Grokking in Neural Networks

Ahmed Mohammed Ahmed[1]

Supervisor: **Dr. Tiffany Vlaar**[1]

[1]African Institute for Mathematical Sciences (AIMS), Ghana

DEEP LEARNING INDABA

Abstract

Typically, as a neural network's performance on training data improves, its performance on (unseen) validation data also improves until the network starts overfitting. However, grokking, a recently discovered phenomenon by Power et al. (2022), defies this norm by showing a sudden and dramatic improvement in validation performance long after overfitting has begun. In this project, we explore the concept of grokking in the context of modular addition, examining its occurrence in different types of neural networks and training setups. We discovered that grokking happens in many types of models and is influenced by the amount of training data used. Specifically, we conduct our experiments on recurrent neural networks (RNNs) and long short-term memory networks (LSTMs).

Double Descent Background

The classical problem in machine learning is to find a special point, often referred to as the "sweet spot", where the loss function is minimized on both training and test datasets. This belief is based on the bias-variance trade-off as shown in Figure 1.

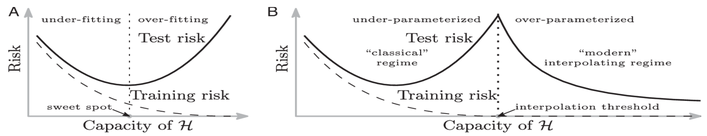

Figure 1. (A) U-shaped curve, classical theory. (B) Double descent curve, where $\mathcal{H}$ is the function class for a specific neural network architecture.

Grokking Phenomenon

Phase 1 - Initial Overfitting: Model memorizes training data (100% accuracy) but fails on validation set.

Phase 2 - Sudden Generalization: After extended training, an unexpected transition to near-perfect validation accuracy occurs.

Figure 2. **Grokking Phenomenon:** X-axis: optimization steps (log scale), Y-axis: accuracy. Red line = training accuracy, Green line = validation accuracy.

Research Questions

- Do different types of neural network architectures exhibit grokking behavior?
- What is the effect of training dataset size on different architectures?
- How does turning regularization off affect the grokking phenomenon?

Mathematical Setup & Experimental Design

Training Data: Let $(\mathbf{x}_i, y_i)_{i=1}^n$ be the training inputs and labels.

- $\mathbf{x}_i \in \mathcal{D}$ represents the sequence of tokens (e.g., ['<sos>', 'x ', '+'," y ', 'mod' ,'=']).
- $y_i \in \mathbb{R}$ represents the result of the arithmetic operation (e.g., $(x + y) \mod 91$).
- y_i is one-hot encoded in $\mathbb{R}^C$, where C is the number of possible results modulo mod.

- $h : \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^C$ represents the hypothesis class of neural networks.
- $\mathbf{w} \in \mathbb{R}^p$ denotes the parameters of the neural network.
- $h(\mathbf{w}, \cdot) : \mathcal{D} \to \mathbb{R}^C$ represents the network's predictions for inputs in $\mathcal{D}$.

Training Loss

$$\mathcal{L}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(\mathbf{w}, \mathbf{x}_i)) \tag{1}$$

Weight Decay Regularization

$$\mathcal{L}_\lambda(\mathbf{w}) := \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{2}$$

The training loss, where $\ell(\cdot, \cdot)$ is the cross-entropy loss function.

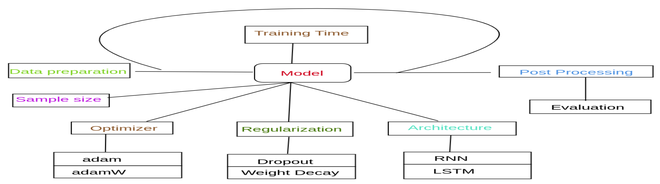

Figure 3. Grokking Experimental Design

Grokking in RNN Models with Varying Training Data Fractions

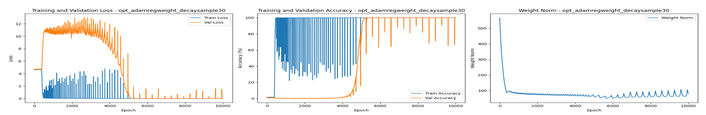

Figure 4. Training and validation loss curves for an RNN model trained on 30% of the dataset.

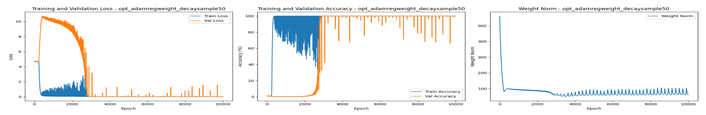

Figure 5. Training and validation loss curves for the RNN model trained on 50% of the data.

Novel Discovery: Late Weight Decay

We aimed to identify the effects of turning off weight decay from the beginning and then reapplying it at different epochs, with either the same or increased weight decay values.

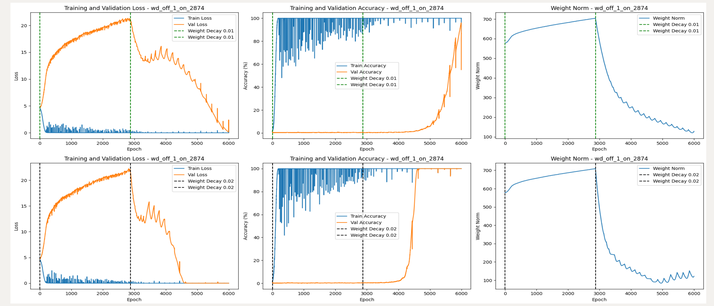

Figure 6. **Dynamic Weight Decay:** Green and black dashed lines indicate points where weight decay of 0.01 and 0.02 were applied or removed.

Impact of Weight Decay on Grokking

Weight Decay Experiments: We discovered that grokking happens in many types of models and is influenced by the amount of training data used. Our results show that using regularization techniques such as weight decay can help the models generalize better.

Novel Finding: Introducing weight decay only late in training still leads to grokking, which is a novel discovery to our knowledge.

- Training without weight decay leads to poor generalization.
- Removing weight decay too early also has the same effect.
- Weight decay is not necessary during the early phases of training.

Conclusion

Our experiments confirmed that weight decay, sample size, and extended training are critical factors influencing the grokking phenomenon in neural networks. Our findings about the impact of training set size on grokking are similar to what Liu et al. (2022) found, even though they used different types of datasets and training configurations.
Future Work: Investigate new architectures (Mamba, XLSTM) and understand when explicit regularization is necessary for grokking.

References

Power et al. (2021): *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets*, International Conference on Learning Representations (ICLR), 2021.

Liu et al. (2022): *OmniGrok: Grokking Beyond Algorithmic Data*, International Conference on Learning Representations (ICLR), 2022.