Dissecting Compositional Generalization: Correlation Analysis on Generalization Capacities of Different Compositional Problems

Yoobin Cheong, Yeong Koh and **Yoon Tae Park**
Center for Data Science
New York University
60 5th Avenue, New York, NY
{yc5206, yk2678, yp2201}@nyu.edu

1 Introduction

There exist many benchmark tests that claim to evaluate the abstract capacity of compositional generalization. However, whether these tests that fall under this abstraction target the same underlying capacity is an open question. (Weißenhorn et al., 2022; Furrer et al., 2020; Li et al., 2019; Liu et al., 2021; Hupkes et al., 2019)

As a means to investigate this idea, we will perform a correlation analysis on the performance of various neural network models trained on multiple tasks that have been proposed to measure compositional generalization. As a potential outcome of our project, grouping of tasks that are identified as being highly correlated will provide insights towards a better characterization of the abstraction compositional generalization, and furthermore be useful for task selection in multitask learning or transfer learning.

2 Experimental Design

Our team will start by finding relevant datasets and models to test. We will test some Sequence-to-Sequence models on widely used benchmark datasets such as SCAN (Lake and Baroni, 2018), CFQ (Keysers et al., 2019), and COGS (Kim and Linzen, 2020). After achieving good performances on given models, we will create a training evaluation pipeline, so that we can further test different models with different datasets or tasks. In this experiment, each task means a different split or generalization subcase of different datasets.

Then, we will compare the performances of the models through correlation analysis. For example, a high correlation between two tasks would be an indication that a network that performs well on one task performs well on another, suggesting that the two tasks recruit similar capacities. On the other hand, a low or no correlation would mean that they may require distinct solutions as a solution for one task does not generalize well to another. At this stage, we will be working on the Greene cluster, to get time efficiency on multiple tests that would be conducted for next step and further developments.

3 Future Extension

This research can be further developed by applying the same experiments to other relevant datasets and models and analyzing the results. If the result shows some highly correlation between two tasks, we will also examine whether this data can be useful for task selection in multitask learning or transfer learning. If the result shows low or no correlation between two tasks, we will conclude that there are no straightforward correlations among different tasks based on our research and further investigation is needed.

4 Collaboration Statement

Our team will work altogether for all tasks in general, but will focus on different parts: (1) Yoobin Cheong will identify a set of compositional generalization datasets and models to test, and organize the different datasets into a single unified format, (2) Yoon Tae Park will build a codebase for the training evaluation pipeline, and (3) Yeong Koh will perform correlation analysis with the evaluation results and document any interesting findings. If correlation analysis shows some coherent, interpretable grouping of tasks, we will also test if this information can be helpful for task selection in multitask learning or transfer learning.

Acknowledgments

This research is supervised and mentored by Najoung Kim, a faculty fellow at CDS, as a part of her ongoing research project.

References

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. 2020. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. Compositionality decomposed: how do neural networks generalise?

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2019. Measuring compositional generalization: A comprehensive method on realistic data.

Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation.

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.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019. Compositional generalization for primitive substitutions.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2021. Challenges in generalization in open domain question answering.

Pia Weißenhorn, Yuekun Yao, Lucia Donatelli, and Alexander Koller. 2022. Compositional generalization requires compositional parsers.