Final Project: Grokking Phenomenon Reproduction

Kecen Sha *
Your ID

Yuziheng Wu *
Your ID

Di Yue *
2100012961

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Grokking [1]

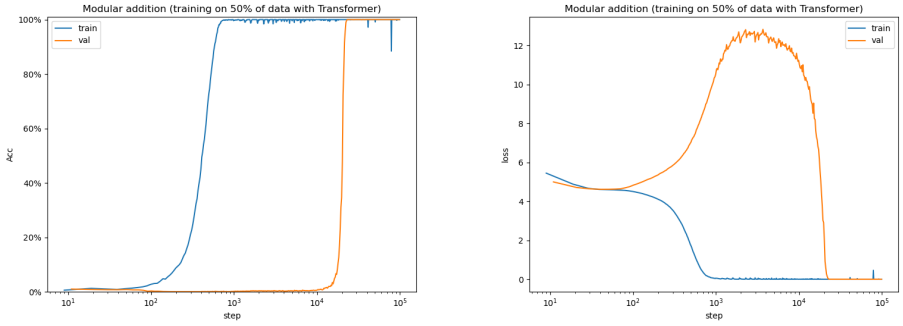
2 Preliminary

3 Experiments

3.1 Reproducing the Grokking Curve

In this task, we investigate the grokking phenomenon in modular addition with the transformer model. Specifically, we use a decoder-only transformer with 2 layers, width $d_{\text{model}} = 128$, 4 attention heads and dropout 0.1. We use the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$, learning rate 10^{-3}, weight decay 0.1, and linear learning rate warmup over the first 10 updates. For $p = 97$ and $\alpha = 0.5$, we randomly separate an α fraction from all p^2 equations in $\mathbb{Z}_p$ as the training set; the rest serves as the validation set. The model is trained with minibatch size 512 for 10^5 steps.

The accuracy and loss throughout the training process are plotted in Figure 1.



(a) Training and validation accuracy (b) Training and validation loss

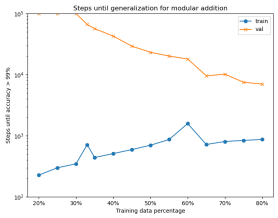
Figure 1: The grokking curves of transformer model

*Equal contribution.

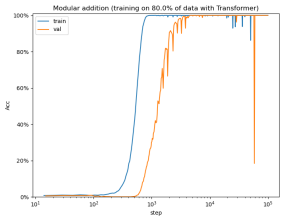
Preprint. Under review.

As is shown in Figure 1a, the model overfits the training data within 10^3 updating steps. Nevertheless, generalization does not happen until after 10^4 steps. Figure 1b further illustrates that the decrease of loss is consistent with the increase of accuracy, both for training and validation phases. Interestingly, while the training loss decreases monotonically, an increase of the validation loss is observed before it begins to converge.

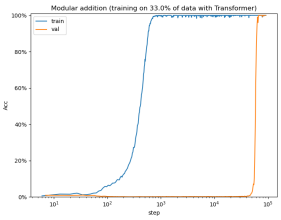
We further study the effect of the training data fraction α. For each α we sample, we train the model on a random α fraction of all p^2 equations for at most 10^5 steps, and determine the minimum number of steps required to achieve validation accuracy $\geq 99\%$. The results are plotted in Figure 2a. As a comparison, we plot the accuracy curves for $\alpha = 80\%$ and $\alpha = 33\%$ in Figures 2b and 2c, respectively.



(a) Minimal number of training steps required to achieve validation accuracy $\geq 99\%$



(b) The accuracy curve for $\alpha = 80\%$



(c) The accuracy curve for $\alpha = 33\%$

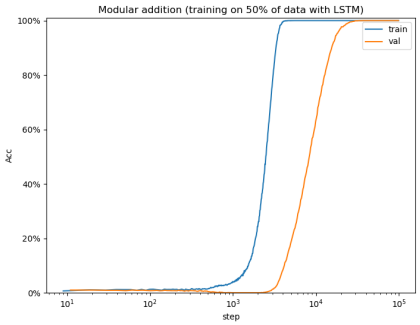
Figure 2: Effect of the training data fraction α

When $\alpha = 80\%$, the model generalizes in 10^4 steps. As α decreases, it becomes easier for the model to overfit the training data, while the number of steps required for generalization increases rapidly. When $\alpha \leq 30\%$, the model would not generalize in 10^5 steps.

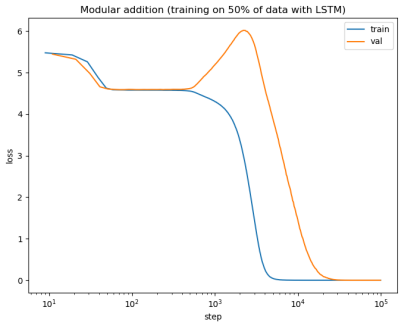
3.2 Grokking Phenomenon of Other Models

We study the grokking phenomenon for two other network architectures, long short-term memory (LSTM) and multilayer perceptron (MLP), suggesting that grokking is not unique for transformer. Specifically, our LSTM has 2 layers, a hidden state of size $d_{\text{hidden}} = 128$, and dropout 0.1. Our MLP has 2 hidden layers of size $d^{(1)}_{\text{hidden}} = d^{(2)}_{\text{hidden}} = 256$, both of which have dropout 0.1. Other hyperparameters remain the same as Section 3.1.

The grokking curves for LSTM and MLP are shown in Figure 3 and Figure 4, respectively.



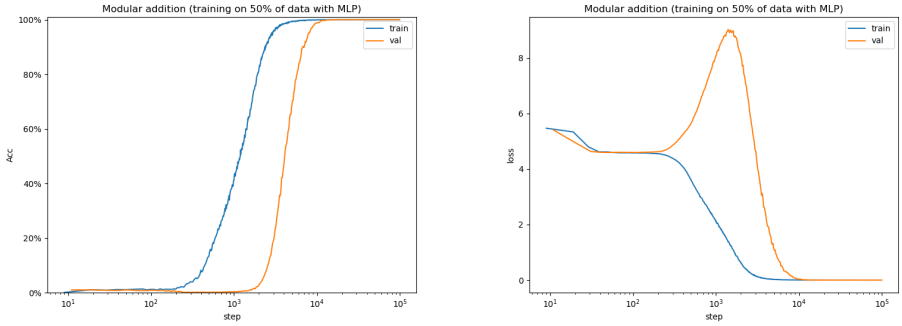
(a) Training and validation accuracy



(b) Training and validation loss

Figure 3: The grokking curves of LSTM model

In Figures 3a and 4a, we again observe a delay of generalization which, however, is not as significant as that in Figure 1a. Besides, a similar increase of validation loss could be observed in both Figure 3b and Figure 4b.



(a) Training and validation accuracy (b) Training and validation loss

Figure 4: The grokking curves of MLP model

3.3 Effects of Different Hyper-Parameters

3.4 Grokking for K-Wise Modular Addition

4 An Explanation of the Grokking Phenomenon

References

[1] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *CoRR*, abs/2201.02177, 2022.

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.