement are - `subtraction` and `addition` to perform subtraction and addition, and `choose` to operate on subsets of objects. Instead of removing all blue spheres, `choose` allows us to remove a random number of a specific type of object, e.g. 2 blue spheres out of 4. The random number generated by `choose` replaces a questions "X" placeholder during generation. Figure 4.8a shows three examples of subtraction, and Figure 4.8b shows a question requiring multihop reasoning. Appendix 4.5.1 includes more samples from the test set.



- (a) (i) Remove all gray spheres. How many spheres are there? (3), (ii) Take away 3 cubes. How many objects are there? (7), (iii) How many blocks must be removed to get 1 block? (2)



- (b) Take away all large green metallic spheres. Now remove all cyan objects. How many objects are left? (4)

Figure 4.8: Example image-question pairs from CLEVR-Math, 4.8a showcase addition and subtraction, and 4.8b shows multihop reasoning. Answers in parenthesis.

Question Categories:

The different question categories are shown in Table 4.2.

- **Remove group:** All objects belonging to a specific group are removed from the scene.
- **Insertion:** A specific number of objects are added to the scene.

- **Count backwards:** The query is about the change - that is the number of objects added/removed from the scene to get a goal state.
- **Remove subset:** A specific number of objects are removed from the scene.
- **Adversarial questions:** These are trick questions where the actions may be performed on one object, but the query is about an object that is not affected by the action. The adversarial actions are always on objects that are seen in the image.
- **Multi-hop:** In contrast to the above questions, multi-hop questions perform sequences of actions (insertion, removal) on the objects. Such questions with chained functions help us test a model's ability to generalise to infinite combinations of operations.

Type	Templates
Remove group	"Remove all <C> <S>s. How many <S>s are there?" "Take away all <Z> <C> <M> <S>s. How many <S>s are there?" "Take away X <C> <S>s. How many objects are there?" "Take away all <C> <S>s. How many objects are there?"
Insertion	"Add X <Z> <C> <M> <S>s. How many <Z> <C> <M> <S>s are here?" "Add X <Z> <C> <M> <S>s. How many objects are there?"
Count backwards	"How many <C> <S>s must be removed to get X <C> <S>s?" "Take away <C> <S>s. How many were removed if there are X <C> <S>s left?"
Multi-hop	"Take away all <Z> <C> <M> <S>s. Remove all <Z2> <C2> <M2> <S2>s. How many objects are left?"
Remove subset	"Remove X <S>s. How many <S>s are there?"
Adversarial questions	"Remove all <C1> <S1>s. Remove all <C2> <S2>s. How many <S1>s are left?" "Remove all <C1> <S1>s. How many <C2> <S2>s are left?"

Table 4.2: An overview of the different templates implemented by CLEVR-Math. $\langle Z \rangle$, $\langle C \rangle$, $\langle M \rangle$, $\langle S \rangle$ are instantiated to size, color, material, and shape during the question generation.

Each problem in the dataset is also annotated with it's equivalent functional program based on the CLEVR functions described in the previous section. For example, consider the question from insertion category and it's program (the arguments of an instruction refer to another instruction - indicating it's input is the output of the referred instruction):

Q: *Add 3 blue cylinders. How many cylinders are there?*

```
Program: 1. scene, 2. choose[3], 3. count(1),
4. filter_cylinder(1), 5. count(4), addition(2, 5)
```

The program contains the `choose` function - `choose[i]` operator returns i ($i = 3$ in this case).

Question generation.

To support greater linguistic variation, we add synonyms for addition and subtraction to the template engine. *Subtract* can be replaced with *remove*, *take away* and *withdraw*, and *addition* with *introduce*, and *insert*. We use the same training and validation scenes as CLEVR, and generate 5000 new scenes as test data. Figure 4.9 show the distribution of attributes, words, templates and answers in CLEVR-Math, aggregated over the training, validation, and test data. The distribution is reflected in each of the splits.

There are 50 words in the CLEVR-Math vocabulary, where the narrow language puts focus on the mathematical reasoning rather than advanced language capabilities. Figure 4.9c show that most questions are 8-9 words long, with a second peak at 13 for the multihop questions.

Template	Train	Validation	Test
Subtraction	229364	49149	3281
Addition	193641	41600	2752
Adversarial	65180	13900	950
Multihop	67897	14553	972

Table 4.3: Distribution of templates in each data split.

Table 4.3 shows the distribution of templates, with an approximately equal amount of questions for subtraction and addition, and similarly for adversarial and multihop questions. The ratios are consistent between splits. To test multihop reasoning and compositional generalisability we generate train-validation-test with only singlehop questions in training and validation, and only multihop questions in the test data. Thus, a model using the `CLEVR-Math-multihop` configuration must solve the multihop questions in a zero-shot fashion.

Examples

Figure 4.10c show some examples of questions from different templates in CLEVR-Math.

Open sourcing data.

We open source CLEVR-Math as a Huggingface dataset ³ with two configurations; `CLEVR-Math` and `CLEVR-Math-multihop`. The extended CLEVR source code is available on Github ⁴. Table 4.4 shows the Huggingface dataset card for

³<https://huggingface.co/datasets/dali-does/clevr-math>

⁴<https://github.com/dali-does/clevr-math>

CLEVR-Math. The template feature allows for filtering to perform, e.g., only singlehop training and multihop testing.

Feature	Type	Example
template	String	subtraction-multihop
id	String	CLEVR_math_test_000010.png
question	String	<i>Remove 5 spheres. How many objects are there?</i>
image	image path	CLEVR_v1.0/images/train/CLEVR_new_000010.png
label	int64, 0-10	5

Table 4.4: Huggingface dataset card for CLEVR-Math.

4.5.2 Experiments

CLIP (Radford et al., 2021) is used as a neural baseline. Questions and images are embedded using CLIP, and an additional classification layer is added to predict the correct answer. Fine tuning CLIP on CLEVR-Math as a masked language task before adding classification gave no significant improvements, while consuming significantly more computational resources. CLIP and this classification layer is trained jointly for 10 epochs with early stopping using a batch size of 64.

NS-VQA (Yi et al., 2018) is used as the neuro-symbolic baseline. Here, a mask-RCNN (He et al., 2017) is trained independently to convert an image to a scene graph. In our experiments, we skip this step and use the actual scene graphs associated with images. The question is parsed into a functional program by a sequence to sequence (Seq2Seq) network based on Bi-LSTM. A quasi-symbolic program executor executes the program generated on the scene graph of the image to return an answer. The Seq2Seq network is pre-trained in a fully supervised fashion by providing it a few examples (around 60 examples) of (*question, program*) pairs. The pre-trained network is then trained further using REINFORCE algorithm that returns a reward based on whether the program generated could derive the expected answer or not. Supervised pretraining and REINFORCE were run for 1000 and 5000 iterations, respectively, with a batch size of 128. Both CLIP and NS-VQA models were trained on a NVIDIA Tesla P100 GPU computing processor.

Each model is evaluated on each question category, and are trained on 2500, 5000, 10000, and 20000 samples to see the influence of the amount of data. For multihop, training and validation sets with and without multihop questions are used, with the latter named *multihop (0-shot)*.

4.5.3 Results

Table 4.5 shows the accuracy of CLIP and NS-VQA on the different categories as well as an aggregated accuracy over the entire dataset. Both the models were trained on 10,000 samples. NS-VQA performs better than CLIP models for

most templates apart from multihop. NS-VQA performs better on subtraction and adversarial problems (both based on ‘subtraction’ CLEVR function) than addition problems. This could be because the functional programs for addition problems always contain a `choose` operator. It is important to identify the argument to `choose` operator from the problem statement (which is mostly one of the numerical quantities in the word problem) to arrive at the correct answer. Unlike this, there are subtraction and adversarial problems (in remove group) that do not have a `choose` operator in the program. Neither of the

Model	All	Addition	Subtraction	Adversarial	Multihop	Multihop (0-shot)
NS-VQA	0.8840	0.9781	0.9948	0.9957	0.286	0.267
CLIP	0.3464	0.5699	0.3019	0.2848	0.272	0.238

Table 4.5: Accuracy on the CLEVR-Math dataset, shown for each template group and aggregated over all templates.

methods perform well on the multi-hop questions, with a clear degradation in the performance for NS-VQA. This is because the question parser of NS-VQA relies on a Seq2Seq network that does not generalize compositionally (B. Lake & Baroni, 2018a). CLEVR focus on visual attribute compositionality, and the multi-hop reasoning introduces higher demands on linguistic compositionality. When multihop questions are included in the training and validation data, naturally both methods improve their performance.

To gain further insight into CLIPs’ performance on CLEVR-Math, Appendix 4.5.6 shows a confusion matrix from training CLIP on 20 000 samples and evaluating on all question categories. These results show that most errors made by CLIP is off by ones. This reflects the generative nature of such models, in how they can get the context correct but sometimes miss out on details. We also see how CLIP focus on learning in the range 1-5, reflecting that these problems represent a majority of the problems.

Model	2500	5000	10000	20000
NS-VQA	0.6283	0.8840	0.6795	0.6118
CLIP	0.2918	0.3184	0.3528	0.3464

Table 4.6: Accuracy over all templates for different dataset sizes.

Table 4.6 shows how different training sizes influence the accuracy. We can see that NS-VQA achieves high accuracy from relatively few examples and plateaus, which is consistent with the original results on CLEVR. It also seems like NS-VQA is overfitting with more data given, and one hypothesis is that more emphasise is put on the program, but that they are similar enough to confound NS-VQA. In CLEVR, the different questions were more distinguishable from a program perspective. CLIP scales with the number of samples, but plateaus at a much lower accuracy. We note that a larger number of samples could lead to similar performance for CLIP, but at the cost of more computational resources.

We randomly sample 20 correct and 20 incorrect answers from the multihop test data for both CLIP and NS-VQA. Appendix 4.5.5 contains a subset of those samples, and Figure 4.11 illustrates two incorrect answers. There are no clear patterns of failures, such as only performing one of the actions, but we notice multiple instances where CLIP fails to perform overlapping subtraction, or subtraction when no objects match the description. Another observation is that half of the 20 incorrect answers from CLIP, where on images with only three objects. Scenes with few objects have a much smaller possible action space associated to it, meaning that there is less room for error. In Figure 4.11a, there are no purple matte blocks to remove, so the corresponding equation is $3 - 0 - 1 = 2$.

4.5.4 Conclusions

We introduced a new dataset, CLEVR-Math, containing word math problems about visual scenes. Our results show that the state-of-the-art NeSy model, NS-VQA, achieves higher accuracy on CLEVR-Math with less data and computational resources, than the neural model, CLIP. This is further evidence that neural methods, such as CLIP, are lacking in reasoning capabilities, even after fine tuning. Given that NS-VQA uses perfect scene graphs, the comparison is not completely fair. We still expect the results of learning end-to-end to be consistent with the current results in alignment with the original results on CLEVR for NS-VQA.

CLEVR-Math successfully introduces a focused benchmark for learning and reasoning in multimodal data. There are a few natural extensions to this work, both on further development of the dataset and on evaluation. Extending the benchmark to answers outside of the range 0-10 would provide a more challenging domain, and providing scene graphs for each step of the reasoning chain could open up for other methods. The empirical results show that neither of the models could generalize to chained actions. Hence, it is also of research interest to design neuro-symbolic models where language perception is tackled in a more generalizable manner. Focus should lie on the representations (symbols) that are learned. Other interesting directions is to introduce a representation that is manipulated internally according to the actions as they are read. Adding longer chains of operations, or chains with alternating subtraction and addition, would put even more emphasise on the reasoning capabilities. Finally, there is an opportunity to add confounding information to test the robustness, e.g. by associating each shape with a fixed color during training and randomise it during testing.

4.5.5 Incorrect answers on multihop questions

CLIP

Figure 4.12 shows random samples of when CLIP fails to answer multihop questions correctly.

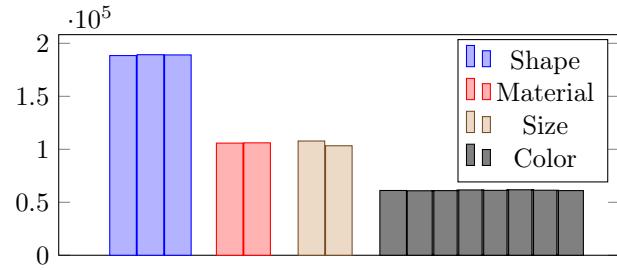
NS-VQA

Figure 4.13 shows random samples of when NS-VQA fails to answer multihop questions correctly.

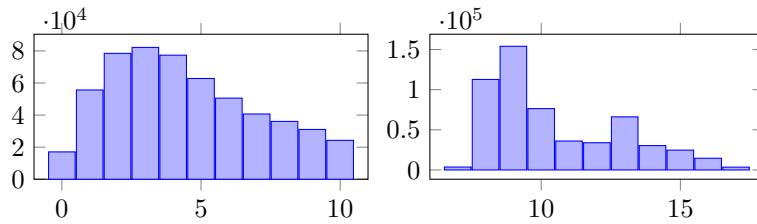
4.5.6 CLIP confusion matrix

Figure 4.14 shows a confusion matrix indicating that CLIP is learning something for all labels. It also shows that when an answer is wrong, it is off by one. The confusion matrix also reflects the distribution over answers, showing that most answers are considered by CLIP to lie in the range 1-5.

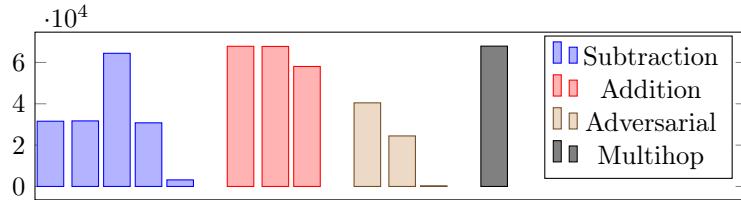
4.6 Challenges and characteristics



(a) Attribute distribution per category, showing even allocations.

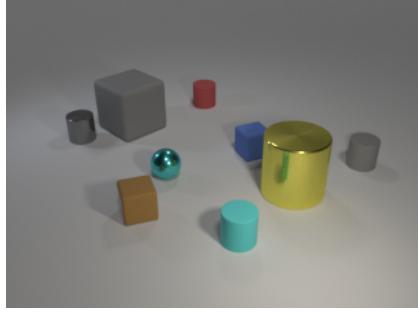


(b) Answer distribution, from 0 to 10.
(c) Distribution of number of words.

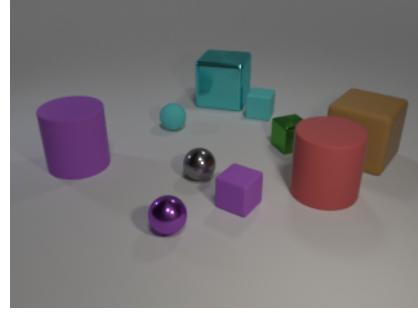


(d) Template distribution over categories of templates. Each bar corresponds to a template in each respective category. We see that subset subtraction (i.e., *remove 2 blue cubes*) is underrepresented.

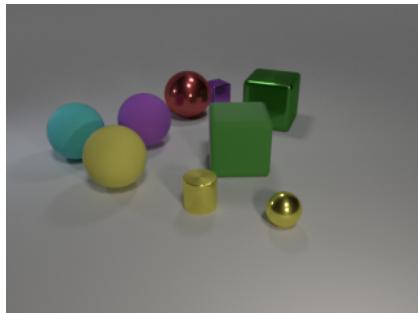
Figure 4.9: The attributes are used evenly throughout the dataset, whereas the answers are biased towards the smaller numbers. The numbers are aggregated over all splits.



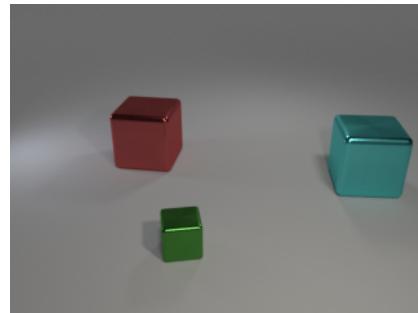
(a) Subtract all gray cylinders. Subtract all gray cubes. How many cylinders are left? (3)



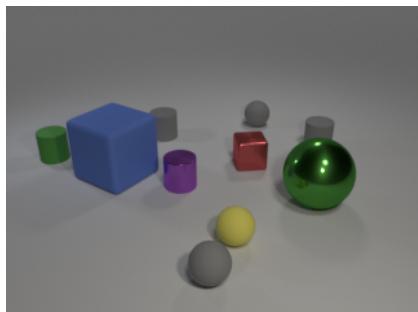
(b) How many cyan cubes must be subtracted to get 1 cyan cubes? (1)



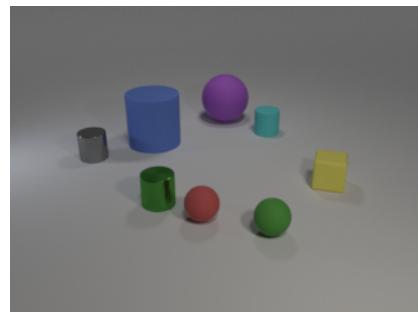
(c) Subtract all purple cylinders. Subtract all yellow blocks. How many cylinders are left? (1)



(d) Add 2 large cubes. How many objects exist? (5)

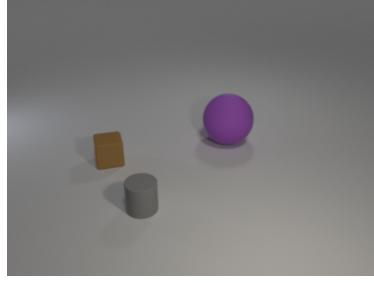


(e) Subtract 4 cylinders. How many cylinders are left? (0)



(f) Subtract 0 brown blocks. How many objects are left (8)

Figure 4.10: Samples from the test set of CLEVR-Math.



(a) *Subtract all small purple matte blocks. Subtract all blocks. How many objects are left?* was answered by CLIP with 3 instead of 2.



(b) *Subtract all red metallic objects. Subtract all yellow objects. How many objects are left?* was answered with 9 instead of 5 by NS-VQA.

Figure 4.11: Examples of when CLIP and NS-VQA fails on multihop questions.

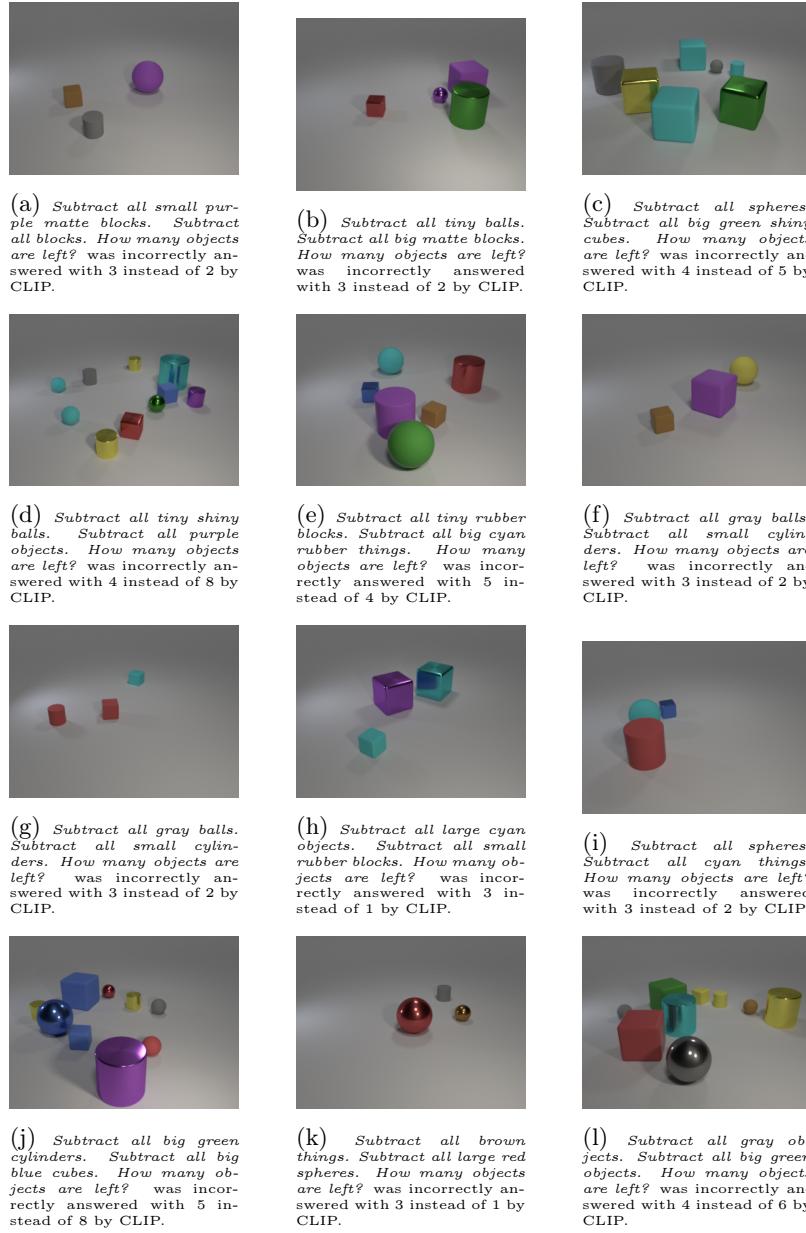


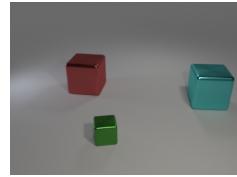
Figure 4.12: Sampling of incorrect answers by CLIP on multihop.



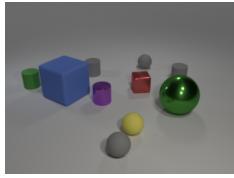
(a) Subtract all large purple objects. Remove all green metallic objects. How many objects are left? was incorrectly answered with 7 instead of 8 by NS-VQA.



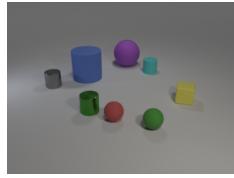
(b) Subtract all red metallic objects. Subtract all yellow objects. How many objects are left? was incorrectly answered with 9 instead of 5 by NS-VQA.



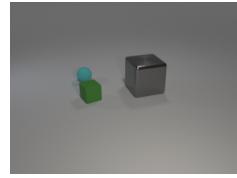
(c) Subtract all tiny green metallic cubes. Subtract all large brown blocks. How many objects are left? was incorrectly answered with 0 instead of 2 by NS-VQA.



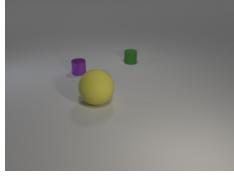
(d) Subtract all small red objects. Subtract all tiny metal cylinders. How many objects are left? was incorrectly answered with 6 instead of 8 by NS-VQA.



(e) Subtract all cylinders. Subtract all purple objects. How many objects are left? was incorrectly answered with 7 instead of 3 by NS-VQA.



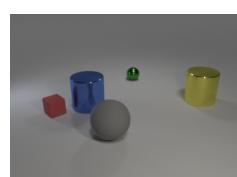
(f) Subtract all blue metal cylinders. Subtract all gray objects. How many objects are left? was incorrectly answered with 1 instead of 2 by NS-VQA.



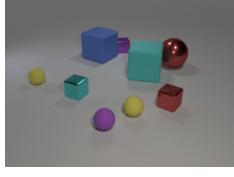
(g) Subtract all large rubber spheres. Subtract all blue blocks. How many objects are left? was incorrectly answered with 3 instead of 2 by NS-VQA.



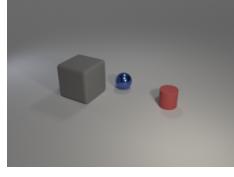
(h) Subtract all big gray blocks. Subtract all large cylinders. How many objects are left? was incorrectly answered with 4 instead of 6 by NS-VQA.



(i) Subtract all tiny blocks. Subtract all small balls. How many objects are left? was incorrectly answered with 4 instead of 3 by NS-VQA.



(j) Subtract all small purple blocks. Subtract all matte objects. How many objects are left? was incorrectly answered with 7 instead of 3 by NS-VQA.



(k) Subtract all tiny red rubber objects. Subtract all small blue balls. How many objects are left? was incorrectly answered with 2 instead of 1 by NS-VQA.



(l) Subtract all cyan cylinders. Subtract all small rubber objects. How many objects are left? was incorrectly answered with 6 instead of 5 by NS-VQA.

Figure 4.13: Sampling of incorrect answers by NS-VQA on multihop.

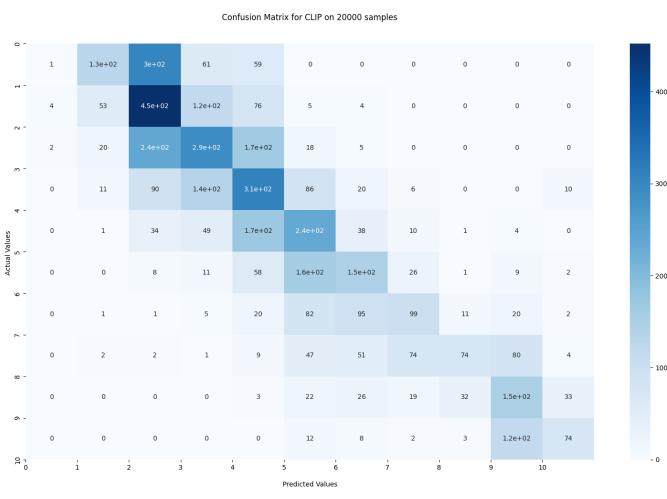


Figure 4.14: Confusion matrix for CLIP trained on 20 000 samples.

Chapter 5

Ontology of Characteristics for multimodal language modelling

Example of how to use quotes at
the beginning of chapters

dali

It is important to explicitly reason about the properties and behaviour we expect out of a system. If we build a model that address the issue of compositional generalisation, but fail to fulfill other critical expectations, the model will be insufficient and flawed by design. We can then reason about the viability and validity of a model by explicitly stating our expectations and reason about them. A trivialising example is a model that perfectly models language, but takes a 100 years to process a sentence making it practically unviable.

Lazaridou et al. gives an account of the pitfalls of static language modeling, and TODO finish

5.1 Ontology of Characteristics

What are the components of an ontology to classify and understand language models?

From Brenden Lake: The characteristics to test for in a benchmark for grounding/multimodal language learning

Key concepts in human-centric AI are communication and collaboration, transparency and explanation, control and adaptivity, and the social dimension.

- Communication and collaboration

- Incremental learning - batch learning will never keep up
- Goal oriented
- Faithfulness
- Definiteness (Lambrecht, 1996)
- Multilinguality
- Concept drift
- Multimodality
- Groundedness
- Specificity
- Gricean speaker
 - * Truthful
 - * Brief
 - * Informative
- Transparency and explanation
 - Access to/control over internals
 - Externalisation of knowledge (e.g. KGs)
- Control and adaptivity
 - Generalise out of distribution
 - Compositionality
 - Amount of examples needed to learn something
 - Externalisation of knowledge (e.g. KGs)
 - Concept drift
- Social dimension
 -
- Technical
 - Scalability
 - Degree of hand-engineering (e.g., domain specific schemas vs. sub-symbolic free-for-all)
 - Explicit vs. implicit mapping to symbols
 - Explicit vs. implicit grounding over modalities (how are observations in e.g. image and text connected)
- Models context (is the model stateful and aware of the current context?)
- Missingness from causality

- Thinking fast and slow

From (J. Chen & Xiao, 2022)

- Diverse
- Descriptive
- Figurative
- Faithful
- Logical
- Controllable
- Knowledgeable
- Explainable
- Fair
- Efficient

Behavior

- Faithfulness
- Truthfulness
 - Tell if something is true or not (truth conditions in formal semantics)
- Groundedness
- Compositionality
- Generalise out of distribution
- Concept drift
- Incremental learning - batch learning will never keep up
- Goal oriented
- Multimodality
- Models context (is the model stateful and aware of the current context?)
- Missingness from causality
- Specificity
- Definiteness (Lambrecht, 1996)

Technical/internal

- Externalisation of knowledge (e.g. KGs)
- Control over internals (transparency and possibility for interventions)
 - Explainable
- Scalability
- Amount of examples needed to learn something
 - Atomic updates vs. gradual adjustment to examples (e.g. knowledge graphs vs. fine-tuning neural networks)
- Degree of hand-engineering (e.g., domain specific schemas vs. sub-symbolic free-for-all)
- Explicit vs. implicit mapping to symbols
- Explicit vs. implicit grounding over modalities (how are observations in e.g. image and text connected)

Brainstormed list of properties

- Causality
- Temporality
- Resolution of observations in domain/modality (object segment vs. image)
- Expressiveness of logic
- Coverage over symbols/concepts (e.g., difficult to construct symbols for every imaginable concept and still have computationally viable logical inference systems, coverage can be a compromise for other properties)

Categories

- Training
- Testing
- Architecture
- Data
- Syntactic information vs. semantic information

Properties of Knowledge Graphs

There are many examples of knowledge graphs that contain multimodal data (Alberts et al., 2021; Hudson & Manning, 2019a; Krishna et al., 2017a). New knowledge can be added with a single operation, and it is possible to extract discrete facts. These two properties are two sides of the same coin, and in a graphical representation of properties they should be connected. We can build a knowledge base from scratch, where four facts are as usable as four million. It is possible to capture temporality and multiple languages (Gottschalk & Demidova, 2018). Knowledge graph structures can be learnt using embeddings (Cai et al., 2018; Q. Wang et al., 2017). However, there are issues with these knowledge graph embeddings, especially when performing link prediction/knowledge base completion (Kadlec et al., 2017; Pezeshkpour et al., 2020; Rossi et al., 2021; Rossi & Matinata, 2020; Y. Wang et al., 2019).

Properties of Transformer embeddings

Word embeddings such as word2vec (Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) showed how language could be represented efficiently in vector spaces while maintaining similarity metrics. They build on the idea that "*you shall know a word by the company it keeps*" (Firth, 1957), and words that are used in the same context are moved closer together in the vector space. One flaw of these methods is that words such as bank or chair, words with multiple distinct meanings, are difficult to disambiguate. Transformers solve this by encoding the context of a word into the embedding, instead of only mapping the word itself onto the vector space. Updating a transformer model usually requires many examples and iteration through the training process, although there are recent studies trying to address this problem (Mitchell et al., 2021). There are multiple studies on how to extract facts directly from transformers such as BERT (Kalo et al., 2020; A. Liu et al., 2019; Yao et al., 2019)

Properties of Logical Representations

Logical frameworks share many properties with knowledge graphs, as the triples in a knowledge graph can be formulated as a logical rule connecting two atoms representing the entities via a relationship. Link prediction in knowledge graphs is an instance of logical inference, to deduce new rules from a set of previous rules (triples). Similarly, logic representations allow for explainability as it is possible to point to explicit rules used to arrive at an answer. Logic systems generally are not scalable to large sets of examples or rules. There are variations, like probabilistic logic programming or fuzzy logic, that relaxes these systems to provide approximate answers (Manhaeve, Dumančić, et al., 2018). There are logical systems that capture temporal aspects. Adding new knowledge to a logical system is an atomic operation, but if consistency is required it involves computation (reference here). Logical systems are commonly used in expert systems, such as medical applications, where expert knowledge is encoded in logical rules. One such example is answer-set programming (Brewka et al.,

2011; Erdem et al., 2016; Lifschitz, 2019; Marek & Truszczyński, 1999; Niemelä, 1999). There are applications of modelling visual contexts using ASP (Khan et al., 2019; Suchan et al., 2018).

5.1.1 Foundational semantics

Go through the following list and outline pros and cons

- Knowledge graphs
- Distributional semantics
- Logical clauses
- Narrative networks

5.2 Chomsky-like hierarchy for grounding and compositionality

The above described case studies should be ordered in a hierarchy/taxonomy/ontology, both as a way to more clearly talk about these related problems, and also device a ladder to climb and construct datasets from. Inspiration can be taken from the Chomsky hierarchy (Chomsky, 1956). Delétang et al. attempts to bridge the Chomsky hierarchy and neural networks. They provide empirical evidence of where to fit RNNs, LSTMs, and transformers into the hierarchy. The authors argue that memory-augmentation is necessary to move up the levels of the hierarchy.

5.3 Framework

This section formalises the previously described properties into an ontology/formal structure.

We can make a distinction between behaviour and functional properties. How do we fit in language grounding and modelling the world into this? Possible to connect this to state, behaviour, and identity from object oriented programming languages.

Procedure for structuring formalism

- Do a rough outline, hierarchy, graph representation of the problem
- Find what is missing in the structure
- How concrete is the formalism, how abstract? Does it have a small footprint (neatly formulated) or a big one (probably too messy)?
- If no clear path from here is visible, move back to writing about the other parts of the thesis

5.4 Ideas

A graph representation with properties and methods where an edge between property and method means the method exhibits that property. This can be used to cluster methods and visualise the interaction between properties and methods.

There is a parallel that can be drawn between the formalisation of properties, and evaluation according to them, and controlled experiments in other natural sciences. In the medical field there is a clear set of steps to create new medication, starting with agar plates and petri dishes, via lab experiments on rodents to human trials. In machine learning, we can place different training schemas, datasets, or evaluation tasks in a similar hierarchy. A highly specific evaluation of how well a model does compositionality in mathematical equations can give useful information in what properties we can assign it. For drug discovery, if the compound does not kill the bacteria in a petri dish, we know that it will not do so in the human body either. However, if it does work on the bacteria we cannot say for certain that it will in human trials, or without side-effects.

Two dependency trees can differ in structure but still represent the same semantic graph (e.g. AMR). In other words we can rearrange the surface form, i.e. the words used, to give it a different structure but still maintain the same deep structure (Chomsky, 1964). This deeper structure conveys the semantics of language. The surface form is governed by syntax such that, e.g., grammatical rules are followed.

Properties can be group according to what level of abstraction they work on. Technical properties, such as surface form, compositionality, or what language a model works on, can be separated from soft properties such as bias. This grouping can be used as a basis for an ontology, and guide the overall formalisation further.

There are many ways to analyse language from a linguistic perspective, and machine learning models from technical perspectives. How do we choose properties to investigate that are useful and not only an intellectualisation of language processing? Is there a hierarchy of importance or impact?

Synthetic data can be useful, but there are examples of when it is problematic. The idealised world it represents might either hide or introduce artificial problems that are not relevant to the real world equivalents. How can this be reflected in a properties framework?

Many of the properties are sliding scales rather than binary, and it is important to be able to described this within the framework. It is also important to be able to study the relations between methods in terms of these properties. E.g., does Method A inherit property X from Method B as it builds on top of that?

Chapter 6

Learning Hierarchical Compositions

Example of how to use quotes at the beginning of chapters

dali

Compositional generalisation in multimodal language models is studied mostly through implicit metrics in unsupervised settings. Recent systems such as CLIP, DALL-E 2, Stable Diffusion, et c., are impressive in the way they process novel combinations of concepts, but it is difficult to verify underlying structures and mechanisms allowing these compositions. One difficulty when analysing these capabilities is that we do not know exactly what data a system was trained on. Therefore, an experimental setup using synthetic data means we can investigate with high control. However, for ecological validity, we also want the data to have enough natural properties to show transferability to natural domains. Previous work does this with abstract 2D concepts and pseudowords (B. M. Lake, 2019; B. M. Lake et al., 2019; Ruis et al., 2020). In this work, we propose a compositional generalisation benchmark in a 3D environment using hierarchical pseudoword concepts. With pseudowords, we can ensure that the specific concept has not been seen before associated with that word. Since vision models can achieve perfect accuracy on the CLEVR dataset, this means that we can assume that the basic properties such as shape and color is already known. The hierarchical aspect means that concepts build on each other, and that we can investigate whether a model learns basic building blocks first before composing more complex concepts. This then means that we can more easily construct curriculum learning setups. This allows us to investigate the impact of relying on such structures.

Pseudoword setups has a long-standing place in linguistics research, most famously with the Wug Test introduced by Berko in 1958. The test involves

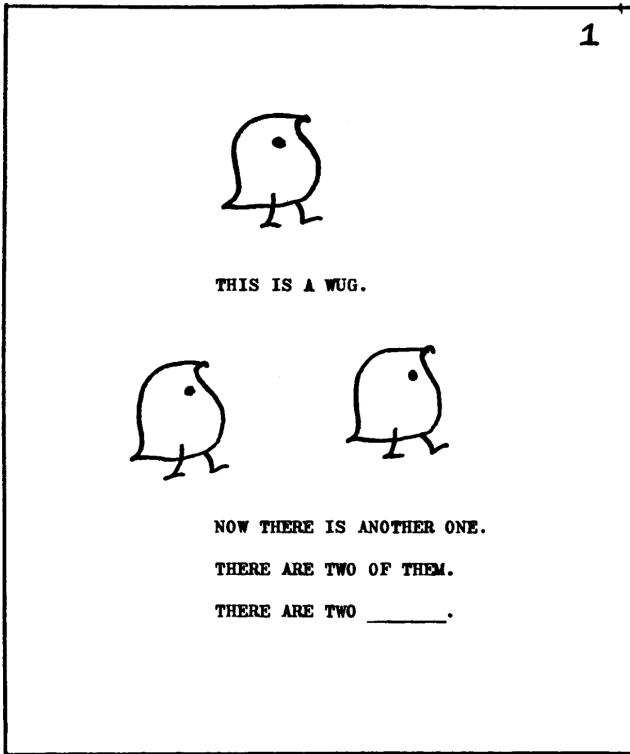


Figure 6.1: Example of card from the Wug Test (Berko, 1958), showing a task of applying morphological rules to a novel (pseudo)word.

27 questions where pseudowords are introduced, and the task is to use it in a novel grammatical role. Each question is posed on a card with an illustration of the pseudoconcept. The example that gave the test its name is seen in Figure 6.1, showing how the word *wug* (denoting a bird) is supposed to be used in plural. Using 56 children age 4–7, the experiments show how the subjects can apply morphological rules to novel words correctly with fairly high degrees of accuracy. Regarding the difficulty of the task, we refer to this statement by the authors; *Answers where willingly, and often insistently, given* (Berko, 1958). One important takeaway from the Wug Test is that humans learn rules that can be applied to novel words in a zero-shot situation, and that we are able to compose previous knowledge to do so.

Carey and Bartlett is another example from developmental psychology, investigating how children learn a single new word. The authors detail the process of acquiring a word by different pieces of information. According to Carey and Bartlett, a learner

- makes a new lexical entry, noting the word and in which language

- learns the syntactic subcategorization, e.g. that it is a verb,
- relate it to other known words through super-, hypo-, and hypernyms
- ground the word in the real world,
- differentiate this concept from previous concepts by e.g. breaking it out as a different species of animal.

In their experiments with 19 children, the subjects were told that *chromium* was the word for the color *olive green*. The procedure involved the following tasks;

- Introduction to the word “chromium”,
- Baseline vocabulary assessment,
- Olive sorting task,
- Olive naming task,
- Chromium comprehension task, and
- “Chromium” hyponym task.

In the sorting task, the children were tasked with using their newly acquired knowledge about “chromium” to solve the physical task of matching colors to boxes among similarly odd colors. With plain red, green, and yellow, it can be expected that the children confused the concept of chromium meaning olive green with it meaning something like *the odd color out*. In the comprehension task, the children were tasked with pointing at three colors, one of which was chromium, controlling whether they had properly learnt a referent for “chromium”. In the hyponym task, the experiment controlled for whether the children had learnt that “chromium” indeed referred to a color. It is important to note here that these tasks cover multiple different aspects of understanding a word, rather than only the textual understanding aspect as in the Wug Test. From their experiments, the authors distinguish between two phases; the *fast mapping* and *drawn out mapping*. Fast mapping takes place in the first few encounters, and gives only a small subset of the information outlined above, such as its language and supernym. A more complete understanding of the word instead requires both more encounters and more time. Their results show that the subjects could use the new word after only one exposure, but that the second encounter was necessary to perform well on the outlined tasks. One takeaway is that we learn a sufficient amount of information about a new word with very little data by utilizing existing understanding, but not enough to understand and use it fully without exposure over longer periods of time. In a sense, it can be expected that we learn certain aspects well at the first encounter, but that more complex notions take more time.

lake2019; B. M. Lake use similar ideas to construct tests for compositional generalisation skills in humans. Their work involves learning words for objects

and functions over objects, constructed as 2D images of colored dots in patterns. While the authors reuse the same approach of using pseudowords that are speakable, they restrict the experiments to the pseudoword domain. In the context of AI, using only pseudowords mitigates the problem of information leaking from the training data. For instance, it is difficult to draw any strong conclusions from performing the Wug Test on GPT-derivatives as this is most likely mentioned many times over in the vast amount of data used during training.

T. Brown et al. show with 6 examples that GPT-3 can acquire new words. This shows how deep learning-based methods can be built to acquire new words, but it does not tell us much about to which extent the new word and associated concept can be understood in relation to existing knowledge. We argue that the only conclusion we can draw is that GPT-3 performs the fast mapping described by Carey and Bartlett.

Similar to the psychology experiments on acquiring a new word by Carey and Bartlett, Eustace performs experiments with learning a complex concept at different hierarchical levels.

What we can learn from the Wug Test, is that we can use new words instantly with little learning, and therefore we should be able to bootstrap to previous knowledge. When designing a benchmark, we can translate the Wug Test to check whether there are internal structures and rules that can be applied to novel words, or if the model relies on something more fuzzy.

From the Chromium experiments, we learn different aspects of a word with different speed, which means that testing should reflect these expectations. For our benchmark, this means that we can expect a pseudoword to be lexically understood, but that things like hyponyms are expected to take longer to learn.

The first hypothesis is that we can expect to see similar behaviour in language models. The second hypothesis is that constructing the training procedure to build on previous knowledge will be beneficial for training times, and that learning syntactic usage should come before more complex tasks.

TODO give formal definition from e.g. (Montague, 1970). TODO define productivity TODO define systematicity TODO rnns are notoriously hard to train, but in theory they should be able to learn anything, therefore we need to construct better ways to train these models

Address the following works

- (B. M. Lake & Murphy, 2021)
- (Ruis & Lake, 2022)
- (Y. Zhou & Lake, 2021)
- (Vong & Lake, 2022)
- (Feinman & Lake, 2020)
- (Ruis et al., 2020)

Figure 6.2: TODO figure where GPT is asked to compute additions.

- (B. M. Lake, 2019)
- (B. M. Lake et al., 2019)
- (B. Lake & Baroni, 2018a)
- (Hupkes et al., 2020b)
- (Hupkes et al., 2023)
- (Valvoda et al., 2022)
- (Bogin et al., 2022)
- (Zhao et al., 2022)
- (Phillips & Wilson, 2016)
- (Aydede, 2010)

B. Lin et al. surveys compositional generalisation in applications. They outline seven different application areas; mathematics, control systems, semantic parsing, image captioning, question answering, automatic translation, and recommendation systems. In this thesis, we focus on the compositionality in mathematics, semantic parsing, and question answering. In *mathematics*, the operators must work compositionally with any novel combination of mathematical elements. As we can see in Figure 6.2, GPT does not exhibit this ability (TODO simple experiment with GPT and addition). The experiments with CLEVR-Math described in Chapter ?? supports the claim that a lack of compositional generalisation severely affects the mathematical capabilities of a model. These insights are echoed by e.g. Y. Lan et al. For *semantic parsing*, compositional generalisation entails generating structured representations of the meaning of a piece of text. We have seen COGS (N. Kim & Linzen, 2020) as an example of a benchmark that tests this. Qiu et al. outlines the limitations of model scaling for compositional generalisation, pointing out that, e.g., fine-tuning has a flat or even negative effect. They find that prompt tuning is the most effective way to improve compositional generalisation. For *image captioning* and *question answering*, compositional generalisation is central to understand and respond to novel combinations of concepts. However, in both cases, current state of the art methods leave big gaps to improve upon. To summarise, in all the mentioned domains B. Lin et al. identifies recent methods leveraging structured representations to improve performance on compositional generalisation benchmarks.

Qiu et al. suggest that *the remaining challenges for gSCAN may not necessarily be related to visual grounding [...]*, and propose an additional task with more complex natural language. While processing more complex language is

a natural extension of the gSCAN dataset, another conclusion is that there is room for a more visually complex dataset.

Z. Wu et al. identifies a set of limitations in gSCAN which they address with ReaSCAN.

Sikarwar et al. extends the work by Qiu et al. with GroCoT, a multimodal transformer model achieving state-of-the-art performance on ReaSCAN. The authors complement their experiments on extened ReaSCAN and GSRR (Qiu et al., 2021) with linear probing classifiers to identify what information the transformer is encoding for each object property. They conclude that their modifications to a multimodal transformer does improve compositional generalisation in the gSCAN domain. Their probing experiments show that identifying the target location is a main challenge for better solving the benchmark.

6.1 Compositional generalisation benchmark using hierarchical pseudoword concepts in CLEVR

TODO naming: CLEVR-composition

Previous work shows us the importance of compositionality for generalisation, and how human intelligence is compositional. We have seen examples of tests for compositional generalisation in language models (e.g. COGS (N. Kim & Linzen, 2020)), and examples for multimodal language models in e.g. (Johnson et al., 2017). However, the benchmarks for multimodal language models have focused mainly on confounding information and n-gram associations (e.g. fixing the color of spheres in training but not testing), rather than complex compositional structures such as those modeled in COGS (N. Kim & Linzen, 2020). This section will detail a compositional generalisation benchmark for hierarchical concepts using some of the ideas from developmental psychology outlined earlier in this chapter. The benchmark is realised with CLEVR using pseudoword concepts that build hierarchically on each other, exemplified in Figure 6.3. In the spirit of the Chromium test by Carey and Bartlett, we device multiple tasks through which the comprehension of these concepts are tested.

- Determining presence or absence of concept in image
- Performing a task – mathematical reasoning
- Hyponym task

These tasks are constructed to cover different aspects of the concepts, not necessarily corresponding to orders of difficulty. However, given the hierarchical ordering of the concepts, we can construct a curriculum learning setup for learning concepts by order in the hierarchy. Curriculum learning has shown to improve generalisability and the convergence rate during training (Bengio et al., 2009; X. Wang et al., 2022). One central challenge of curriculum learning is how to estimate difficulty in order to create a curriculum. In our case, we can use the hierarchical structure to reflect the complexity of a concept. With a curriculum,

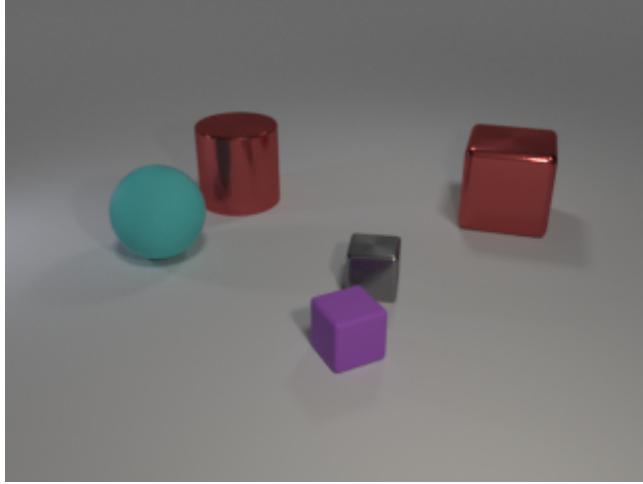


Figure 6.3: A simple example of data generated in CLEVR, where we see two pseudoconcepts; a) a *blargh* – two small cubes next to each other, and b) a *perde* – a large cyan sphere.

we can then compare differences in task performance when training on randomly ordered concepts versus using a curriculum. Beyond performance, we can hypothesise about how a curriculum affects the internal structure of a model to better allow for compositional generalisation.

Another motivation behind this work is TODO CITATION OF IBM, where abstraction, composition, and recursion are three central characteristics necessary to generalise. We can also look to examples from reinforcement learning, where Zhao et al. propose a reinforcement learning method that achieves compositional generalisation in a object oriented domain. The authors borrow ideas from curriculum learning, as they describe three stages (object extraction, action binding, and transition modeling between properties) of learning using their model. However, this work does not address hierarchical compositions but only a flat hierarchy.

Keyser et al. formalise *distribution-based compositionality assessment* (DBCA) as a method to *assess the adequacy* of a dataset split for measuring compositional generalisation. They introduce two guiding principles; 1) similar atom distribution, and 2) different compound distribution, and argue together with Saxton et al. that automated rule-based generation of data brings the control necessary to adhere to these principles. Hence, we build on these ideas when constructing our benchmark.

Another important continuation is to provide tasks with increasing difficulty, either as more steps, higher scene complexities, or more complicated operations to learn. A task that is difficult in general is the introduction of new symbols. From a grounding perspective, it is key that a word is associated with the correct real world referent. From a compositional perspective, it is key that

a new concept is composed of previously known concepts where appropriate. A key component of intelligence is that we can generalise beyond the domain knowledge we have. For AI systems, this means that the domain knowledge encoded at creation might be important inductive bias, but cannot restrict a models' ability to interact and to learn.

Fundamentals

- Mapping n-grams of properties (blue, small) to "natural"* pseudowords[0,4,5] as pseudoconcepts - small blue sphere -> blargh
- Compose a hierarchy of pseudoconcepts
- Hierarchy should reflect (arbitrary) categories
- How is this complementary to gSCAN (Ruis et al., 2020), Tangrams (A. Ji et al., 2022)? Can we use those domains as well/instead?
 - gSCAN talks about “composition of references and another involving composition of attributes”, since a small red square can have referents “square”, “red square”, “small red square”, et c.
- None of the compositional generalisation benchmarks seem to cover hierarchical compositions of concepts, this is a main point of novelty
- Important to have uncertainty in concepts, e.g. by defining conditional properties such as wug=sphere AND (red OR green)

Tasks and experiments

- Did the model learn category abstractions?
- What is the effect of curriculum learning (X. Wang et al., 2022) by acquiring pseudoconcepts from the bottom up (i.e. 1-gram pseudoconcepts) rather than sampled randomly from hierarchy?
 - Do we learn faster/with less?
 - Is the resulting internal structure of the model different? (E.g. investigate using probing experiments of Lovering and Pavlick)
- “Maximize Presupposition”-like setup using, e.g., an image with only one large object: “What color is the large cube?” vs. “What color is the large object?” (Schlenker, 2012)
- Different levels of comprehension - Wug Test (Berko, 1958), Chromium test (Carey & Bartlett, 1978), CLEVR-Math (Lindström & Abraham, 2022)
 - Detecting the presence of an object

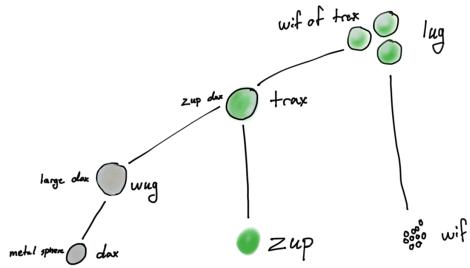


Figure 6.4: Example of a pseudoconcept hierarchy over the CLEVR vocabulary.

- Point out concept in line up
- Use it in a task, e.g. mathematical reasoning or regular CLEVR (Johnson et al., 2017)
- Investigate the effects on comp.gen. when, e.g., gray scaling the image or dimming the lighting, and other visual changes
- Important to do splits across multiple dimensions to test compositionality
 - Recursive depth, do we learn a mechanism/algorithm, or just pattern matching?
 - Properties, shape color et c., do we strongly associate colors and shapes, or do we disentangle them?
- How do we construct a rich enough hierarchy?

The hierarchical constructions are on the following forms

- No parent – base/given knowledge, in this case shape/color et c.
- Naming of an unknown concept, i.e. analogous to the Chromium experiments
- Renaming of a color, e.g. blargh as a synonym for red

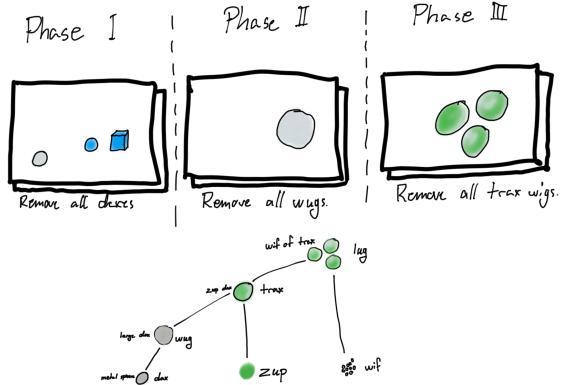


Figure 6.5: Illustration of how a curriculum can be used to investigate the effects of training on compositional generalisation.

- Creating a parent for a 2-gram, e.g. a blue cylinder is called a blik
- Creating a parent for pseudoconcepts, e.g. two small bisks are called a fnik

6.1.1 Different levels of resolution compositionality

Humans adapt our language to context in many ways, as our understanding of concepts are built compositionally. For example, in a certain concept it is only necessary to identify an object by its color, e.g., when instructing which object to pick up. In a different context, it is more important to communicate that the same object is a ball. Or similarly, talking about an animal in a picture, versus talking about the orange house cat. Apart from the context determining the granularity, in many situations a higher-level description carries the same information. If the orange house cat is the only animal in the image, then using *animal* is sufficient. This ability to construct complete understanding of a communication based in a hierarchical composition of knowledge is central to human communication. In these experiments we construct scenes where the language resolution can be varied without affecting the information gain.

- synecdoche
- There's definiteness sort of from pragmatics, which is the term for how clear a referent is. So I guess the example is of two sentences where the referents have the same definiteness. And low definiteness means that we are violating Grice's maxim of quantity. But that's not a term for the sentence pair in the example itself.

One thing that seems relevant is the Maximize Presupposition! (Schlenker, 2012; Singh, 2011) theory in semantics. The theory says that, if we have two

statements p and q that are equally informative in context c , but p makes a stronger presupposition than q , then we should prefer to say q . If we rephrase your example to use statements (p =“The cat is big” vs. q =“The animal is big”), these conditions apply, because:

1. The context c has exactly one cat and exactly one animal, so p/q are equally informative.
2. p presupposes “there is exactly one contextually relevant cat” whereas q presupposes the weaker “there is exactly one contextually relevant animal”.

Thus, the theory would predict saying p is better than saying q .

If we are talking about questions instead of statements, then “informative” is ill-defined, so it is a bit more complicated. Intuitively, it seems more okay to ask “What color is the animal?” than “The animal is big”.

If we consider synthetically generated data in CLEVR, we can construct pairs of sentences p, q for scenes defining the context c . Importantly, p, q only differ in one word, signifying different levels of definiteness but conveying the same information under c . Given that we have a defined scope given by the objects in the scene, we can test whether a system is better at recognising p or q . If there is no difference in how well a system performs on a task (e.g. classification or reasoning), we can draw the conclusion that the system understands the semantic relationship between p and q .

6.1.2 Hierarchical compositional concepts conditioned on context

Concepts are compositional by nature, we build our understanding of things in a hierarchical fashion – a book is made of pages, pages contain text, text is made up of sentences, sentences are made up of words, and so on. However, similar to how words can have different meanings (e.g. river bank vs. financial bank, or ball as a sphere vs. ball as a dance), there is no one true compositional hierarchy for a given concept. The meaning and usage of a concept is highly dependent on the context. For instance, we talk about cars very differently when talking about traveling to work versus when designing them. Therefore, similar to how transformers showed great improvement over distributional semantics by modeling context, there cannot be one hierarchically compositional definition of a concept grounded in previous knowledge but definitions conditioned on some abstract representation of context. For computational reasons, we argue that tree-based hierarchical structures are most useful.

$$\text{concept} = \text{tree_sum}(\text{concepts}) + \text{sum}(\text{sentence}) \quad (6.1)$$

Learning new concept as a composition of previously known ones

Consider the case of learning a pseudoword *blargh* to mean a set of three blue spheres. We can learn to recognise blargh without understanding its components, which is arguably how deep learning methods approach the problem. We

can understand this process as having to look at thousands of examples, and gradually learning the new concept. However, if we can construct our understanding as a composition, we can bootstrap this new concept to already known concepts, given that we can count and that we know exactly what *blue* and *sphere* means. Since we already know how to identify the blue cubes, the learning process should not have to include as many examples and would ideally be close to zero-shot learning.

Mutual exclusivity without resolution

When we see a new word, we assume that it refers to a new concept. How do we situate this new concept in our internal representation?

Mutual exclusivity with resolution to already known concept

How do we ground a new word in the meaning of a known word? Distributional semantics would learn this by similar usage, which makes sense. This is where more traditional methods might not be as strong.

Meta abstraction by consolidating concepts into a category

One powerful mechanism in learning is the ability to abstract concepts into categories, as a way to e.g. understand that similar objects have similar affordances and can therefore be talked about in the same way. There is also the problem of level of resolution, where grouping cars and bikes is useful in one context (talking about traffic planning) but not in others (e.g. as means to travel to another country). This ability to consider categories under different contexts is important, otherwise a representation becomes too static.

TODO Write about IBM work on abstraction, composition, and recursion

Chapter 7

Conclusions

Example of how to use quotes at
the beginning of chapters

dali

Summarise challenges, opportunities, how current work address those, and future work.

Bibliography

- Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR, abs/1608.04207*.
- Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *5th International Conference on Learning Representations, ICLR*.
- Alberts, H., Huang, N., Deshpande, Y., Liu, Y., Cho, K., Vania, C., & Calixto, I. (2021). VisualSem: A high-quality knowledge graph for vision and language. *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 138–152. <https://doi.org/10.18653/v1/2021.mrl-1.13>
- Andreas, J., Baroni, M., Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A., Devlin, J., Fyshe, A., Wehbe, L., et al. (2019). Measuring compositionality in representation learning. *International Conference on Learning Representations*, 375, 2227–2237.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Arras, L., Osman, A., & Samek, W. (2022). Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81, 14–40.
- Aydede, M. (2010). The language of thought hypothesis. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*.
- Badreddine, S., d'Avila Garcez, A., Serafini, L., & Spranger, M. (2022). Logic tensor networks. *Artificial Intelligence*, 303, 103649.
- Badreddine, S., Garcez, A. d., Serafini, L., & Spranger, M. (2021). Logic Tensor Networks [arXiv: 2012.13635]. *arXiv:2012.13635 /cs*. Retrieved October 29, 2021, from <http://arxiv.org/abs/2012.13635>
- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2), 423–443.
- Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 178–186.

- Basu, K., Murugesan, K., Atzeni, M., Kapanipathi, P., Talamadupula, K., Klinger, T., Campbell, M., Sachan, M., & Gupta, G. (2021). A hybrid neuro-symbolic approach for text-based games using inductive logic programming. *Combining Learning and Reasoning: Programming Languages, Formalisms, and Representations*.
- Beinborn, L., Botschen, T., & Gurevych, I. (2018). Multimodal grounding for language processing. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics, COLING 2018* (pp. 2325–2339). Association for Computational Linguistics.
- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021a). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021b). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
- Bender, E. M., & Koller, A. (2020a). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bender, E. M., & Koller, A. (2020b). Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.
- Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48. <https://doi.org/10.1145/1553374.1553380>
- Bennett, M. T. (2022). Symbol emergence and the solutions to any task. In B. Goertzel, M. Iklé, & A. Potapov (Eds.), *Artificial general intelligence* (pp. 30–40). Springer International Publishing.
- Ben-Younes, H., Cadene, R., Cord, M., & Thome, N. (2017). Mutan: Multimodal tucker fusion for visual question answering. *Proceedings of the IEEE international conference on computer vision*, 2612–2620.
- Berko, J. (1958). The child's learning of english morphology. *Word*, 14(2-3), 150–177.
- Berrendorf, M., Faerman, E., Vermue, L., & Tresp, V. (2020). Interpretable and fair comparison of link prediction or entity alignment methods with adjusted mean rank. *CoRR*, abs/2002.06914. <https://arxiv.org/abs/2002.06914>
- Besold, T. R., Garcez, A. d., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K.-U., Lamb, L. C., Lowd, D., Lima, P. M. V., et al. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

- Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. <https://doi.org/10.18653/v1/2020.emnlp-main.703>
- Bogin, B., Gupta, S., & Berant, J. (2022). Unobserved local structures make compositional generalization hard. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2731–2747. <https://aclanthology.org/2022.emnlp-main.175>
- Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29: Annual conference on neural information processing systems, NIPS 2016* (pp. 4349–4357).
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2020). Inducing relational knowledge from BERT. *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7456–7463. <https://aaai.org/ojs/index.php/AAAI/article/view/6242>
- Brewka, G., Eiter, T., & Truszczyński, M. (2011). Answer set programming at a glance. *Communications of the ACM*, 54(12), 92–103.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual*. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>
- Brunet, M., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. S. (2019). Understanding the origins of bias in word embeddings. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning (ICML 2019)* (pp. 803–811).

- Cai, H., Zheng, V. W., & Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9), 1616–1637.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. *Science*, 356(6334), 183–186.
- Carey, S., & Bartlett, E. (1978). Acquiring a single new word.
- Casper, S., Rauker, T., Ho, A., & Hadfield-Menell, D. (2023). Sok: Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. *First IEEE Conference on Secure and Trustworthy Machine Learning*. <https://openreview.net/forum?id=8C5zt-0Utdn>
- Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4427–4442.
- Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., & Bisk, Y. (2021). WebQA: Multihop and Multimodal QA [arXiv: 2109.00590]. *arXiv:2109.00590 /cs*. Retrieved October 29, 2021, from <http://arxiv.org/abs/2109.00590>
- Chaplot, D. S., Sathyendra, K. M., Pasumarthi, R. K., Rajagopal, D., & Salakhutdinov, R. (2018). Gated-attention architectures for task-oriented language grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.
- Chen, J., & Xiao, Y. (2022). Harnessing knowledge and reasoning for human-like natural language generation: A brief review. *arXiv preprint arXiv:2212.03747*.
- Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O. M., Antonyrajah, D., & Horrocks, I. (2021). OWL2Vec*: Embedding of OWL ontologies. *Machine Learning*, 110(7), 1813–1845. <https://doi.org/10.1007/s10994-021-05997-6>
- Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141, 112948. <https://doi.org/https://doi.org/10.1016/j.eswa.2019.112948>
- Chen, X., Liang, C., Yu, A. W., Song, D., & Zhou, D. (2020). Compositional generalization via neural-symbolic stack machines. *arXiv preprint arXiv:2008.06662*.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3), 113–124.
- Chomsky, N. (1964). *Current issues in linguistic theory*. De Gruyter Mouton.
- Chomsky, N. (1975). *The logical structure of linguistic theory*. Springer.
- Chomsky, N. (2014). *Aspects of the theory of syntax* (Vol. 11). MIT press.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, volume 1: Long papers* (pp. 2126–2136).

- Dahlgren, A., Björklund, J., & Drewes, F. (2021). Perception, memory, and inference: The trinity of machine learning. <https://openreview.net/group?id=ijcai.org/IJCAI/2021/Workshop/NSNLI>
- Dai, W.-Z., Muggleton, S. H., & Zhou, Z.-H. (2015). Logical vision: Meta-interpretive learning for simple geometrical concepts. *ILP (Late Breaking Papers)*.
- Dantsin, E. (1992). Probabilistic logic programs and their semantics. In *Logic programming* (pp. 152–164). Springer.
- De Raedt, L., Kimmig, A., & Toivonen, H. (2007). Problog: A probabilistic prolog and its application in link discovery. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2462–2467.
- De Raedt, L., Manhaeve, R., Dumancic, S., Demeester, T., & Kimmig, A. (2019). Neuro-symbolic= neural+ logical+ probabilistic. *NeSy'19@ IJCAI, the 14th International Workshop on Neural-Symbolic Learning and Reasoning*.
- Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Hutter, M., Legg, S., & Ortega, P. A. (2022). Neural networks and the chomsky hierarchy. *arXiv preprint arXiv:2207.02098*.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, volume 1: Long and short papers* (pp. 4171–4186).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Dillon, S. (2020). The eliza effect and its dangers: From demystification to gender critique. *Journal for Cultural Research*, 24(1), 1–15.
- Donadello, I., Serafini, L., & Garcez, A. D. (2017). Logic tensor networks for semantic image interpretation. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1596–1602.
- Dong, H., Mao, J., Lin, T., Wang, C., Li, L., & Zhou, D. (2019). Neural Logic Machines [arXiv: 1904.11694]. *arXiv:1904.11694 [cs, stat]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1904.11694>
- Duch, W., Adamczak, R., & Grąbczewski, K. (1998). Extraction of Logical Rules from Neural Networks. *Neural Processing Letters*, 7(3), 211–219. <https://doi.org/10.1023/A:1009670302979>
- Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173, 43–59.
- Dwivedi, V. P., & Bresson, X. (2020). A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*.

- Elliott, D., & Kádár, Á. (2017). Imagination improves multimodal translation. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 130–141.
- Ellis, K. M., Morales, L. E., Sablé-Meyer, M., Solar Lezama, A., & Tenenbaum, J. B. (2018). Library learning for neurally-guided bayesian program induction.
- Erdem, E., Gelfond, M., & Leone, N. (2016). Applications of answer set programming. *AI Magazine*, 37(3), 53–68.
- Erdem, E., Kuyu, M., Yagcioglu, S., Frank, A., Parcalabescu, L., Plank, B., Babii, A., Turuta, O., Erdem, A., Calixto, I., Lloret, E., Apostol, E.-S., Truica, C.-O., Sandrih, B., Gatt, A., Martincic-Ipsic, S., Berend, G., & Korvel, G. (2022). Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning [to appear]. *Journal of Artificial Intelligence Research (JAIR)*.
- Eustace, B. W. (1969). Learning a complex concept at differing hierarchical levels. *Journal of Educational Psychology*, 60(6p1), 449.
- Evans, J. (1996). Rationality and reasoning. *Cognititive Psychology*.
- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate [PMID: 26172965]. *Perspectives on Psychological Science*, 8(3), 223–241. <https://doi.org/10.1177/1745691612460685>
- Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2018). VSE++: improving visual-semantic embeddings with hard negatives. *British Machine Vision Conference 2018, BMVC 2018*, 12.
- Feinman, R., & Lake, B. M. (2020). Generating new concepts with hybrid neuro-symbolic models.
- Felix, R., Kumar, V. B. G., Reid, I., & Carneiro, G. (2018). Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision*.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.
- Fodor, J. D., Fodor, J. A., & Garrett, M. F. (2013). 12. the psychological unreality of semantic representations. In *Volume ii readings in philosophy of psychology, volume ii* (pp. 238–252). Harvard University Press.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3–71.
- Foucault, M. (1991). *The foucault effect: Studies in governmentality*. University of Chicago Press.
- Frank, S., Bugliarello, E., & Elliott, D. (2021). Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers [arXiv: 2109.04448 version: 1]. *arXiv:2109.04448 [cs]*. Retrieved November 1, 2021, from <http://arxiv.org/abs/2109.04448>
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). DeViSE: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger

- (Eds.), *Advances in neural information processing systems 26: 27th annual conference on neural information processing systems, NIPS 2013* (pp. 2121–2129).
- Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016a). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
- Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016b). Multimodal compact bilinear pooling for visual question answering and visual grounding. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 457–468.
- Gadamer, H. (1975). Hermeneutics and social science, cultural hermeneutics 2: 307–316.
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4), 1–37.
- Garcez, A., Gori, M., Lamb, L., Serafini, L., Spranger, M., & Tran, S. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4), 611–631.
- Garcez, A. d., Bader, S., Bowman, H., Lamb, L. C., de Penning, L., Illuminoo, B., Poon, H., & Gerson Zaverucha, C. (2022). Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342, 1.
- Garcez, A. d., Besold, T. R., De Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K.-U., Lamb, L. C., Miikkulainen, R., & Silver, D. L. (2015). Neural-symbolic learning and reasoning: Contributions and challenges. *2015 AAAI Spring Symposium Series*.
- Garcez, A. S. d., Broda, K., Gabbay, D. M., et al. (2002). *Neural-symbolic learning systems: Foundations and applications*. Springer Science & Business Media.
- Garcez, A. S. d., Broda, K. B., & Gabbay, D. M. (2012). *Neural-symbolic learning systems: Foundations and applications*. Springer Science & Business Media.
- Geffner, H. (2018). Model-free, Model-based, and General Intelligence [arXiv: 1806.02308]. *arXiv:1806.02308 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1806.02308>
- Geffner, H. (2021). Target Languages (vs. Inductive Biases) for Learning to Act and Plan [arXiv: 2109.07195]. *arXiv:2109.07195 [cs]*. Retrieved November 1, 2021, from <http://arxiv.org/abs/2109.07195>
- Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. *Proceedings of the NAACL Student Research Workshop*, 8–15. <https://doi.org/10.18653/v1/N16-2002>
- Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental psychology*, 28(1), 99.

- Gottschalk, S., & Demidova, E. (2018). Eventkg: A multilingual event-centric temporal knowledge graph. *European Semantic Web Conference*, 272–287.
- Goyal, A., Didolkar, A., Ke, N. R., Blundell, C., Beaudoin, P., Heess, N., Mozer, M., & Bengio, Y. (2021). Neural Production Systems [arXiv: 2103.01937]. *arXiv:2103.01937 [cs, stat]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/2103.01937>
- Greff, K., Van Steenkiste, S., & Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*.
- Hamilton, K., Nayak, A., Božić, B., & Longo, L. (2022). Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, (Preprint), 1–42.
- Hammer, B., & Hitzler, P. (2007). *Perspectives of neural-symbolic integration* (Vol. 77). Springer.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335–346.
- Harnad, S. (1993). Problems, problems: The frame problem as a symptom of the symbol grounding problem. *Psychology*, 4(34).
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. corr abs/1703.06870. *arXiv preprint arXiv:1703.06870*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- Henderson, J. (2020). The unstoppable rise of computational linguistics in deep learning. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6294–6306.
- Henderson, L. (2020). The Problem of Induction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University.
- Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743.
- Hinton, G. (2021). How to represent part-whole hierarchies in a neural network. *arXiv preprint arXiv:2102.12627*.
- Hinton, G. E., et al. (1986). Learning distributed representations of concepts. *Proceedings of the eighth annual conference of the cognitive science society*, 1, 12.
- Hitzler, P., Eberhart, A., Ebrahimi, M., Sarker, M. K., & Zhou, L. (2022). Neuro-symbolic approaches in artificial intelligence [nwac035]. *National Science Review*, 9(6). <https://doi.org/10.1093/nsr/nwac035>
- Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books.
- Hohenecker, P., & Lukasiewicz, T. (2020a). Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research (JAIR)*, 68, 503–540. <https://doi.org/10.1613/jair.1.11661>

- Hohenecker, P., & Lukasiewicz, T. (2020b). Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68, 503–540.
- Holzinger, A., Kickmeier-Rust, M., & Müller, H. (2019). Kandinsky patterns as iq-test for machine learning. *International cross-domain conference for machine learning and knowledge extraction*, 1–14.
- Huang, D., Shi, S., Lin, C.-Y., Yin, J., & Ma, W.-Y. (2016). How well do computers solve math word problems? large-scale dataset construction and evaluation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 887–896.
- Hudson, D. A., & Manning, C. D. (2019a). Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.
- Hudson, D. A., & Manning, C. D. (2019b). Learning by Abstraction: The Neural State Machine [arXiv: 1907.03950]. *arXiv:1907.03950 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1907.03950>
- Hudson, D. A., & Manning, C. D. (2019c). Learning by Abstraction: The Neural State Machine [arXiv: 1907.03950]. *arXiv:1907.03950 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1907.03950>
- Hume, D. (1739). *A treatise of human nature*. Oxford University Press.
- Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020a). Compositionality decomposed: How do neural networks generalise? *J. Artif. Intell. Res.*, 67, 757–795.
- Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020b). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67, 757–795.
- Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., et al. (2023). State-of-the-art generalisation research in nlp: A taxonomy and review.
- Ito, T., Klinger, T., Schultz, D. H., Murray, J. D., Cole, M. W., & Rigotti, M. (2022). Compositional generalization through abstract representations in human and artificial neural networks. *arXiv preprint arXiv:2209.07431*.
- Janssen, T. M. (1986). Foundations and applications of montague grammar: Philosophy, framework, computer science.
- Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R. D., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*.
- Ji, J., Krishna, R., Fei-Fei, L., & Niebles, J. C. (2020). Action genome: Actions as compositions of spatio-temporal scene graphs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10236–10247.
- Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8, 423–438. https://doi.org/10.1162/tacl_a_00324
- Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional

- language and elementary visual reasoning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.
- Kadlec, R., Bajgar, O., & Kleindienst, J. (2017). Knowledge base completion: Baselines strike back. *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 69–74. <https://doi.org/10.18653/v1/W17-2609>
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kakas, A. C., & Michael, L. (2020). Abduction and argumentation for explainable machine learning: A position survey. *CoRR, abs/2010.12896*. <https://arxiv.org/abs/2010.12896>
- Kalo, J.-C., Fichtel, L., Ehler, P., & Balke, W.-T. (2020). KnowlyBERT - Hybrid Query Answering over Language Models and Knowledge Graphs. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (pp. 294–310). Springer International Publishing. https://doi.org/10.1007/978-3-030-62419-4_17
- Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, 57(2), 129–191.
- Kannan, A. V., Fradkin, D., Akrotirianakis, I., Kulahcioglu, T., Canedo, A., Roy, A., Yu, S., Malawade, A. V., & Faruque, M. A. A. (2020). Multimodal knowledge graph for deep learning papers and code. In M. d'Aquin, S. Dietze, C. Hauff, E. Curry, & P. Cudré-Mauroux (Eds.), *CIKM '20: The 29th ACM international conference on information and knowledge management, virtual event, ireland, october 19-23, 2020* (pp. 3417–3420). ACM. <https://doi.org/10.1145/3340531.3417439>
- Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137.
- Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
- Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, 4(6), 533–550.
- Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., Tsarkov, D., Wang, X., van Zee, M., & Bousquet, O. (2020). Measuring compositional generalization: A comprehensive method on realistic data. *International Conference on Learning Representations*. <https://openreview.net/forum?id=SygcCnNKwr>
- Khan, A., Bozzato, L., Serafini, L., & Lazzerini, B. (2019). Visual reasoning on complex events in soccer videos using answer set programming. *GCAI*, 42–53.
- Kiela, D., Conneau, A., Jabri, A., & Nickel, M. (2018). Learning visually grounded sentence representations. *Proceedings of NAACL-HLT*, 408–418.
- Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9012–9020.
- Kim, N., & Linzen, T. (2020). Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*.
- Klinger, T., Adjodah, D., Marois, V., Joseph, J., Riemer, M., Pentland, A., & Campbell, M. (2020). A study of compositional generalization in neural models. *arXiv preprint arXiv:2006.09437*.
- Kobayashi, G., Kurabayashi, T., Yokoi, S., & Inui, K. (2020). Attention is not only a weight: Analyzing transformers with vector norms. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7057–7075. <https://doi.org/10.18653/v1/2020.emnlp-main.574>
- Kocijan, V., Lukasiewicz, T., Davis, E., Marcus, G., & Morgenstern, L. (2020). A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.
- Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., & Hajishirzi, H. (2016). Mawps: A math word problem repository. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1152–1157.
- Kottur, S., Moura, J. M., Parikh, D., Batra, D., & Rohrbach, M. (2019). Cleverdialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017a). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1), 32–73.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017b). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1), 32–73.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
- Kruglanski, A. W., & Gigerenzer, G. (2018). Intuitive and deliberate judgments are based on common principles. In *The motivated mind* (pp. 104–128). Routledge.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day american english*. Brown University Press.
- Kuhn, T. S., & Hawkins, D. (1963). The structure of scientific revolutions. *American Journal of Physics*, 31, 554–555.
- Lake, B., & Baroni, M. (2018a). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International conference on machine learning*, 2873–2882.
- Lake, B., & Baroni, M. (2018b). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference*

- on machine learning* (pp. 2873–2882). PMLR. <https://proceedings.mlr.press/v80/lake18a.html>
- Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. *arXiv preprint arXiv:1906.05381*.
- Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. *41st Annual Meeting of the Cognitive Science Society: Creativity+ Cognition+ Computation, CogSci 2019*, 611–617.
- Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological review*.
- Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge university press.
- Lan, Y., Wang, L., Jiang, J., & Lim, E.-P. (2022). Improving compositional generalization in math word problem solving. *arXiv preprint arXiv:2209.01352*.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. <https://openreview.net/forum?id=H1eA7AEtvS>
- Law, M., Russo, A., Broda, K., & Bertino, E. (2021). Scalable Non-observational Predicate Learning in ASP. *2*, 1936–1943. <https://doi.org/10.24963/ijcai.2021/267>
- Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kociský, T., Ruder, S., et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, *34*, 29348–29363.
- Lecue, F. (2019). On the role of knowledge graphs in explainable AI. *Semantic Web*, *11*, 1–11. <https://doi.org/10.3233/SW-190374>
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th Annual International Conference on Systems Documentation*, 24–26.
- Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 552–561.
- Li, T., & Srikumar, V. (2020). Augmenting Neural Networks with First-order Logic [arXiv: 1906.06298]. *arXiv:1906.06298 [cs, stat]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1906.06298>
- Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Van Durme, B., & Yuille, A. (2022). Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. *arXiv preprint arXiv:2212.00259*.
- Liang, P. P., Zadeh, A., & Morency, L.-P. (2022). Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

- Lifschitz, V. (2002). Answer set programming and plan generation. *Artificial Intelligence*, 138(1-2), 39–54.
- Lifschitz, V. (2019). *Answer set programming*. Springer Berlin.
- Lin, B., Bouneffouf, D., & Rish, I. (2023). A survey on compositional generalization in applications. *arXiv preprint arXiv:2302.01067*.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, 740–755.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. J. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Proceedings of the 13th european conference on computer vision, ECCV 2014, part V* (pp. 740–755).
- Lindström, A. D., & Abraham, S. S. (2022). Clevr-math: A dataset for compositional language, visual and mathematical reasoning.
- Ling, W., Yogatama, D., Dyer, C., & Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.
- Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5210–5217.
- Liu, A., Huang, Z., Lu, H., Wang, X., & Yuan, C. (2019). BB-KBQA: BERT-Based Knowledge Base Question Answering. In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese Computational Linguistics* (pp. 81–92). Springer International Publishing.
- Liu, F., Ye, R., Wang, X., & Li, S. (2020). Hal: Improved text-image matching by mitigating visual semantic hubs. *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.
- Liu, R., Liu, C., Bai, Y., & Yuille, A. L. (2019). Clevr-ref+: Diagnosing visual reasoning with referring expressions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liu, S., & Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 730–734.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.
- Lovering, C., & Pavlick, E. (2022). Unit testing for concepts in neural networks. *arXiv preprint arXiv:2208.10244*.
- Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363.

- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.
- Luong, M.-T., Kayser, M., & Manning, C. D. (2015). Deep neural language models for machine translation. *19th CoNLL*, 305–309.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627*.
- Maker, M. H. (2006). Ai@ 50: Ai past, present, future. *Dartmouth College*. http://www.engagingexperience.com/2006/07/ai50_ai_past_pr.html.
- Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.
- Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems*.
- Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & Raedt, L. D. (2018). Deepproblog: Neural probabilistic logic programming. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, neurips 2018, december 3-8, 2018, montréal, canada* (pp. 3753–3763). <https://proceedings.neurips.cc/paper/2018/hash/dc5d637ed5e62c36ecb73b654b05ba2a-Abstract.html>
- Manhaeve, R., Dumančić, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). DeepProbLog: Neural Probabilistic Logic Programming [arXiv: 1805.10872]. *arXiv:1805.10872 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1805.10872>
- Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019a). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.
- Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019b). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. *International Conference on Learning Representations*. <https://openreview.net/forum?id=rJgMlhRctm>
- Marcus, G., & Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Vintage.
- Marcus, G. F. (2003). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.
- Marek, V. W., & Truszczyński, M. (1999). Stable models and an alternative logic programming paradigm. In *The logic programming paradigm* (pp. 375–398). Springer.
- Marino, K., Salakhutdinov, R., & Gupta, A. (2017). The more you know: Using knowledge graphs for image classification. *2017 IEEE Conference on*

- Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 20–28. <https://doi.org/10.1109/CVPR.2017.10>
- McCarthy, J. (1960). Programs with common sense.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115–133.
- Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences.
- Merrill, W., Warstadt, A., & Linzen, T. (2022). Entailment semantics can be extracted from an ideal language model. *arXiv preprint arXiv:2209.12407*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26: 27th annual conference on neural information processing systems, NIPS 2013* (pp. 3111–3119).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013. proceedings of a meeting held december 5-8, 2013, lake tahoe, nevada, united states* (pp. 3111–3119). <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>
- Miller, G. A. (1992). WordNet: A lexical database for English. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. <https://aclanthology.org/H92-1116>
- Minervini, P., Bošnjak, M., Rocktäschel, T., Riedel, S., & Grefenstette, E. (2020). Differentiable reasoning on large knowledge bases and natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), 5182–5190. <https://doi.org/10.1609/aaai.v34i04.5962>
- Mitchell, E., Lin, C., Bosselut, A., Finn, C., & Manning, C. D. (2021). Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.
- Mitra, A., & Baral, C. (2016). Learning to use formulas to solve simple arithmetic problems. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2144–2153.
- Montague, R. (1970). Pragmatics and intensional logic. *Synthese*, 22(1-2), 68–94.
- Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19, 629–679.
- Narasimhan, M., Lazebnik, S., & Schwing, A. (2018). Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31.
- Navigli, R., Blloshmi, R., & Lorenzo, A. C. M. (2022). Babelnet meaning representation: A fully semantic formalism to overcome language barriers.
- Ng, R., & Subrahmanian, V. S. (1992). Probabilistic logic programming. *Information and computation*, 101(2), 150–201.

- Niemelä, I. (1999). Logic programs with stable model semantics as a constraint programming paradigm. *Annals of mathematics and Artificial Intelligence*, 25(3), 241–273.
- Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. In A. T. Ihler & D. Janzing (Eds.), *Proceedings of the 32nd conference on uncertainty in artificial intelligence, UAI 2016* (pp. 557–566).
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, 11(6), 988–1010.
- Parcalabescu, L., Trost, N., & Frank, A. (2021). What is multimodality? *Proceedings of the First Workshop on Multimodal Semantic Representations (MMSR)*. <https://iwcs2021.github.io/proceedings/mmsr/pdf/2021.mmsr-1.1.pdf>
- Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2022). Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena [to appear]. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. <https://arxiv.org/abs/2112.07566>
- Patel, A., Bhattacharya, S., & Goyal, N. (2021). Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.
- Patel, R., & Pavlick, E. (2021). Mapping language models to grounded conceptual spaces. *International Conference on Learning Representations*.
- Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, 8(1), 447–471. <https://doi.org/10.1146/annurev-linguistics-031120-122924>
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60.
- Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. <https://doi.org/10.3115/v1/D14-1162>
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019a). Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473.
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019b). Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. <https://doi.org/10.18653/v1/D19-1250>
- Pezeshkpour, P., Chen, L., & Singh, S. (2018). Embedding multimodal relational data for knowledge base completion. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3208–3218. <https://doi.org/10.18653/v1/D18-1359>

- Pezeshkpour, P., Tian, Y., & Singh, S. (2020). Revisiting evaluation of knowledge base completion models. *Automated Knowledge Base Construction*.
- Phillips, S., & Wilson, W. H. (2016). Systematicity and a categorical theory of cognitive architecture: Universal construction in context. *Frontiers in psychology*, 7, 1139.
- Poerner, N., Waltinger, U., & Schütze, H. (2019). BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, 101–108.
- Qiu, L., Hu, H., Zhang, B., Shaw, P., & Sha, F. (2021). Systematic generalization on gscan: What is nearly solved and what is next? *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2180–2188.
- Qiu, L., Shaw, P., Pasupat, P., Nowak, P., Linzen, T., Sha, F., & Toutanova, K. (2022). Improving compositional generalization with latent structure and data augmentation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4341–4362. <https://doi.org/10.18653/v1/2022.naacl-main.323>
- Qu, M., Chen, J., Xhonneux, L.-P., Bengio, Y., & Tang, J. (2021a). Rnnlogic: Learning logic rules for reasoning on knowledge graphs. *International Conference on Learning Representations*.
- Qu, M., Chen, J., Xhonneux, L.-P., Bengio, Y., & Tang, J. (2021b). RNN-Logic: Learning Logic Rules for Reasoning on Knowledge Graphs [arXiv: 2010.04029]. *arXiv:2010.04029 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/2010.04029>
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *International conference on machine learning*, 4218–4227.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision [arXiv: 2103.00020]. *arXiv:2103.00020 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/2103.00020>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019a). *Language models are unsupervised multitask learners* (tech. rep.). OpenAI.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019b). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Raedt, L. D., Dumancic, S., Manhaeve, R., & Marra, G. (2020). From statistical relational to neuro-symbolic artificial intelligence. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on*

- artificial intelligence, IJCAI 2020* (pp. 4943–4950). ijcai.org. <https://doi.org/10.24963/ijcai.2020/688>
- Raedt, L. d., Dumančić, S., Manhaeve, R., & Marra, G. (2020). From statistical relational to neuro-symbolic artificial intelligence [Survey track]. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 4943–4950). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2020/688>
- Raedt, L. D., & Kersting, K. (2008). Probabilistic inductive logic programming. In *Probabilistic inductive logic programming* (pp. 1–27). Springer.
- Rehurek, R., & Sojka, P. (2011). Gensim—python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
- Reiter, R. (1981). On closed world data bases. In *Readings in artificial intelligence* (pp. 119–140). Elsevier.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144.
- Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., Qian, H., Fagin, R., Barahona, F., Sharma, U., et al. (2020). Logical neural networks. *arXiv preprint arXiv:2006.13155*.
- Rieger, L., Singh, C., Murdoch, W., & Yu, B. (2020). Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 8116–8126). PMLR. <https://proceedings.mlr.press/v119/rieger20a.html>
- Robaidek, B., Koncel-Kedziorski, R., & Hajishirzi, H. (2018). Data-driven methods for solving algebra word problems. *arXiv preprint arXiv:1804.10718*.
- Rogers, A., Hosur Ananthakrishna, S., & Rumshisky, A. (2018). What's in your embedding, and how it predicts task performance. *Proceedings of the 27th International Conference on Computational Linguistics*, 2690–2703.
- Rogers, C. R., & Carmichael, L. (1942). Counseling and psychotherapy: Newer concepts in practice.
- Rosenbloom, P. S. (2010). Combining procedural and declarative knowledge in a graphical architecture. *Proceedings of the 10th International Conference on Cognitive Modeling*, 205–210.
- Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2), 1–49.
- Rossi, A., & Matinata, A. (2020). Knowledge graph embeddings: Are relation-learning models learning relations? *EDBT/ICDT Workshops*.
- Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., & Lake, B. M. (2020). A benchmark for systematic generalization in grounded language un-

- derstanding. *Advances in neural information processing systems*, 33, 19861–19872.
- Ruis, L., & Lake, B. (2022). Improving systematic generalization through modularity and augmentation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).
- Salewski, L., Koepke, A. S., Lensch, H. P., & Akata, Z. (2022). Clevr-x: A visual reasoning dataset for natural language explanations. *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, 69–88.
- Sampat, S. K., Kumar, A., Yang, Y., & Baral, C. (2021). CLEVR_HYP: A challenge dataset and baselines for visual question answering with hypothetical actions over images. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3692–3709. <https://www.aclweb.org/anthology/2021.naacl-main.289>
- Saqur, R., & Narasimhan, K. (2020). Multimodal graph networks for compositional generalization in visual question answering. *Advances in Neural Information Processing Systems*.
- Sarker, M. K., Zhou, L., Eberhart, A., & Hitzler, P. (2021). Neuro-symbolic artificial intelligence. *AI Communications*, 34(3), 197–209.
- Saxton, D., Grefenstette, E., Hill, F., & Kohli, P. (n.d.). Analysing mathematical reasoning abilities of neural models. *International Conference on Learning Representations*.
- Schlag, I., Smolensky, P., Fernandez, R., Jojic, N., Schmidhuber, J., & Gao, J. (2019). Enhancing the transformer with explicit relational encoding for math problem solving. *arXiv preprint arXiv:1910.06611*.
- Schlenker, P. (2012). Maximize presupposition and gricean reasoning. *Natural language semantics*, 20, 391–429.
- Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., & Bethge, M. (2020). Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417–424.
- Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American*, 262(1), 25–31.
- Sen, P., de Carvalho, B. W., Riegel, R., & Gray, A. (2022). Neuro-symbolic inductive logic programming with logical neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8), 8212–8219.
- Serafini, L., & Garcez, A. d. (2016). Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv:1606.04422*.
- Shaw, P., Chang, M.-W., Pasupat, P., & Toutanova, K. (2020). Compositional generalization and natural language variation: Can a semantic parsing approach handle both? *arXiv preprint arXiv:2010.12725*.

- Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., & Bernardi, R. (2017). FOIL it! Find one mismatch between image and language caption. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, 255–265.
- Shen, F., Zhou, X., Yu, J., Yang, Y., Liu, L., & Shen, H. T. (2019). Scalable zero-shot learning via binary visual-semantic embeddings. *IEEE Transactions on Image Processing*, 28(7), 3662–3674.
- Shi, H., Mao, J., Xiao, T., Jiang, Y., & Sun, J. (2018). Learning visually-grounded semantics from contrastive adversarial samples. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics, COLING 2018* (pp. 3715–3727). Association for Computational Linguistics.
- Shi, X., Padhi, I., & Knight, K. (2016). Does string-based neural MT learn source syntax? *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1526–1534.
- Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4613–4621.
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International Conference on Machine Learning*, 3145–3153.
- Sikarwar, A., Patel, A., & Goyal, N. (2022). When can transformers ground and compose: Insights from compositional generalization benchmarks. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 648–669. <https://aclanthology.org/2022.emnlp-main.41>
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550. <http://dx.doi.org/10.1038/nature24270>
- Singh, R. (2011). Maximize presupposition! and local contexts. *Natural Language Semantics*, 19, 149–168.
- Sinha, K., Sodhani, S., Dong, J., Pineau, J., & Hamilton, W. L. (2019). CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text [arXiv: 1908.06177]. *arXiv:1908.06177 [cs, stat]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1908.06177>
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2019). How can we fool lime and shap? adversarial attacks on post hoc explanation methods. *CoRR, abs/1911.02508*. <http://arxiv.org/abs/1911.02508>
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1), 3.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, 11(1), 1–23.

- Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). Neurocompositionality: From the central paradox of cognition to a new generation of ai systems. *arXiv preprint arXiv:2205.01128*.
- Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. Y. (2013). Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, 935–943.
- Sorokin, D., & Gurevych, I. (2017). Context-aware representations for knowledge base relation extraction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1784–1789. <https://doi.org/10.18653/v1/D17-1188>
- Stammer, W., Schramowski, P., & Kersting, K. (2021). Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3619–3629.
- Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, 23(5), 645–665.
- STEELS, L., VERHEYEN, L., & VAN TRIJP, R. (2022). An experiment in measuring understanding. *Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence (HHAI)*. IOS Press, Amsterdam.
- Suchan, J., Bhatt, M., Wałega, P., & Schultz, C. (2018). Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Sugiyama, M., & Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.
- Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).
- Sun, R., & Bookman, L. A. (1994). Computational architectures integrating neural and symbolic processes: A perspective on the state of the art.
- Sundaram, S. S., & Abraham, S. S. (2018). Solving simple arithmetic word problems precisely with schemas. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 542–547.
- Sundaram, S. S., Deepak, P., & Abraham, S. S. (2020). Distributed representations for arithmetic word problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 9000–9007.
- Sundaram, S. S., & Khemani, D. (2015). Natural language processing for solving simple word problems. *Proceedings of the 12th International Conference on Natural Language Processing*, 394–402.
- Szabó, Z. G. (2012). The case for compositionality.

- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for Fact Extraction and VERification [arXiv: 1803.05355]. *arXiv:1803.05355 [cs]*. Retrieved November 3, 2021, from <http://arxiv.org/abs/1803.05355>
- Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. *arXiv preprint arXiv:2204.03162*.
- Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., & Zadeh, A. (2022). Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77, 149–171.
- Valvoda, J., Saphra, N., Rawski, J., Williams, A., & Cotterell, R. (2022). Benchmarking compositionality with formal languages. *Proceedings of the 29th International Conference on Computational Linguistics*, 6007–6018.
- van Bekkum, M., de Boer, M., van Harmelen, F., Meyer-Vitali, A., & ten Teije, A. (2021). Modular design patterns for hybrid learning and reasoning systems: A taxonomy, patterns and use cases. <https://arxiv.org/abs/2102.11965>
- Varghese, D., Bauer, R., Baxter-Beard, D., Muggleton, S., & Tamaddoni-Nezhad, A. (2021). Human-like rule learning from images using one-shot hypothesis derivation. *International Conference on Inductive Logic Programming*, 234–250.
- Vong, W. K., & Lake, B. M. (2022). Cross-situational word learning with multimodal neural networks. *Cognitive science*, 46(4), e13122.
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 6(2).
- Wald, J., Dhamo, H., Navab, N., & Tombari, F. (2020). Learning 3d semantic scene graphs from 3d indoor reconstructions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970.
- Wang, P., Wu, Q., Shen, C., van den Hengel, A., & Dick, A. (2017). Fvqa: Fact-based visual question answering.
- Wang, P.-W., Donti, P., Wilder, B., & Kolter, Z. (2019a). SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. *International Conference on Machine Learning*, 6545–6554. Retrieved October 29, 2021, from <https://proceedings.mlr.press/v97/wang19e.html>
- Wang, P.-W., Donti, P., Wilder, B., & Kolter, Z. (2019b). SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 6545–6554). PMLR. <https://proceedings.mlr.press/v97/wang19e.html>
- Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724–2743.

- Wang, X., Chen, Y., & Zhu, W. (2022). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 4555–4576. <https://doi.org/10.1109/TPAMI.2021.3069908>
- Wang, Y., Ruffinelli, D., Gemulla, R., Broscheit, S., & Meilicke, C. (2019). On evaluating embedding models for knowledge base completion. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 104–112. <https://doi.org/10.18653/v1/W19-4313>
- Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4), 964–994.
- Weber, L., Minervini, P., Münchmeyer, J., Leser, U., & Rocktäschel, T. (2019). Nlprolog: Reasoning with weak unification for question answering in natural language. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6151–6161.
- Weiss, G., Goldberg, Y., & Yahav, E. (2021). Thinking like transformers. *International Conference on Machine Learning*, 11080–11090.
- Weissenhorn, P., Donatelli, L., & Koller, A. (2022). Compositional generalization with a broad-coverage semantic parser. *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, 44–54.
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- Werning, M. E., Hinzen, W. E., & Machery, E. E. (2012). *The oxford handbook of compositionality*. Oxford University Press.
- Winters, T., Marra, G., Manhaeve, R., & De Raedt, L. (2021). Deepstochlog: Neural stochastic logic programming. *arXiv preprint arXiv:2106.12574*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019a). Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019b). Huggingface’s transformers: State-of-the-art natural language processing.
- Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., & Ma, W.-Y. (2019). Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 6609–6618.
- Wu, Z., Kreiss, E., Ong, D., & Potts, C. (2021). ReaSCAN: Compositional Reasoning in Language Grounding. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/2838023a778dfaecd212708f721b788-Paper-round1.pdf>

- Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., & Liu, H. (2021). Graph learning: A survey. *IEEE Transactions on Artificial Intelligence*, 2(2), 109–127. <https://doi.org/10.1109/TAI.2021.3076021>
- Xie, S., Morcos, A., Zhu, S.-C., & Vedantam, R. (2022). Coat: Measuring object compositionality in emergent representations. *International Conference on Machine Learning*, 24388–24413.
- Xie, Z., & Sun, S. (2019). A goal-driven tree-structured neural model for math word problems. *IJCAI*, 5299–5305.
- Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., & Schütze, H. (2019). Probing for semantic classes: Diagnosing the meaning content of word embeddings. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, volume 1: Long papers* (pp. 5740–5753).
- Yao, L., Mao, C., & Luo, Y. (2019). Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.
- Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., & Tenenbaum, J. B. (2020). CLEVRER: CoLlision Events for Video REpresentation and Reasoning [arXiv: 1910.01442]. *arXiv:1910.01442 [cs]*. Retrieved October 29, 2021, from <http://arxiv.org/abs/1910.01442>
- Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.
- Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4), 3208–3216.
- Zhang, B., Hu, H., Qiu, L., Shaw, P., & Sha, F. (2021). Visually Grounded Concept Composition [arXiv: 2109.14115]. *arXiv:2109.14115 [cs]*. Retrieved November 3, 2021, from <http://arxiv.org/abs/2109.14115>
- Zhang, J., Wang, L., Lee, R. K.-W., Bin, Y., Wang, Y., Shao, J., & Lim, E.-P. (2020). Graph-to-tree learning for solving math word problems.
- Zhao, L., Kong, L., Walters, R., & Wong, L. L. (2022). Toward compositional generalization in object-oriented world modeling. *International Conference on Machine Learning*, 26841–26864.
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1, 57–81. <https://doi.org/https://doi.org/10.1016/j.aiopen.2021.01.001>
- Zhou, Y., & Lake, B. (2021). Flexible compositional learning of structured visual concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Zhu, Y., Zhang, C., Ré, C., & Fei-Fei, L. (2015). Building a large-scale multi-modal knowledge base for visual question answering. <http://arxiv.org/abs/1507.05670>