Information-Theoretic Progress Measures reveal Grokking is an Emergent Phase Transition

Kenzo Clauw [1] **Daniele Marinazzo** [1] **Sebastiano Stramaglia** [2]

Abstract

This paper studies emergent phenomena in neural networks by focusing on grokking where models suddenly generalize after delayed memorization. To understand this phase transition, we utilize higher-order mutual information to analyze the collective behavior (synergy) and shared properties (redundancy) between neurons during training. We identify distinct phases before grokking allowing us to anticipate when it occurs. We attribute grokking to an emergent phase transition caused by the synergistic interactions between neurons as a whole. We show that weight decay and weight initialization can enhance the emergent phase.

1. Introduction

Grokking is a phenomenon where neural networks during training suddenly generalize after prolonged memorization with limited progress in the loss. Understanding this phase transition is vital for AI safety and alignment. While (Wei et al., 2022) claim these models exhibit emergent properties, others attribute it to the choice of evaluation metric (Schaeffer et al.). This lack of consensus poses a key question: Are neural networks internally learning emergent behavior?

To understand and predict emergent phase transitions, one approach is to identify hidden progress measures – metrics causally related to the loss (Barak et al., 2023) - to predict and understand grokking. Recent work found progress measures based on mechanistic interpretability by reverse engineering models into interpretable components (i.e. circuits of subnetworks) (Nanda et al., 2023). However, this approach has several limitations. First, the reliance on human effort to identify circuits prevents scaling to larger models and it is susceptible to subjective bias. Second, these measures are task-specific. Lastly, it ignores the emergent interactions between components.

This work proposes information theory as a task-independent tool to identify emergent sub-networks in neural networks for mechanistic interpretability. Pairwise mutual information has been shown to identify important features (Liu et al., 2018). However, it can only measure the dependency between two variables. Multivariate mutual information provides a fine-grained analysis of the statistical interactions between multiple variables by decomposing the dependencies into synergy and redundancy. Synergy refers to the cooperative behavior between variables as a whole, where their combined statistical interactions exceed the sum of their contributions in isolation. Redundancy is the shared information between variables. Motivated by recent work identifying synergistic sub-networks for feature importance (Clauw et al., 2022), we study higher-order interactions between neurons to understand grokking.

In this paper, we hypothesize that grokking is a phase transition caused by the emergence of a generalizing sub-network due to the collective interactions between neurons as a whole, which cannot be quantified using pairwise metrics. To illustrate this, we study the simplest setting where grokking is observed - fully connected neural networks on modular arithmetic - as a case study for emergence. To understand grokking, we utilize the O-Information - a multivariate information theory measure that scales to multiple variables - to quantify the synergy and redundancy in a network (Rosas et al., 2019). We study two strategies that impact grokking: weight decay and high initialization of the weights (Liu et al., 2022). Our contributions are as follows :

- Using the synergy and redundancy as progress measures, we identify three key phases during training: **Feature Learning** of low-level patterns, **Emergence** of a generalizing sub-network, and **Decoupling** for compression.

- We observe that a low weight decay value results in an additional **divergent** and **delayed emergence** phase. We find that weight initialization and increasing weight decay directly leads to an emergent phase and reduces

*Equal contribution [1]Department of Data Analysis, Ghent University, Belgium [2]Physics Department, University of Bari, Italy. Correspondence to: Kenzo Clauw <kenzoclauw@gmail.com>, Daniele Marinazzo <daniele.marinazzo@ugent.be>, Sebastiano Stramaglia <sebastiano.stramaglia@uniba.it>.

the delayed generalization gap

- We present preliminary findings indicating that the sub-networks at the emergent phase may be causally related to delayed generalization.

- We show that early peaks of synergy can predict if grokking occurs

2. Related Work

Cause of Grokking Several studies attribute the cause of grokking to properties related to the difficulty of representation learning the generalizing sub-network, including delayed feature learning (Kumar et al., 2023), sparsity (Merrill et al., 2023), and compression (Liu et al.). While these properties are observed in emergent complex systems, these interpretations do not consider the relationships between components as a whole.

Progress Measures Several progress measures have been proposed based on L2 norm (Liu et al., 2022) or Fourier gap (Barak et al., 2023). However, these are mostly heuristics. One study uses pairwise mutual information for measuring progress (Tan & Huang, 2023). None of these metrics quantify higher-order interactions between neurons.

3. Methodology

Setup We study the modular addition operation $\mathbb{Z}_p$ by generating equations of the form $(a + b)\%p = c$ where $p = 97$. To reduce confounding factors, we study a simple 2-layer fully connected network with ReLU activation functions. The input $(a, b) \in \mathbb{Z}_p \times \mathbb{Z}_p$ is a vector $x \in \mathbb{R}^{2p}$ represented by the concatenation of the one-hot representations of a and b. Our model processes the input as $f(\theta; x) = W_2 Z^1$ where $Z^1 = \text{ReLU}(W_1 x + b_1) \in \mathbb{R}^{s \times n}$ is a matrix of feature vectors $z_1^1, .., z_n^1$ in the first hidden layer where $n = 250$ is the number of neurons in the layer, and each feature vector z_n^1 has a dimension equal to the number of samples s in the input data. For each experiment, we train 5 models with different random seeds using full-batch AdamW optimization with a 0.03 learning rate and 40 % of the data (Loshchilov & Hutter, 2018).

O-Information Given a collection of n random variables $\mathbf{Z} = \{X_1, .., X_n\}$, the O-Information (Rosas et al., 2019) is defined as :

$$\Omega_n(\mathbf{Z}) = (n - 2)H(\mathbf{Z}) + \sum_{j=1}^{n}[H(Z_j) - H(\mathbf{Z}\backslash Z_j)] \quad (1)$$

where H is the entropy, and $\mathbf{Z}\backslash Z_j$ is the complement of Z_j with respect to $\mathbf{Z}$. If $\Omega_n(\mathbf{Z}) > 0$ the system is redundancy-dominated, while if $\Omega_n(\mathbf{Z}) < 0$ it is synergy-dominated.

We estimate the entropy terms in 1 using Gaussian Copula transformation (Ince et al., 2017) allowing an efficient closed-form solution (see papers for details).

Quantify synergy and redundancy We first reduce the dimensions of the features $Z^1 \in \mathbb{R}^{s \times 250}$ by grouping similar feature vectors into 10 bins using agglomerative clustering based on the standard configuration in sklearn resulting in a feature matrix $\tilde{Z}^1 \in \mathbb{R}^{s \times 10}$ where each bin in $\tilde{Z}^1$ consist of a set of similar feature vectors $\tilde{z}_1^k$. To quantify the synergy and redundancy in a network, we perform an exhaustive search of each combination of multiplets (i.e. subsets of 2 to k bins where the largest combination $k = 10$) using the O-Information $\Omega_k(\tilde{z}_1^1, .., \tilde{z}_1^k)$ to find the optimal synergy (lowest) and redundancy (highest) value. In our plots, we normalize both synergy and and redundancy between 0 and 1 while inverting the synergy for comparison.

4. Evaluating grokking

In this experiment, we evaluate the impact of design choices to simulate grokking by training models according to two strategies:

Role of weight decay Figure 1 illustrates models trained with low (0.01) and high (2.0) weight decay. We observe that a low weight decay results in delayed generalization. However, increasing weight decay reduces the generalization gap and increases its sharpness.

Role of initialization To increase the validity of our findings, we investigate grokking beyond weight decay. (Liu et al., 2022) observed that scaling the initialization weights of the model with a factor $\alpha > 0$ while constraining the L2 weight norm to be constant during training can induce grokking. Here, we train a model with a high initialization factor ($\alpha = 8$) with zero weight decay. Similar to (Lyu et al., 2023), we set the initialization weights in the last layer to zero to avoid learning instabilities when using a large α. Figure 1 (right) verifies that this strategy reduces the generalization gap while increasing its sharpness.
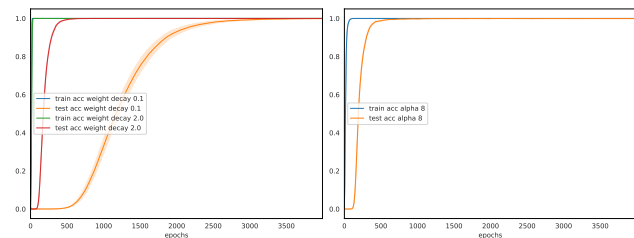

Figure 1. Left: accuracy for weight decay 0.1 and 2.0, Right: accuracy for alpha initialization 8.

5. Grokking is an emergent phase transition

To understand grokking, we plot the normalized synergy and redundancy with the loss, as this reflects the training dynamics, to measure the model's progress during training. Each measure is on a log scale for visualization purposes. We include Pareto plots to illustrate the trade-off between synergy and redundancy as this provides a more fine-grained analysis of the progress measures.

5.1. Low weight decay delays the emergence phase

We first study the most common setting of grokking for a model trained with low weight decay of 0.1. From figure 1 (left) we observe that training can be classified into 5 distinct phases:

Feature Learning We initially observed low synergy and high redundancy. The lack of coupling between features suggests the network is independently learning features, while the high redundancy indicates it is extracting basic patterns with similar properties.

Emergence During this phase, the Pareto front in figure 2 (right) reveals that the model rapidly trades off redundancy for synergy. At the same time, the size of the synergistic sub-network increases, and a peak is observed in the test loss. This suggests a critical phase transition from memorization where the model attempts to combine features to emerge a generalizing synergistic sub-network that is growing as it explores.

Divergence For low weight decay, the emergent phase does not directly lead to a generalizing solution. During a divergent phase, both the synergy and redundancy of the model drop, while the synergistic sub-network decreases in size with a small peak in the test loss. This indicates the model is overfitting due to a lack of sufficient features to combine. We hypothesize this is due to a low weight decay resulting in a difficult loss landscape making it difficult to learn generalizing.

Delayed Emergence During this phase, we observe a rapid increase in synergy, redundancy, and the size of the synergistic sub-network. This suggests the model escaped the sub-optimal solution and is now able to combine features to form a generalizing solution.

Decoupling In this phase, synergy decreases, redundancy increases, and test accuracy increases. This suggests the model has found a generalizing sub-network but not all features are relevant so it removes coupled interactions for compression.
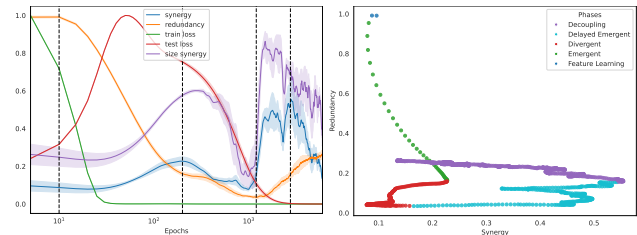

Figure 2. Baseline model with weight decay 0.1

5.2. Increasing weight decay enhances emergence

From Figure 3 (left), we observe that increasing weight decay directly results in an emergent phase with a larger peak in synergy and size of the synergistic sub-network. Moreover, the decoupling phase occurs immediately after this phase, with no divergent and delayed emergence phase.

This suggests that weight decay acts as regularization either by reducing the capacity of the network, which promotes learning shared features that are more robust by reducing the number of active features and thereby decreasing overfitting, or by smoothing the loss landscape, making it easier for optimization to find regions where more features can be combined.

We additionally observe a finalizing phase with minimal synergy and redundancy changes after test loss convergence. We attribute this to compression of the representation which is interesting for transfer learning but is not explored in this work.
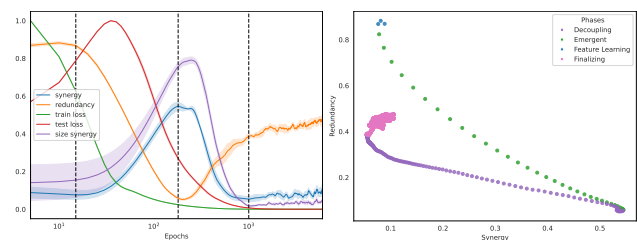

Figure 3. Baseline model with weight decay 2.0

5.3. The role of weight initialization

To isolate the contribution of weight decay regularization, we evaluate a model trained without weight decay using high weight initialization and a constrained norm. Figure 4 (left) illustrates that a model trained with high weight initialization without weight decay directly results in an emergent phase. However, from Figure 4 (right) we initially observe a rapid drop in redundancy followed by a delayed rapid increase in synergy. We hypothesize that this observation is due to the high convexity of the loss landscape making it

easier for optimization to find a solution (Fort & Scherlis, 2019; Vysogorets et al., 2024).
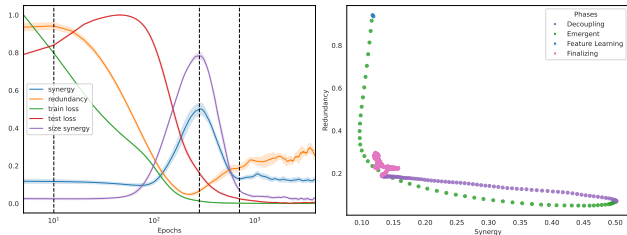

Figure 4. Baseline model with alpha = 8 and weight decay = 0

6. Early synergy peaks predict grokking

In this experiment, we plot the synergy for a variety of models trained with weight decay and weight initialization. We empirically provide evidence that the synergy can predict grokking if there is a small peak early in training. For comparison, we additionally train the models with the same parameters from section 4 but with higher values for both weight decay and high weight initialization.

Weight Decay We additionally train models with weight decay values of 10 and 50, which do not achieve grokking (see figure 9 in the appendix). From figure (left) we observe that both low (0.1) and higher (2) weight decay have a peak prior to grokking occurring. However, if weight decay is too high (i.e. 10 and 50) then the model is not able to grok indicated by lower synergy.

Initialization We additionally train models with weight decay values of 10 and 50, which do not achieve grokking (see figure 10 in the appendix). From figure (right) we observe that high weight initialization (8) results in a synergistic peak. However, if alpha is too low (1) or too high (5) then the synergy rapidly drops and the model never recovers.

These results indicate that early synergy peaks might be indicative of grokking. However, further experimentation with a variety of models and statistical tests is necessary to verify this hypothesis.
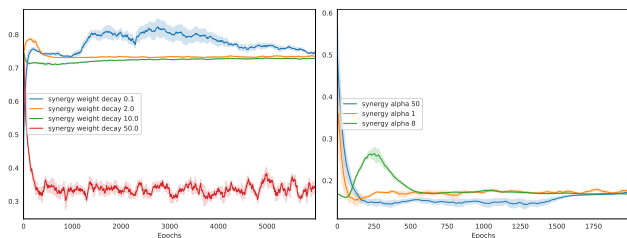

Figure 5. Left: synergy for weight decay (0.1, 2, 10, 50) Right: synergy for alpha (1, 8, 50)

7. Are the emergent synergistic sub-networks causally related to generalization?

This experiment investigates whether sub-networks with high synergy during the emergence phase contribute to delayed generalization. We focus on models trained with high weight decay (2.0) and weight initialization (8).

First, we identify and extract sub-networks with high synergy during the emergence phase. We then train these sub-networks in isolation, setting all other neurons to zero. We compare the test accuracy of these isolated sub-networks with both the original test accuracy and the accuracy of their inverse sub-networks. For comparison, we also evaluate sub-networks identified during the delayed emergence phase of a model trained with low weight decay (0.1).

Role of weight decay From figure 6 left, we find that the synergistic network for low weight decay 0.1 achieves similar performance to the original model but the peak in test accuracy is less sudden indicating the model relies less on emergent interactions. On the other hand, the inverse model

In contrast, figure 6 right shows that synergistic sub-network for weight decay 2.0 groks in fewer epochs. We hypothesize this is due to an emergent phase as indicated by a sharper test accuracy peak and higher synergy.
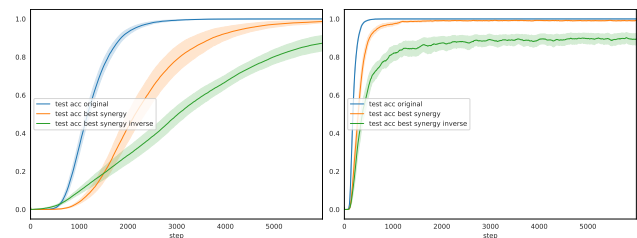

Figure 6. Left: test accuracy for synergistic sub-network with weight decay 0.1, Right: test accuracy for synergistic sub-network with weight decay 2.0

Role of initialization Figure 7 shows that weight initialization 8 results in a similar grokking period than the original model. It should be noted that we observe similar results for the models with an emergent phase. These findings strengthen our belief in synergistic interactions and suggest that grokking may result from improper hyperparameter tuning.

These preliminary findings indicate that the synergistic sub-networks are causally related to generalization. However, for the models with an emergent phase, the contrast between the synergistic sub-networks and its inverse is not that large. We argue that alternative methods are needed to provide a more accurate estimate of the clusters to identify the sub-networks.
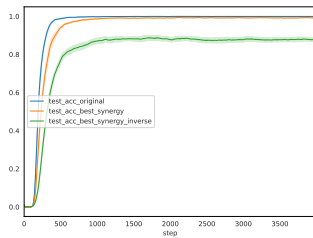
Figure 7. test accuracy for synergistic sub-network with alpha 8

8. Limitations

A key limitation of our work is the simplicity of the model and benchmark. To validate the generality of the emergent phase transitions, we aim to extend this work to 1) realistic architectures based on transformers (Nanda et al., 2023), or toy encoder-decoder models (Doshi et al., 2023) if scalability of our method is an issue. This allows us to disentangle the memorizing from the generalizing solution to better understand if these models learn isolated sub-networks, 2) evaluate alternative toy benchmarks, including sparse parity (Barak et al., 2023), XOR (Xu et al., 2023), regression (Kumar et al., 2023), realistic benchmarks like MNIST to study the impact of representation learning (Liu et al., 2022) and settings that can control grokking to increase the variety of models (Miller et al., 2024), 3) scaling the width and depth of our models. We hypothesize this might provide us insights if the relationship between grokking and double descent can be explained via synergy and redundancy (Huang et al., 2024; Nakkiran et al., 2021), and 4) explore the role of optimization using gradient descent as our results might be influenced due to the internal regularization of AdamW (Loshchilov & Hutter, 2018).

Another limitation is the difficulty of scaling the O-Information to larger networks. Our method relies on agglomerative clustering to group similar features into bins which may not fully capture the structure between features. We aim to explore alternative similarity measures such as spectral clustering (Hod et al., 2022) and the Hessian of the loss (Lange et al., 2022).

9. Conclusion and Discussion

In this work, we present - to our knowledge - the first method to quantify emergent properties in neural networks for explaining grokking without relying on supervision, heuristics, or task-specific quantities. Our findings reveal an emergent phase transition where the model collectively combines features to solve arithmetic tasks. We illustrate that synergy peaks early during training could predict grokking.

Through this preliminary work, we advocate for a paradigm shift in interpretability research from decomposing networks into the sum of isolated components (reductionism) to recognizing the emergent properties that arise from interactions between these components (emergentism). We emphasize the importance of developing methods that can quantify these higher-order interactions, as they are crucial for understanding training dynamics, generalization, and interpretability. As this is ongoing work, we conclude with future research directions.

Loss Geometry Our experiments illustrated that grokking might be caused by the difficulty of optimization to navigate the loss landscapes. We plan on further exploring this by visualizing these landscapes for each phase (Li et al., 2018) and exploring the connectivity between optima (Garipov et al., 2018). This might give insights on geometric properties like sharp minima giving rise to generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017).

Synergistic Dropout Is the synergy between neurons a necessary and/or sufficient condition for grokking? To study if synergy causes grokking we aim to regularize for the synergy via dropout of the synergistic subnetworks (Srivastava et al., 2014).

Phase Transitions Recent work observed distinct phases during optimization when training neural networks (Kalra & Barkeshli, 2023) related to the edge of stability (Cohen et al., 2020). We aim to investigate if these phases can be related via the training dynamics to the emergent phases in this work. This might explain why previous work on grokking discovered oscillation in the loss (Notsawo et al., 2023).

10. Author Contributions

Kenzo carried out the experiments and wrote the paper. Daniele and Sebastiano supervised the project.

References

Barak, B., Edelman, B. L., Goel, S., Kakade, S., Malach, E., and Zhang, C. Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit, January 2023. URL http://arxiv.org/abs/2207.08799. arXiv:2207.08799 [cs, math, stat].

Clauw, K., Stramaglia, S., and Marinazzo, D. Higher-order mutual information reveals synergistic sub-networks for multi-neuron importance, November 2022. URL http://arxiv.org/abs/2211.00416. arXiv:2211.00416 [cs, math, q-bio].

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Confer-*

ence on Learning Representations, 2020. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Doshi, D., Das, A., He, T., and Gromov, A. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets, October 2023. URL http://arxiv.org/abs/2310.13061. arXiv:2310.13061 [cond-mat, stat].

Fort, S. and Scherlis, A. The Goldilocks Zone: Towards Better Understanding of Neural Network Loss Landscapes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3574–3581, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33013574. URL https://ojs.aaai.org/index.php/AAAI/article/view/4237.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, January 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1.

Hod, S., Filan, D., Casper, S., Critch, A., and Russell, S. Quantifying Local Specialization in Deep Neural Networks, February 2022. URL http://arxiv.org/abs/2110.08058. arXiv:2110.08058 [cs].

Huang, Y., Hu, S., Han, X., Liu, Z., and Sun, M. Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition, February 2024. URL http://arxiv.org/abs/2402.15175. arXiv:2402.15175 [cs].

Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., and Schyns, P. G. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping*, 38(3):1541–1573, March 2017. ISSN 1065-9471, 1097-0193. doi: 10.1002/hbm.23471. URL https://onlinelibrary.wiley.com/doi/10.1002/hbm.23471.

Kalra, D. S. and Barkeshli, M. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width, October 2023. URL http://arxiv.org/abs/2302.12250. arXiv:2302.12250 [cond-mat].

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, February 2017. URL http://arxiv.org/abs/1609.04836. arXiv:1609.04836 [cs, math].

Kumar, T., Bordelon, B., Gershman, S. J., and Pehlevan, C. Grokking as the Transition from Lazy to Rich Training Dynamics, October 2023. URL http://arxiv.org/abs/2310.06110. arXiv:2310.06110 [cond-mat, stat].

Lange, R. D., Rolnick, D. S., and Kording, K. P. Clustering units in neural networks: upstream vs downstream information, March 2022. URL http://arxiv.org/abs/2203.11815. arXiv:2203.11815 [cs, stat].

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

Liu, K., Amjad, R. A., and Geiger, B. C. Understanding individual neuron importance using information theory. *arXiv preprint arXiv:1804.06679*, 19:5171–5180, 2018. URL https://www.researchgate.net/profile/Rana-Ali-Amjad/publication/324600151_Understanding_Individual_Neuron_Importance_Using_Information_Theory/links/5b7ab84d4585151fd12239c1/Understanding-Individual-Neuron-Importance-Using-pdf.

Liu, Z., Zhong, Z., and Tegmark, M. Grokking as Compression: A Nonlinear Complexity Perspective.

Liu, Z., Michaud, E. J., and Tegmark, M. Omnigrok: Grokking Beyond Algorithmic Data. *arXiv preprint arXiv:2210.01117*, 2022.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. 2018. URL https://openreview.net/forum?id=SJs7uYYeM.

Lyu, K., Jin, J., Li, Z., Du, S. S., Lee, J. D., and Hu, W. Dichotomy of Early and Late Phase Implicit Biases Can Provably Induce Grokking, November 2023. URL http://arxiv.org/abs/2311.18817. arXiv:2311.18817 [cs].

Merrill, W., Tsilivis, N., and Shukla, A. A Tale of Two Circuits: Grokking as Competition of Sparse and Dense Subnetworks. *arXiv preprint arXiv:2303.11873*, 2023.

Miller, J., O'Neill, C., and Bui, T. Grokking Beyond Neural Networks: An Empirical Exploration with Model Complexity, March 2024. URL http://arxiv.org/abs/2310.17247. arXiv:2310.17247 [cs, stat].

Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. Optimal Regularization Can Mitigate Double Descent, April 2021. URL http://arxiv.org/abs/2003.01897. arXiv:2003.01897 [cs, math, stat].

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability, January 2023. URL http://arxiv.org/abs/2301.05217. arXiv:2301.05217 [cs].

Notsawo, P. J. T., Zhou, H., Pezeshki, M., Rish, I., and Dumas, G. Predicting Grokking Long Before it Happens: A look into the loss landscape of models which grok, June 2023. URL http://arxiv.org/abs/2306.13253. arXiv:2306.13253 [cs].

Rosas, F. E., Mediano, P. A. M., Gastpar, M., and Jensen, H. J. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical review E*, 100(3-1):032305, September 2019. ISSN 2470-0053. doi: 10.1103/physreve.100.032305. URL https://doi.org/10.1103/PhysRevE.100.032305.

Schaeffer, R., Miranda, B., and Koyejo, S. Are Emergent Abilities of Large Language Models a Mirage?

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. URL https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer,. Publisher: JMLR. org.

Tan, Z. and Huang, W. Understanding Grokking Through A Robustness Viewpoint, November 2023. URL http://arxiv.org/abs/2311.06597. arXiv:2311.06597 [cs].

Vysogorets, A., Dawid, A., and Kempe, J. Deconstructing the Goldilocks Zone of Neural Network Initialization, February 2024. URL http://arxiv.org/abs/2402.03579. arXiv:2402.03579 [cs, math].

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent Abilities of Large Language Models, October 2022. URL http://arxiv.org/abs/2206.07682. arXiv:2206.07682 [cs].

Xu, Z., Wang, Y., Frei, S., Vardi, G., and Hu, W. Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data, October 2023. URL http://arxiv.org/abs/2310.02541. arXiv:2310.02541 [cs, stat].

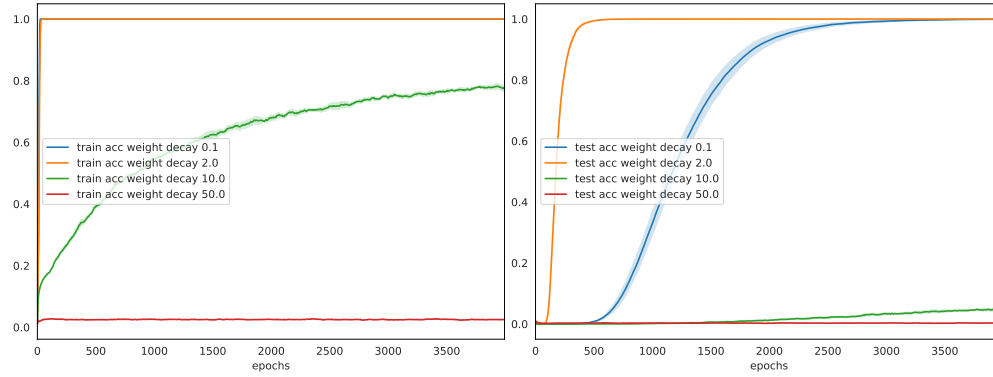## A. Training results for baseline model with weight decay

Figure 8. Left: train accuracy for weight decay (0.1, 2, 10, 50) Right: test accuracy for weight decay (0.1, 2, 10, 50)

B. Training results for baseline model with weight initialization alpha
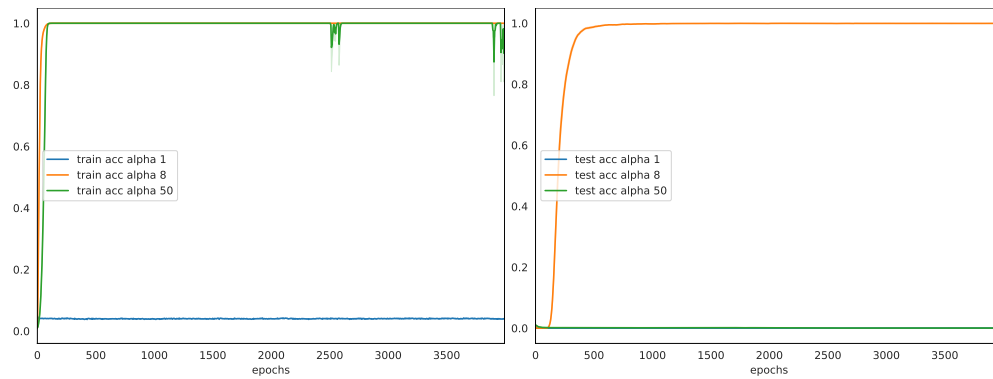

Figure 9. Left: train accuracy for alpha (1, 8, 50) Right: test accuracy for alpha (1, 8, 50)