GROKKING - WHAT/WHEN/WHY?

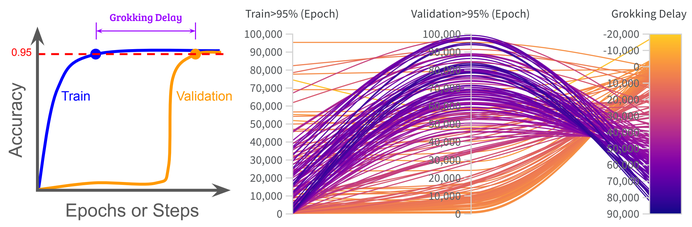
Huanran Li†, Sadman Sakib‡

Department of Electrical Engineering† / Computer Science‡, Wisconsin Institute for Discovery†, University of Wisconsin-Madison†,‡

WHAT is Grokking?

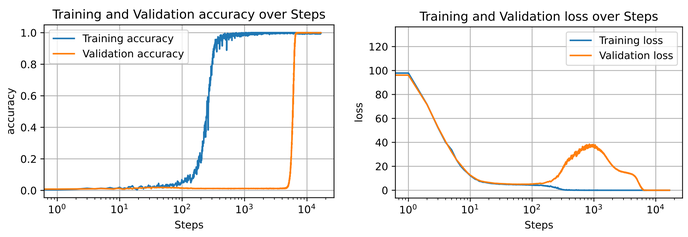
Grokking is a phenomenon when the validation accuracy continues to improve significantly after the training accuracy has plateaued. Grokking usually occurs **when the dataset size of training/validation ratio is small** (less than 40/60). For later analysis, we denote the **grokking delay** as the number of steps/epochs between the points when training and validation accuracies reach 95%.

Grokking is NOT wanted. Right figure: In practice, we observe that grokking is primarily **due to the delay in validation** rather than early convergence of training.



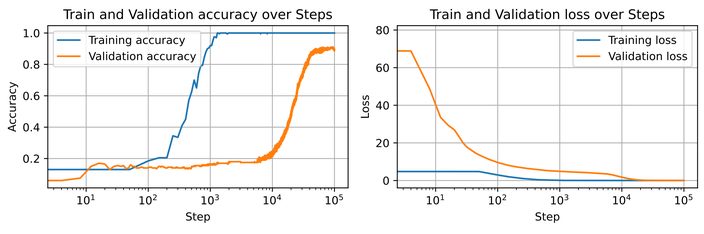
Grokking on Transformers (Arithmetic Data)

We trained a **decoder-only transformer** model on an arithmetic dataset following [3]. The dataset includes operations of the form $(x \circ y) \bmod p = z$, where $\circ$ is an arithmetic operation **(addition, subtraction, multiplication, division)** and p is a prime number. The model we used for the experiments is a 2-layer transformer model with 4 heads in one transformer block. For optimization, we utilized the AdamW optimizer with Cross Entropy loss and Weight Decay regularization.



Grokking on MLPs (MNIST)

Following [2], we replicated their experiment results with the **MNIST** dataset. We employed a **3-layer Multi-Layer Perceptron (MLP)** with a width of 200 and ReLU activation functions. For optimization, we utilized the AdamH optimizer with Mean-Square-Error (MSE) loss function and Weight Decay regularization. The batch size was set to 200. Notably, the network weights were initialized 8x larger than usual. The figures presented below are the results of training with **1k images**.

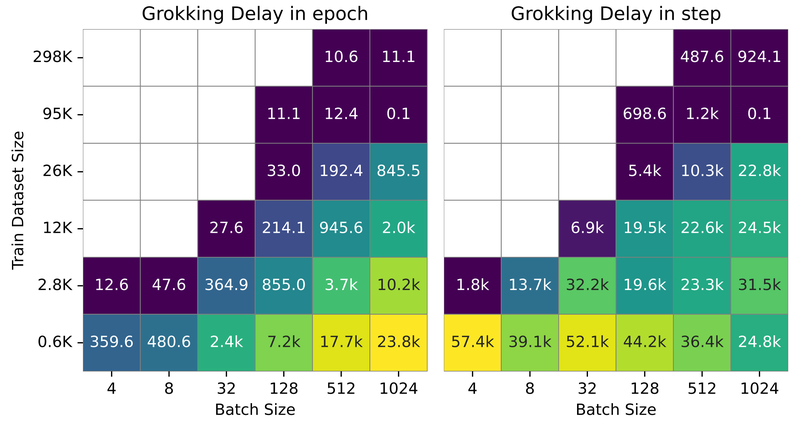


WHEN are Transformers not Grokking?

The RATIO (training dataset size) / (batch size) MATTERS!

We trained the **transformers** on a **division** dataset with varying sizes of training datasets and batch sizes. The size of the training dataset is controlled by the scale of the prime number p. The train/validation ratio is consistently at 30/70 for optimal observation of grokking.

We plotted the **average delay** in **epochs (first graph)** and **steps (second graph)** for runs that achieved 95% accuracy in both training and validation. Each cell represents the average of 10 runs with various levels of input noise, weight initialization scales, and weight decay rates. White regions indicate that none of the 10 runs reached 95% accuracy in 100k steps.

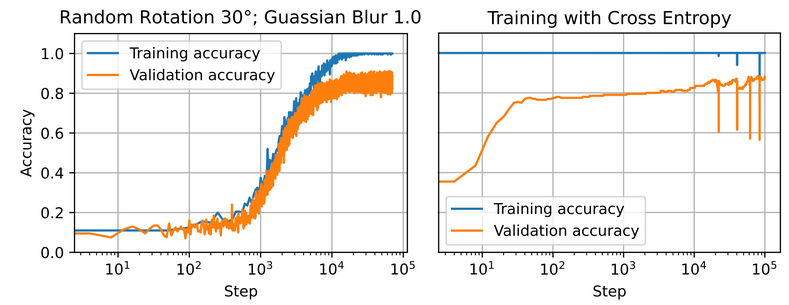


Remarkably, in both graphs, we identified a **dark diagonal region**, where the **grokking delays are shorter** than the rest. We named this region the **"comprehension band"**, within which the ratio of training dataset size to batch size consistently falls within the range of **100 - 800**.

WHEN are MLPs not Grokking?

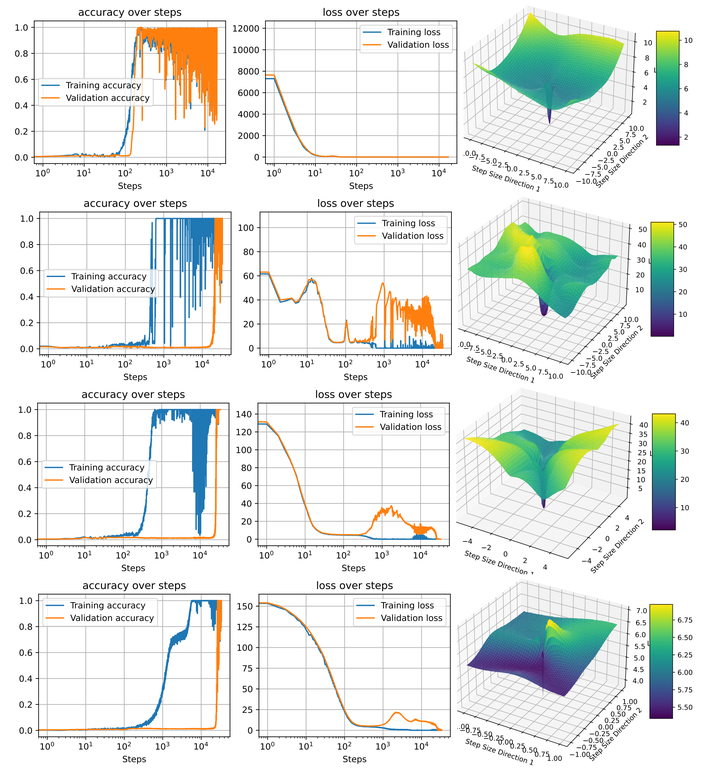
Loss function and augmentation eliminate grokking delay on MNIST!

We trained the same **MLP on 1k MNIST images** but made separate changes to the training dataset size, augmentation strategy, loss function, optimizer, batch size, regularizer, and weight initialization scale. Except for the findings from [2] on weight decay rates and weight initialization scale, we observed that switching the **loss function** or adding **augmentation** could completely **eliminate the grokking delay**. We plotted the results of two runs: 1) added **random rotation** and **Gaussian blur** to the input images; 2) switched the **loss function** from MSE to **Cross Entropy**.

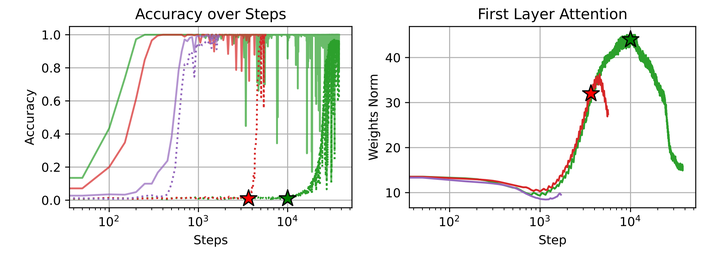


Still Open: WHY are Transformers Grokking?

Landscape of Loss. Following [1], we plotted the contour of the loss around the optima by stepping in two sets of randomly-generated gradients. Compared to the non-grokking (comprehension) optima, the **landscape of grokking optima is more uneven**, and the range of loss could be 8x larger.



Explosion of Attention Weights. We observe a **massive increase** in attention weight's norm **when training accuracy is high** and **validation accuracy is low** (pre-grokking stage). Then, as the validation accuracy improves, the weights start to drop down.



References

[1] Hao Li et al. "Visualizing the loss landscape of neural nets". In: *Advances in neural information processing systems* 31 (2018).

[2] Ziming Liu, Eric J Michaud, and Max Tegmark. "Omnigrok: Grokking beyond algorithmic data". In: *The Eleventh International Conference on Learning Representations*. 2022.

[3] Alethea Power et al. "Grokking: Generalization beyond overfitting on small algorithmic datasets". In: *arXiv preprint arXiv:2201.02177* (2022).