Optimizing Grokking Efficiency: Structured Learning Dynamics and Regularization Strategies in Neural Networks

RESEARCH PROPOSAL

Maryam Taj
2022907

November, 8th 2024

I. Introduction

Problem Statement

Grokking, a term introduced in recent studies, refers to the emergent generalization seen in neural networks after prolonged training on tasks like modular arithmetic or sparse parity (Power et al., 2022; Merrill et al., 2023). Initially, models memorize patterns, achieving high training accuracy without generalizing. Over time, however, they transition to a state of high test accuracy, reflecting a deeper understanding of the task (Nanda et al., 2023). While grokking demonstrates powerful generalization, it comes with high computational costs and lengthy training times. This project aims to address these inefficiencies by exploring structured learning dynamics and targeted regularization to accelerate the grokking process.

Objectives

I. Research Question 1

What role do structured learning dynamics play in achieving efficient grokking?

II. Research Question 2

How can regularization and curriculum learning optimize learning stages for grokking?

III. Research Question 3

Can targeted architectural modifications facilitate early generalization in neural networks?

II. Literature Review

Mechanistic Interpretability (Nanda et al., 2023): Neural networks go through phases—memorization, circuit formation, and cleanup—before grokking. Understanding these phases can help accelerate learning.

Sparse Networks (Merrill et al., 2023): Grokking involves a phase transition driven by sparse subnetworks. Targeted pruning can improve training efficiency.

Regularization & Weight Decay (Thilak et al., 2022): Weight decay enhances grokking by prioritizing generalization over memorization, suggesting controlled regularization can speed up the process.

Curriculum Learning (Barak et al., 2022): Gradually increasing task difficulty improves generalization and could reduce resources needed for grokking.

Learning Rate Schedules (Ganguli et al., 2022): Dynamic learning rates enhance stability during training, potentially reducing time for grokking.

Fourier Interpretability (Liu et al., 2023): Fourier analysis helps identify key components for targeted pruning, improving grokking efficiency.

Research Gaps

The reviewed literature points to several strategies for encouraging generalization but lacks a cohesive framework that combines these approaches to accelerate grokking. This project will address this gap by synthesizing structured training methods, such as curriculum learning and targeted sparsification, with regularization strategies to achieve efficient grokking.

III. Methodology

Approach

This research will investigate the dynamics of grokking across various neural network architectures, analyzing stages within the training process that correspond to memorization, circuit formation, and cleanup. Using interpretability techniques, the study will develop metrics to monitor progress within these phases.

IV. Expected results

This research expects to identify efficient pathways for achieving grokking by using curriculum learning, regularization, and targeted pruning to encourage early formation of sparse, generalizable structures. These findings could lead to practical guidelines for enhancing neural network generalization with minimal resource investment, potentially benefiting tasks requiring rapid, reliable learning.

References

1. Nanda, N., et al. "Progress measures for grokking via mechanistic interpretability." *ICLR 2023*.
2. Merrill, W., et al. "A tale of two circuits: Grokking as competition of sparse and dense subnetworks." *ICLR Workshop on Understanding Foundation Models*, 2023.
3. Thilak, V., et al. "The slingshot mechanism: An empirical study of adaptive optimizers and the Grokking Phenomenon." *NeurIPS 2022 Workshop*.
4. Power, A., et al. "Grokking: Generalization beyond overfitting on small algorithmic datasets." *arXiv preprint arXiv:2201.02177*, 2022.
5. Barak, B., et al. "Hidden progress in deep learning: SGD learns parities near the computational limit." *Advances in Neural Information Processing Systems*, 2022.
6. Engel, A., & Van den Broeck, C. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.
7. Liu, Z., et al. "Omnigrok: Grokking beyond algorithmic data." *International Conference on Learning Representations*, 2023.