Learning Compositional Hierarchies using Curriculum Learning for Compositional Generalisation

Adam Dahlgren Lindström
Umeå University
dali@cs.umu.se

Abstract

Compositional generalisation 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 generalisation. 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 generalisation.

1 Introduction

In this work, we describe a research direction investigating how knowledge-based curricula can help models achieve compositional generalisation. 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 generalisation 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 generalisation 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 new words – 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. 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 investigates how children learn a single new word. The authors detail the process of acquiring a word using different pieces of information for different tasks [Carey and Bartlett, 1978]. According to them, a learner

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

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 of the word “chromium”,
- baseline vocabulary assessment,
- a sorting task and a naming task based on object colors,
- a comprehension task, and a hyponym task.

In the sorting task, the children used 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*. To test comprehension, the children were tasked with pointing at three colors, one of which was chromium, controlling whether they had properly learnt a referent for "chromium". The hyponym task 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 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 investigates learning complex concepts 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. Using only pseudowords mitigates the problem of information leaking from the training data, which is why such investigations can be useful. 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 learn new words, indicating that deep learning-based methods can be built to acquire new words. It does not tells us much about to which extent the new word and associated concept is understood in relation to existing knowledge. The only conclusion we can draw is that GPT-3 performs the fast mapping described by Carey and Bartlett.

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 using a flat hierarchy, as they describe three stages (object extraction, action binding, and transition modeling between properties) of learning using their model.

2.2 Curriculum Learning for Learning Concepts

We have covered some of the dynamics of how *a* new word is acquired, but not how we can build the internal structures necessary to fit that word into. 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 to estimate difficulty in order to create a curriculum. With a curriculum, we can then compare differences in task performance when training on randomly ordered concepts versus using a curriculum. Using synthetic data, we can use known hierarchical structures to reflect the complexity of a concept. Beyond performance, we can hypothesise about how a curriculum affects the internal structure of a model to better allow for compositional generalisation.

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 intuition 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. They use predefined cross-modal embeddings and curriculum learning to reduce the cost of training and the amount of training data while still achieving good accuracy [Aissa *et al.*, 2023]. 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 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.

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. See, e.g., [Hospedales *et al.*, 2022; Vinyals *et al.*, 2016; Snell *et al.*, 2017; Cao *et al.*, 2021] for more on how meta- and concept learning provide efficient learning mechanisms and more interpretable models. However, these examples only models one level of abstraction.

3 Compositional generalisation using hierarchical concepts in CLEVR

Compositional generalisation is now extensively studied in language models (e.g. COGS [Kim and Linzen, 2020], SCAN [Lake and Baroni, 2018]), 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 describe the blueprint of a compositional generalisation benchmark for hierarchical concepts using the ideas from developmental psychology previously outlined. The suggested benchmark used designed with CLEVR using pseudoword concepts that build hierarchically on each other, exemplified in Figure 1.

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. As shown by Carey and Bartlett, we learn different aspects of a word with different speed, and testing should reflect these expectations. Our benchmark can be used to investigate whether this means that we can expect a pseudoword to be lexically understood, and that learning hyponyms relations will take longer to learn. We propose a similar approach to that of Lovering and Pavlick, using evaluations of internal structure and external behaviour complementary. 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. Evaluating models according to these ideas will help us understand how to better address compositional generalisation in language models.

We now devise multiple tasks through which the comprehension of these concepts are tested. In the spirit of the Wug [Berko, 1958] and Chromium [Carey and Bartlett, 1978] tests, we devise multiple tasks to test different levels of concept comprehension;

- Can the model confidently recognise other words of the same type, e.g. other colors?
- Determine the concept type of the pseudoword. Is it an attribute, object, or action?
- Determining presence or absence of concept in image.
- Where in an image is the concept present?

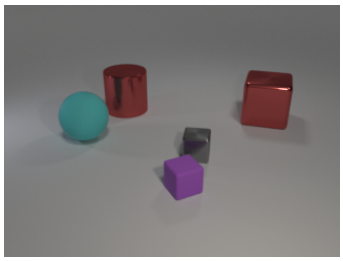

Figure 1: A simple example of data generated in CLEVR, where we see two pseudoconcepts; a) a *blargh* – two small cubes next to each other, and b) a *perde* – a large cyan sphere.

- Determine whether a known word is a hypernym or not, using the given hierarchy.
- Using the concept in a complex mathematical reasoning task.

Following the ideas of Lovering and Pavlick, a model can be evaluated on all these tasks throughout training to investigate the relationship between the different capabilities. From the description in 2.2, we can investigate how curriculum learning affects these capabilities. 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)
- How do we construct a rich enough hierarchy in a synthetic domain?

While we should not anthropomorphise language models, we can still use insights from human learning mechanisms to achieve the capabilities we want and need from our models. The proposed benchmark combines ideas from many different research directions, including curriculum learning, neurosymbolic methods and concept learning, to improve our understanding of how to build models that achieve better compositional generalisation. In doing so, knowledge is an integral part of transferring hierarchical structures from teacher to student.

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.

[Cao *et al.*, 2021] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *International Conference on Learning Representation (ICLR)*, 2021.

[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.

[Hospedales *et al.*, 2022] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169, 2022.

[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 and Baroni, 2018] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR, 7 2018.

[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.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[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.