Compositional Cores: Persistent Attention Patterns in Compositionally Generalizing Subnetworks

Michael Y. Hu* **Chuan Shi*** **Tal Linzen**
Center for Data Science, New York University
{michael.hu, cs5526, linzen}@nyu.edu

Abstract

Transformer language models have shown improvements on compositional generalization benchmarks, but we lack understanding of how these models actually implement compositional generalization. In this work, we propose a method to identify the *compositional core*–the key subnetwork that models use to generalize compositionally. We compare this compositional core against subnetworks from models that simply memorize tasks or rely on shallow distributional patterns. Our analysis reveals that the attention mechanisms in compositionally generalizing subnetworks behave distinctively, with a notable focus on the end-of-sequence (EOS) token. This finding suggests that language models may be using special tokens like EOS as registers to hold and manipulate sentence representations.

Extended Abstract

Compositional generalization in natural language, or the ability to understand and produce new utterances using known primitives (Fodor and Pylyshyn, 1988), is a hallmark of human learning and a known challenge for language models (Kim and Linzen, 2020; Lake and Baroni, 2018; Li et al., 2023). A successful recent strategy for improving the compositional generalization capabilities of transformer language models (LMs) is *scaling*, or increasing the size of LMs and their pretraining data (Orhan, 2022; Zhou et al., 2023; Press et al., 2023). However, one detriment of scaling is that we do not understand how these models implement compositional generalization internally; we only know that these LMs perform better on compositional generalization benchmarks.

In this work, we aim to find persistent properties of how LMs implement seq-to-seq compositional generalization: properties that directly impact performance and are invariant to randomness in the fine-tuning process. To pinpoint where compositional generalization is implemented in the LM, we apply structured pruning (Xia et al., 2022) to trained models, removing the components of transformers that contribute minimally to task performance. Next, we aggregate over several training and subsequent pruning runs and only consider properties that arise in all cases.[1]

Our experimental procedure is as follows:

1. Take a pretrained language model and fine-tune it using several random seeds on a compositional generalization task.

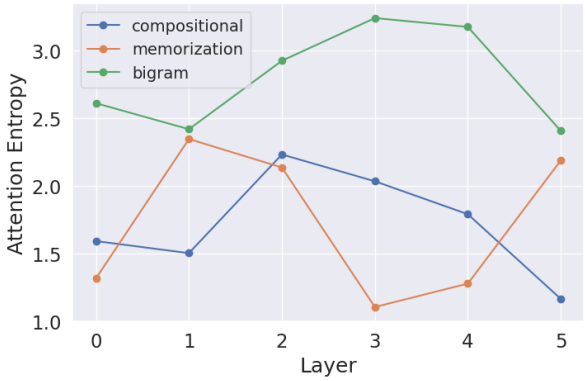
2. For each fine-tuned model, use CoFi pruning (Xia et al., 2022) to find the subnetwork driving task performance by selectively removing unused attention heads. For each fine-tuned model, prune using several random seeds.

3. The result of pruning is a mask over attention heads, indicating which can be kept or deleted (1 or 0). We take the intersection of these masks to find the attention heads that are kept over all instances of fine-tuning and pruning. Following Bhaskar et al. (2024), we call this set of attention heads the *heuristic core*.

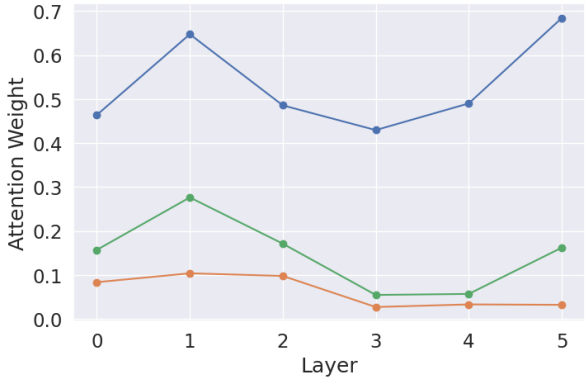
Next, we find the attention head properties that are unique to the heuristic core of compositionally generalizing LMs. To do so, we construct coupled tasks that share inputs with the compositional generalization task but differ in outputs. In the **memorization** task, the output sequence is scrambled, so the only way to learn the training task is to memorize the examples. In the distributional or **bigram** task, the language model is trained to predict the output of a bigram model of the training corpus instead of the next word.

*Equal contribution.

[1] In our experiments, we find that the attention patterns of LMs fine-tuned on the same task with different random seeds are nearly identical to each other.



(a) Layerwise attention entropy trends for T5-small.



(b) Attention weight trends on the EOS token. The compositional core places the most attention weight on the EOS token, especially towards the later layers.

Since the **memorization**, **bigram**, and **compositional** tasks share inputs, we can examine how the same model architecture processes these tasks differently, given the same input sequence.

Results

We study the heuristic cores of T5-small, an encoder-decoder transformer model that performed well in previous work on compositional generalization tasks (Petty et al., 2024; Orhan, 2022). We compute heuristic cores for the memorization, bigram, and compositional (i.e., original) versions of COGS (Kim and Linzen, 2020), a semantic parsing task testing compositional generalization. In COGS, words in the training and test sets appear in disjoint grammatical roles; models solving COGS must apply grammatical roles compositionally to predict output from input. We observe three main findings:

- Models trained on the memorization task cannot be pruned to high sparsity. The memorization core uses most of the LM's attention heads.
- The bigram heuristic core has higher cross attention entropy, indicating more diffuse attention across all tokens.
- In the compositional task, models use cross attention heads that attend to the end-of-sequence (EOS) token more.

Both the bigram and compositional models can be pruned to 60% sparsity in attention heads, causing a ∼17% drop in performance. In other words, the remaining 40% of the attention heads recover 80% of the original performance. However, the memorizing model can only be pruned to around 30% sparsity before dramatic reductions in ability to memorize. Thus, memorization requires more model parameters than bigram or compositional generalization on this task.

To understand the behavior of the remaining attention heads, we analyze their attention patterns (Clark et al., 2019) during the decoding of the first output token. Specifically, we examine the entropy and weights of the cross attention heads in T5-small. Figure 1a shows that the bigram model exhibits higher average cross attention entropy across all layers compared to both the memorization and compositional cores, indicating a more uniform attention distribution over all input tokens.

In contrast, the compositional core's cross attention demonstrates a different distinctive pattern: it allocates significantly more weight to the EOS token. This behavior differs from the bigram and memorization cores, which primarily attend to tokens within the sentence. Since attention weight on the EOS token is higher across all layers, we hypothesize that the EOS token's activation holds a representation for the full sentence, which may be especially useful in the compositional task variant.

We plan to further study this hypothesis in three ways. First, if we remove the EOS token in all tasks, we would expect performance on COGS to decrease but remain roughly the same for the memorization and bigram variants. Second, we can probe the compositional core's EOS token activation for intermediate sentence representations, such as parts of speech or components of the sentence's parse tree. Last, we can attempt to improve compositional generalization by adding more special tokens that the model can use for manipulating intermediate representations (Pfau et al., 2024).

References

Adithya Bhaskar, Dan Friedman, and Danqi Chen. 2024. The heuristic core: Understanding subnetwork generalization in pretrained language models. *Preprint*, arXiv:2403.03942.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

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

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A structural generalization benchmark for semantic parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.

A. Emin Orhan. 2022. Compositional generalization in semantic parsing with pretrained transformers. *Preprint*, arXiv:2109.15101.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7239–7252, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *Preprint*, arXiv:2404.15758.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1513–1528, Dublin, Ireland. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.