Learning Compositional Hierarchies using Curriculum Learning for Compositional Generalization

Author Name
Affiliation
email@example.com

Abstract

Compositional generalization is an integral part of human intelligence. We describe insights from developmental psychology in how humans acquire new words and associated concepts. One key component in such experiments is using pseudowords to avoid pollution from previous experience. Current machine learning approaches for language differ vastly from the process outlined by experiments with humans. We argue for a research direction investigating how curriculum learning and concept learning using pseudowords can affect central capabilities such as compositional generalization. Previous work mostly focus on learning hierarchies of concepts with only one layer of abstraction. Relying on knowledge of more complex hierarchies can help language models learn more efficiently, but also achieve internal structure better suited for compositional generalization.

1 Introduction

In this work, we describe a research direction where knowledge is used to create curricula that can help models achieve compositional generalization. Recent multimodal models of language and vision, such as CLIP, DALL-E 2, and Stable Diffusion, are impressive in the way they process novel combinations of visual concepts. Empirically, there are examples suggesting that these models achieve compositional generalization to some degree. However, it is difficult to verify underlying structures and mechanisms allowing these compositions. Another difficulty is that we do not know exactly what data a system was trained on. Hence we cannot say for sure whether a prompt that seems novel was not seen during training. Compositional generalization is often benchmarked using synthetic data, allowing for high degrees of control. Previous work does this with abstract 2D concepts and pseudowords [Lake *et al.*, 2019; Lake, 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 enforce that the specific concept is not part of any pretraining procedure. 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.

2 Learning pseudowords – insights from developmental psychology

Pseudoword setups has a long-standing place in linguistics research, most famously with the Wug Test introduced by Berko in 1958. The test involves 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. 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 refered 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 that more complex notions take more time. 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.

2.1 In artificial intelligence

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.

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 tells us much about to which extent the new word and associated concept can be understood in relation to existing knowlegdge. We argue that the only conclusion we can draw is that GPT-3 performs the fast mapping described by Carey and Bartlett.

The Wug Test tells us 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 is 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.

3 Compositional generalisation using hierarchical concepts in CLEVR

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 [Kim and 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 [Kim and 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 1. 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; 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, we can then compare differences in task peformance when training on randomly ordered

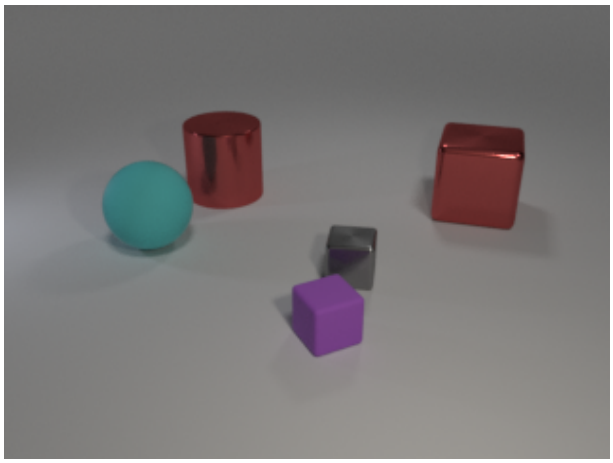

Figure 1: A simple example of data generated in CLEVR, where we see two pseudoconcepts; a) a *blargh* – two small cubes next to each other, and b) a *perde* – a large cyan sphere.

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.

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.

Askarian *et al.* look at the effects of three different curriculum learning strategies on performance in relation to amount of data and training costs. Their claim is that *"curriculum learning effectively improves low data VQA"*, showing on subsets of CLEVR how CL and L2-norm regularisation can drastically improve performance when training with only 20% of the original data. They define three different curriculum learning strategies using complexity criteria based on program length, answer hierarchy, and hard examples. The *first strategy* is based on the intuition that the length of a question is an indication of how difficult it is to answer. As a proxy for length, this strategy measures the length of the program as given in the CLEVR dataset (i.e., `filter_colorblue` counting as one operation. The *second strategy* uses an answer hierarchy created by the authors themselves. The intuition is that a learner first learns the answer type, e.g. that a question requires a number as its answer. From this intution, Askarian *et al.* constructs a hierarchy of the answer types, we refer to [Askarian *et al.*, 2021] for illustrations. Hardness is then defined as how far from the hierarchy root an answer is. Their *third strategy* uses examples that yield high learner loss. This makes it the only strategy to have a dynamic hardness criteria, since the loss will change for hard examples over time as they become easy for the model to answer.

As further insight into the benefits of curriculum learning for visual question answering, Aissa *et al.* proposes a Neural Module Network (NMN) method for Visual Question Answering that uses predefined cross-modal embeddings and curriculum learning to reduce the cost of training and the amount of training data while still achieving good accuracy. They show how their curriculum learning strategies allow the NMN model to achieve the same performance using half of the data and 18 times less compute. Their main hardness criteria is a combination of the number of objects in a scene and the program length of a given question. They complement this hardness criteria with pretraining on random examples, and two weighting strategies to 1) achieve uniform distribution over the different answer types, and 2) weigh examples proportinal to the sum of the avergae losses of the program modules corresponding to the quesition (this focuses the model on hard examples). These strategies are all follow the same spirit as the strategies presented by Askarian *et al.*. Aissa *et al.* move away from CLEVR into the more natural domain of GQA, to provide a more challenging and complex setup.

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.

Following the longstanding historical debate on what concepts are and how they are useful for artificial intelligence, as exemplified by, e.g., Fodor, there is a growing literature of concept- and meta learning that can provide further insights into how we can construct architectures that address the challenges outlined in this paper.

4 Tasks and experiments

The following research questions and challenges should be addressed:

- Does models learn category abstractions, and can curriculum learning help enforce that?
- What is the effect of curriculum learning [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)
- Different levels of comprehension, similar to the Wug [Berko, 1958] and Chromium [Carey and Bartlett, 1978] tests
  - Detecting the presence of an object – image classification

  - Point out concept in line up – object recognition
  - Use it in a task, e.g. mathematical reasoning
- How do we construct a rich enough hierarchy?

References

[Aissa *et al.*, 2023] Wafa Aissa, Marin Ferecatu, and Michel Crucianu. Curriculum learning for compositional visual reasoning. *ArXiv*, abs/2303.15006, 2023.

[Askarian *et al.*, 2021] Narjes Askarian, Ehsan Abbasnejad, Ingrid Zukerman, Wray Buntine, and Gholamreza Haffari. Curriculum learning effectively improves low data vqa. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 22–33, 2021.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[Berko, 1958] Jean Berko. The child's learning of english morphology. *Word*, 14(2-3):150–177, 1958.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Carey and Bartlett, 1978] Susan Carey and Elsa Bartlett. Acquiring a single new word. 1978.

[Eustace, 1969] Barbara W Eustace. Learning a complex concept at differing hierarchical levels. *Journal of Educational Psychology*, 60(6p1):449, 1969.

[Fodor, 1998] Jerry A Fodor. *Concepts: Where cognitive science went wrong*. Oxford University Press, 1998.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[Kim and Linzen, 2020] Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*, 2020.

[Lake *et al.*, 2019] Brenden M Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. In *41st Annual Meeting of the Cognitive Science Society: Creativity+ Cognition+ Computation, CogSci 2019*, pages 611–617. The Cognitive Science Society, 2019.

[Lake, 2019] Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. *arXiv preprint arXiv:1906.05381*, 2019.

[Lovering and Pavlick, 2022] Charles Lovering and Ellie Pavlick. Unit testing for concepts in neural networks. *arXiv preprint arXiv:2208.10244*, 2022.

[Ruis *et al.*, 2020] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. A benchmark for systematic generalization in grounded language understanding. *Advances in neural information processing systems*, 33:19861–19872, 2020.

[Wang *et al.*, 2022] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022.

[Zhao *et al.*, 2022] Linfeng Zhao, Lingzhi Kong, Robin Walters, and Lawson LS Wong. Toward compositional generalization in object-oriented world modeling. In *International Conference on Machine Learning*, pages 26841–26864. PMLR, 2022.