Relating Regularization and Generalization through the Intrinsic Dimension of Activations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Given a pair of models with similar training set performance, it is natural to assume that the model that possesses simpler internal representations would exhibit better generalization. In this work, we provide empirical evidence for this intuition through an analysis of the intrinsic dimension (ID) of model activations, which can be thought of as the minimal number of factors of variation in the model's representation of the data. First, we show that common regularization techniques uniformly decrease the last-layer ID (LLID) of validation set activations for image classification models and show how this strongly affects model generalization performance. We also investigate how excessive regularization decreases a model's ability to extract features from data in earlier layers, leading to a negative effect on validation accuracy even while LLID continues to decrease and training accuracy remains near-perfect. Finally, we examine the LLID over the course of training of models that exhibit grokking. We observe that well after training accuracy saturates, when models "grok" and validation accuracy suddenly improves from random to perfect, there is a co-occurent sudden drop in LLID, thus providing more insight into the dynamics of sudden generalization.

1 Introduction

Occam's razor states that given a set of plausible explanations, the simplest one should be preferred. The benefits of applying this idea to machine learning models are intuitive: any method whose decision criteria are too complex can learn to overfit to noise and biases in the data that are unrelated to the problem at hand. In contrast, methods that are able to fit the observed data with the fewest number of assumptions are more likely to have found generalizable patterns in the data applicable to unseen examples. The successful application of this idea can be seen throughout machine learning (e.g. [35, 1, 5, 29, 11, 10]). Perhaps most notably, Bayesian inference itself can be seen through the lens of Occam's razor [26], with the likelihood and prior encouraging plausibility and simplicity, respectively.

Our work can be seen as another application of Occam's razor, where we analyze how the simplicity and plausibility of learned neural network representations are affected by regularization and correlated with generalization. Specifically, we elucidate this idea using the intrinsic dimension (ID) of the validation data's activations, which can be thought of as the minimal number of factors of variation in their representation. We introduce two measures of representational compexity for neural networks: last-layer ID (LLID) and peak ID (PID). LLID is the intrinsic dimension of the validation set's activations at the last layer of activations and PID is the maximum intrinsic dimension of the validation set's activations over all layers of the model (which for all experiments is always at early layers of the network). As we will show, under the framework of Occam's razor, a neural network is considered "simpler" if it has a low LLID and is "plausible" if training accuracy and PID are sufficiently high.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Given the link between representational simplicity and generalization, the success of vastly over-parameterized neural networks [38] seems counterintuitive. We highlight two explanations for this apparent contradiction. Firstly, Ansuini et al. [2] and Recanatesi et al. [29] show that despite having a large ambient space, activations live on orders of magnitude lower dimensional manifolds. Secondly, although DNNs are trained to maximize performance on the training set, they are also regularized. Regularization techniques bias the model towards simpler and more general solutions, thus reducing overfitting. In this work we unify these two explanations, showing that common regularization techniques decrease the dimensionality of activation manifolds deeper in the network and allow image classification models to better generalize to the validation set. Nevertheless, one cannot arbitrarily decrease LLID through regularization and expect a guaranteed improvement in generalization. We hypothesize that this is partly because regularization also decreases the ID of initial layers in the network, and thus decrease PID, inhibiting sufficiently complex/varied features from being learned. We show that achieving good generalization requires a trade-off between decreasing the LLID and keeping the PID sufficiently high.

To strengthen our understanding of the relationship between simplicity of representation and generalization, we also study a task where models are able to learn to perfectly generalize. Specifically, we analyze the internal representations of small transformer models trained to perform modular division. Power et al. [28] found surprising behaviour on these tasks where well after training accuracy saturates at 100%, models are able to "grok" the problem and suddenly increase their validation performance from random guessing to 100%. Remarkably, we find a co-occurrent decrease in LLID when the model groks. This suggests that there is a strong relationship between finding generalizable patterns in the data and the model's simplicity of representation, further confirming the relevance of measuring LLID.

We hope that our findings linking widespread regularization techniques to measurable proxies for simplicity motivate the community to continue to study the implications of LLID and PID both as indicators of improved generalization and targets for optimization. Specifically, disentangling the effect of regularization techniques invites improved algorithms to enhance and diminish the positive and negative consequences of current techniques respectively. We will release the code publicly upon acceptance to facilitate this future research.

2 Experiments

Despite living in high dimensional space, activations in neural networks have been shown to live on manifolds of much lower intrinsic dimension [2, 29]. In this work, we leverage this to introduce two measures of representational complexity: the intrinsic dimension of the validation set's last layer of activations (LLID) and the maximum intrinsic dimension of the validation set's activations over all layers in the network (PID). In this section, we analyze the implications of LLID and PID in image classification (Section 2.1) and transfomer models (Section 2.2) using the TwoNN estimator of Facco et al. [13]. We include a more thorough discussion of intrinsic dimension and related work in Appendix A.

2.1 Regularization and Generalization in Image Classification

Setup: To analyze the relationship between LLID and regularization we train ResNet-18 models [16] on the CIFAR-10 and CIFAR-100 [19] image classification datasets while independently sweeping over values of dropout and weight decay [31, 20]. Training details can be found in Appendix B. The LLID for these models is defined on the activations after the average pooling layer, and the PID is consistently on the activations after the first ResNet block. We note that all models trained achieve perfect accuracy on the training set.

Regularization decreases LLID: In Figure 1, we plot the final LLID for models with different amounts of regularization trained to convergence. Clearly, increasing the strength of the regularization (dropout probability or weight decay coefficient) decreases LLID as confirmed by the Pearson's correlation coefficients, r, and p-values for independence in a t-test shown in the figure titles. This strong relationship confirms the intuition that regularization decreases the complexity of neural representations.

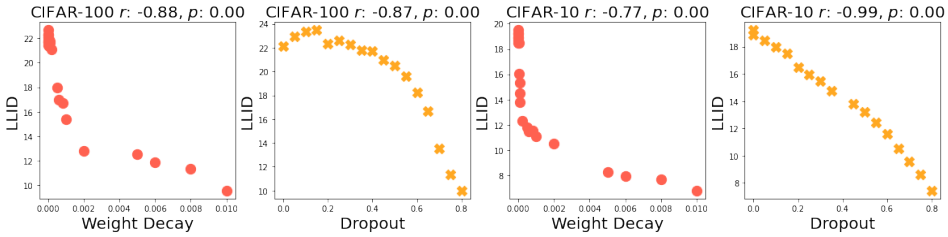

Figure 1: LLID for ResNet-18 models on CIFAR-100/CIFAR-10 datasets with varying values of weight decay and dropout percentage. Increasing regularization strength decreases LLID.

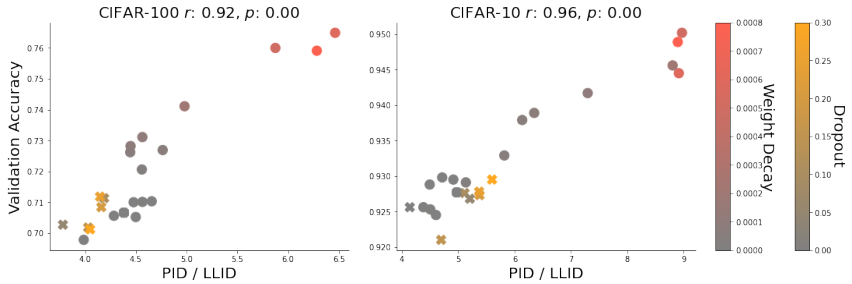

Figure 2: Relationship between PID / LLID for models trained with various values of dropout (indicated with X markers) and weight decay (indicated with circle markers). Increased intensity of red and yellow color denotes increased weight decay value and dropout percentage respectively.

Generalization is correlated with PID and LLID: We have shown that decreasing LLID is one of the effects of regularization that is shared between both methods studied. However, different regularization techniques clearly have different consequences that may either aid or harm generalization and adjusting the strength of regularization requires a trade-off between these effects. We study one such negative impact that we observe in our models – decreasing PID – and leave a more detailed study of the other impacts of regularization to future work.

Figures 5 and 6 in Appendix E show that achieving higher validation accuracy is correlated with decreasing LLID and increasing PID. This indicates that although it is advantageous to have a simple representation of the data at deeper layers, it should not come at the expense of the ability to extract rich features in earlier layers. Figure 2 highlights this trade-off, showing the correlation between validation accuracy and the ratio between PID and LLID (i.e. how much the model reduces the ID of activations across layers). There is a clear trend where a higher PID/LLID ratio indicates better validation performance, highlighting the necessity of balancing the effect of decreasing the complexity of the last layer of activations (which are a single linear layer away from the logits and thus form the basis for the model's ultimate classification) without penalizing the model for extracting the potentially complex features in earlier layers required to form this representation.

The relevance of PID becomes more dramatic as the strength of regularization is pushed beyond normal bounds. In Figure 4 of Appendix E, we plot how PID changes with excessive regularization. We see that soon after regularization is increased beyond the value that gives optimal accuracy (indicated by the red line), there is a rapid decrease in both PID and accuracy. This is particularly true for the more challenging CIFAR-100 dataset where we hypothesize that the increase in the number of classes requires more varied features to be learned in earlier layers. Using the framework of Occam's Razor, we see that excessive regularization prevents "plausibility", which PID is a strong indicator of.

2.2 LLID and Grokking

Setup: We last explore the relationship between LLID and the grokking phenomenon observed in Power et al. [28]. Grokking characterizes a sudden rise in validation performance long after the model has perfectly memorized the training dataset. This effect was specifically observed on a small algorithmic dataset using a two-layer transformer model [36]. We reproduce this training setup

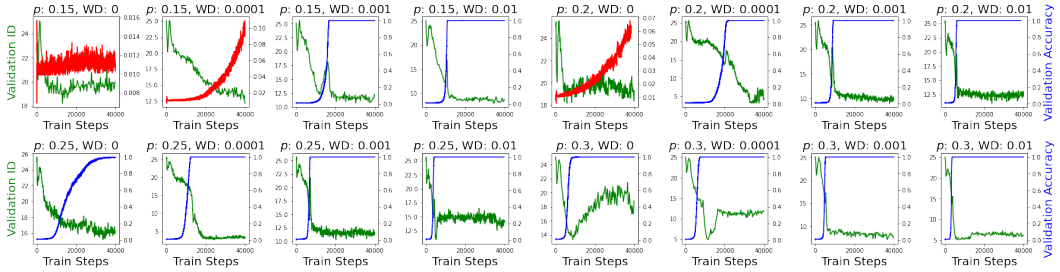

Figure 3: Validation accuracy and LLID throughout training of transformer models on modular division tasks (plot titles show percentage of training data used, p, and weight decay coefficient, WD). We observe co-occurent ID drops when transformers 'grok'. Runs with models that don't achieve perfect validation accuracy in the allotted compute budget of 40k training steps are shown in red.

while monitoring training and validation set LLID over the course of training. We train 16 different models, sweeping over the strength of weight decay regularization and the size of the training dataset, since these parameters strongly affect when models begin to grok and the duration of the grokking period. For these two-layer transformer models, we define the LLID as the intrinsic dimension of the activations after the second transformer block's MLP (we also measure the ID of activations after the first transformer block and find similar patterns which we hypothesize is due to the small network size). More training details can be found in Appendix B.

LLID drops co-occur with grokking: As shown in Figure 3 (we show the same plot alongside the training ID and accuracy of each run in Figure 7 of Appendix E), we observe a sudden drop in LLID co-occurent with the sudden rise in validation accuracy. Notably, models that grok have an average minimum LLID of 8.53, much lower than the average minimum LLID – 16.16 – of models that remain at a low validation accuracy in the allotted compute budget, once again confirming the relevance of the simplicity of representations for generalization as measured by LLID. Curiously, we also observe that for lower amounts of training data when grokking occurs later in training (ex. plot with $p = 0.15$,weight decay$= 0.01$ in Figure 3), there is an initial LLID spike when the model begins grokking followed by a descent into a lower LLID than the previous plateau. This hints at an escape from a local minima, and the onset of this phenomena at higher weight decay values indicates that regularization helps with this jump. This relationship between LLID and grokking is exciting, as it both confirms the validity of measuring simplicity using LLID and shows that it is directly correlated with generalization. Moreover, optimizing for LLID with the aim of replicating this increase in generalization is an exciting direction for future optimization and regularization methods.

3 Conclusion

In this work we explored the link between internal representation simplicity and generalization performance by analyzing the intrinsic dimension of model activations. We identified a strong correlation between increasing amounts of regularization with decreased LLID in image classification models. We also demonstrated that decreased LLID and increased PID correlates with improved generalization, until a point is reached where excessive regularization reduces the capability of models to extract meaningful features in earlier layers. Lastly, we established a co-occurence between a sudden rise in validation accuracy (i.e. grokking) and a sudden drop in LLID on small models trained on algorithmic datasets. We hope that our work motivates the use of ID (and other metrics for internal representation simplicity) as indicators of a model's capability for generalization. We are specifically interested in future work that analyzes the ID of self-supervised language and vision models which are able to quickly adapt to downstream tasks unseen during training. Additionally, we hope that our analysis of the undesirable consequences of excessive regularization can motivate the development of novel regularization techniques which can promote simpler internal representations in later model layers without diminishing a model's ability to extract useful features in earlier layers.

References

[1] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. 2016. doi: 10.48550/ARXIV.1612.00410. URL https://arxiv.org/abs/1612.00410.

[2] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan. Intrinsic dimension of data representations in deep neural networks. *CoRR*, abs/1905.12784, 2019. URL http://arxiv.org/abs/1905.12784.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives, 2012. URL https://arxiv.org/abs/1206.5538.

[4] R. Bhalodia, S. Elhabian, L. Kavan, and R. Whitaker. Coopsubnet: Cooperating subnetwork for data-driven regularization of deep networks under limited training budgets, 2019. URL https://arxiv.org/abs/1906.05441.

[5] T. Birdal, A. Lou, L. Guibas, and U. Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks, 2021. URL https://arxiv.org/abs/2111.13171.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[7] L. Blier and Y. Ollivier. The description length of deep learning models, 2018. URL https://arxiv.org/abs/1802.07044.

[8] A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin. Intrinsic dimension estimation using wasserstein distances, 2021. URL https://arxiv.org/abs/2106.04018.

[9] B. C. A. Brown, A. L. Caterini, B. L. Ross, J. C. Cresswell, and G. Loaiza-Ganem. The union of manifolds hypothesis and its implications for deep generative modelling, 2022. URL https://arxiv.org/abs/2207.02862.

[10] C. Chen, X. Ni, Q. Bai, and Y. Wang. A topological regularizer for classifiers via persistent homology, 2018. URL https://arxiv.org/abs/1806.10714.

[11] V. D'Amario, S. Srivastava, T. Sasaki, and X. Boix. The foes of neural network's data efficiency among unnecessary input dimensions, 2021. URL https://arxiv.org/abs/2107.06409.

[12] B. Dherin, M. Munn, M. C. Rosca, and D. G. T. Barrett. Why neural networks find simple solutions: the many regularizers of geometric complexity, 2022. URL https://arxiv.org/abs/2209.13083.

[13] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, Sep 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y. URL https://doi.org/10.1038/s41598-017-11873-y.

[14] T. Gao and V. Jojic. Degrees of freedom in deep neural networks, 2016. URL https://arxiv.org/abs/1603.09260.

[15] M. Gomtsyan, N. Mokrov, M. Panov, and Y. Yanovich. Geometry-aware maximum likelihood estimation of intrinsic dimension, 2019. URL https://arxiv.org/abs/1904.06151.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[17] D. Kienitz, E. Komendantskaya, and M. Lones. The effect of manifold entanglement and intrinsic dimensionality on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7160–7167, Jun. 2022. doi: 10.1609/aaai.v36i7.20676. URL https://ojs.aaai.org/index.php/AAAI/article/view/20676.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

[19] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] A. Krogh and J. Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf.

[21] J. Kukačka, V. Golkov, and D. Cremers. Regularization for deep learning: A taxonomy, 2017. URL https://arxiv.org/abs/1710.10686.

[22] D. H. Lee and S.-Y. Chung. Unsupervised embedding adaptation via early-stage feature reconstruction for few-shot classification, 2021. URL https://arxiv.org/abs/2106.11486.

[23] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[24] U. Lim, V. Nanda, and H. Oberhauser. Tangent space and dimension estimation with the wasserstein distance. *arXiv preprint arXiv:2110.06357*, 2021.

[25] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S.-T. Xia, S. Wijewickrema, and J. Bailey. Dimensionality-driven learning with noisy labels, 2018. URL https://arxiv.org/abs/1806.02612.

[26] D. J. MacKay, D. J. Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[27] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning, 2021. URL https://arxiv.org/abs/2104.08894.

[28] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

[29] S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown. Dimensionality compression and expansion in deep neural networks, 2019. URL https://arxiv.org/abs/1906.00443.

[30] C. Snell. unofficial re-implementation of "grokking: Generalization beyond overfitting on small algorithmic datasets". https://github.com/Sea-Snell/grokking, 2022.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[32] P. Tempczyk, R. Michaluk, L. Garncarek, P. Spurek, J. Tabor, and A. Golinski. Lidl: Local intrinsic dimension estimation using approximate likelihood. In *International Conference on Machine Learning*, pages 21205–21231. PMLR, 2022.

[33] V. Thilak, E. Littwin, S. Zhai, O. Saremi, R. Paiss, and J. Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon, 2022. URL https://arxiv.org/abs/2206.04817.

[34] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[35] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

[37] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019. URL https://arxiv.org/abs/1910.03771.

[38] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization, 2016. URL https://arxiv.org/abs/1611.03530.

[39] W. Zhu, Q. Qiu, J. Huang, R. Calderbank, G. Sapiro, and I. Daubechies. Ldmnet: Low dimensional manifold regularized neural networks, 2017. URL https://arxiv.org/abs/1711.06246.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Appendix C
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Appendix D
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see code in supplementary materials
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Appendix B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Section 5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite all the datasets we used.
   (b) Did you mention the license of the assets? [No] Because all the assests we use have public licenses allowing us to use them.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include code which, as mentioned in Section 1, will be made publicly available at publication.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] But only because all the data we use is publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Because the datasets we use are standard in machine learning and do not contain such information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Background and Related Work

Intrinsic dimension The manifold hypothesis states that despite having high ambient dimension, data of interest lives on a manifold of much lower dimension, denoted as the data's intrinsic dimension (ID) [3]. This hypothesis has been experimentally confirmed for both image data and activations in deep neural networks [27, 2, 9]. For this study, we use an estimate of the intrinsic dimension of the last layer of activations, LLID, as a proxy for the simplicity of representation of a neural network. As it is intractable to calculate the exact intrinsic dimension, we use the TwoNN estimator of Facco et al. [13] following the implementation described in Ansuini et al. [2] which assumes that the density around each datapoint is constant and maximizes the likelihood of the observed distances of every datapoint's two nearest neighbours. Although more recent ID estimators exist [15, 8, 24, 32, 23], we opt for this one as it is practically reliable [2] and computationally efficient for calculating ID over our high ambient dimension activations.

Intrinsic dimension of activations Although other measures of simplicity have been proposed [12, 14, 7], we choose intrinsic dimension as our proxy for representational complexity as previous works have found that activations live on non-linear manifolds and intrinsic dimension estimators have been insightful at revealing important properties of the geometric structure of data representations [2, 29, 22, 25, 17]. In particular: Ansuini et al. [2] show that LLID is decreased for larger networks and is correlated with generalization, Lee and Chung [22] use the ID of activations as an early stopping criteria for few shot embedding adaptation; and Ma et al. [25] use activation ID as a criterion for decreasing confidence in noisy labels to avoid overfitting.

Notably, Ansuini et al. [2] show that the ID of activation manifolds shows a hunchback shape in classification models where the intrinsic dimension quickly peaks in early layers and continues to decrease until the final activation layer. We extend this study by showing how attributes of this curve, PID and LLID, are affected by regularization and correlated with generalization, thus opening the door for more targeted regularization techniques and improved hyperparameter tuning. We further confirm the importance of these measures by demonstrating how LLID is strongly coupled with the perfect generalization phenomena of grokking.

Grokking Recent work has identified training regimes where models experience a period of rapid generalization long after the model has perfectly memorized the training dataset, also known as the grokking. Power et al. [28] discovered this phenomenon by training two-layer transformer models trained on small procedurally generated algorithmic datasets. Notably, the amount of weight decay applied during training had a large impact on the number of steps required for the model to generalize; stronger regularization caused grokking to occur significantly earlier in training. Our work finds that sudden generalization in these networks is accompanied by a corresponding sharp drop in LLID, with more strongly regularized networks experiencing this drop earlier in training.

Thilak et al. [33] further explore the dynamics of sudden generalization, identifying "slingshot effects" that occur when training the transformer models discussed above without regularization. Each slingshot is characterized by alternating period of stability and instability during training, where in each period of instability the norm of the weights in the model's last layer grows rapidly, which is accompanied by a spike in training loss. Each slingshot is followed by a boost in generalization performance. The authors find that regularized networks are able to generalize without this training instability.

Regularization Regularization has been critical to the success of deep learning methods [21, 31, 6, 34, 20]. Of particular interest to our work are regularization techniques that attempt to decrease the complexity of activations [4, 39]. Our work quantifies this decrease in complexity, demonstrates that prominent regularization techniques [31, 20] also have this effect, and calls for more targeted regularization techniques that are layer-dependent.

B Training Details

CIFAR Image Classification For our regularization sweeps, we independently sweep over weight decay with values of ae^{-b} for all combinations of $a \in \{1, 2, 5, 6, 8\}$ and $b \in \{-6, -5, -4, -3, -2\}$ as well as dropout [31, 20] in between each residual block with dropout probabilities ranging from

0.05 to 0.8 with increments of 0.05. At the end of training, we estimate the intrinsic dimension across all validation datapoints after each of the ResNet blocks (when the channel dimension changes) and final average pooling layer. To calculate the intrinsic dimension, we flatten the shape of the activations and calculate ID on this vector using the TwoNN ID estimator [13]. To train all models, we use the SGD optimizer with a momentum value of 0.9 and a learning rate of 0.1 decayed by a factor of 0.2 at epochs $60, 120$ and 060 for a total of 200 epochs. The data is augmented by applying random crops, random horizontal flips with probability of 0.5 and random rotations between -15 and 15 degrees.

Modular Division Tranformer Models To train all models, we use a 2 layer transformer decoder model with 4 heads and embedding dimension of 128 obtained from the huggingface transformers library [37, 36]. The ADAM optimizer [18] is used with beta values of $(0.9, 0.98)$ and a constant learning rate of 0.001 for $40k$ training steps. The modular division modulo 97 dataset is obtained from [30].

C Limitations

Although the trends observed support the correlations observed for LLID and PID with regularization and generalization, we only compared results on ResNet-18 and two-layer transformers models on three datasets and two regularization techniques. Future work performing the same analysis on more varied architectures, tasks and regularization strategies would strengthen the conclusions. Moreover, to assess the complexity of representations we use an intrinsic dimension estimator on the activations of a neural network. We note that this estimator is imperfect and only a proxy for the true intrinsic dimension. In particular, we note that intrinsic dimension cannot increase across neural network layers which are functions of only their previous activations, yet the ID estimates of the initial activations are often smaller than the PID (the maximum ID over all activations). We attribute this inconsistency to errors in the estimators used and since we only require that trends and ID comparisons are consistent (not exact), we find that this is nevertheless a useful estimator.

D Negative Societal Impacts and Compute Acknowledgement

Due to the nature of contributions, we do not foresee any potential negative societal consequences. For this project, we trained all models discussed using an internal cluster with a combination of Tesla V100, Quadro RTX 6000 and TITAN V GPUs.

E Additional Figures

Due to space constraints, in this section we highlight additional results.

CIFAR Results: In Figure 4 we show the effect of excessive regularization on the PID of ResNet-18 models. We note that as regularization is pushed beyond typical bounds, PID and validation accuracy decrease. In Figure 5 and 6 we further decompose the relationship shown in Figure 2 by displaying the correlation between LLID and PID with validation accuracy respectively.

Grokking Results: In Figure 7 we show a larger and more detailed version of Figure 3 including both training accuracy and estimates of the intrinsic dimension of the training set's last layer of activations.

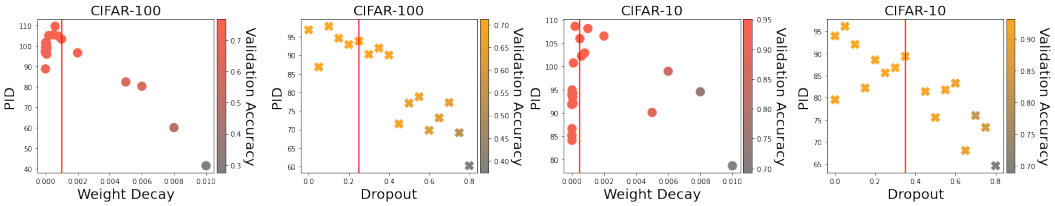

Figure 4: Relationship between PID and large values of regularization. Stronger red and yellow color indicates increased validation accuracy. Beyond a certain regularization strength, PID and validation accuracy decrease rapidly. The vertical red line in each plot indicates the regularization strength of the best performing model.

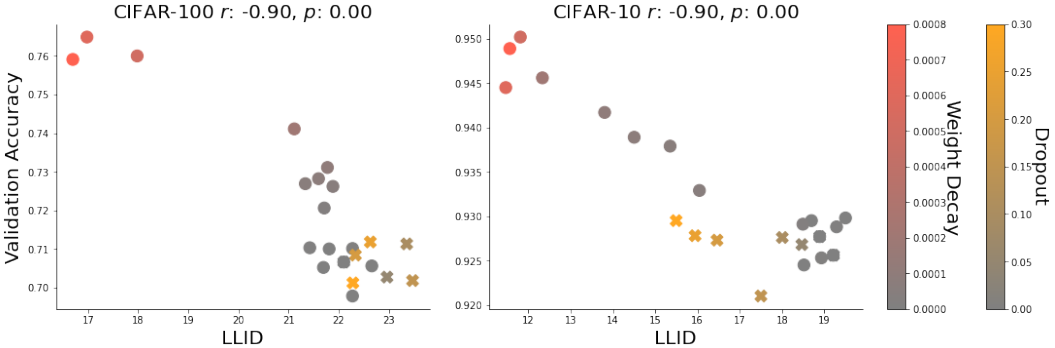

Figure 5: The correlation between LLID and validation accuracy for datasets CIFAR-100 (left) and CIFAR-10 (right) for various dropout and weight decay values.

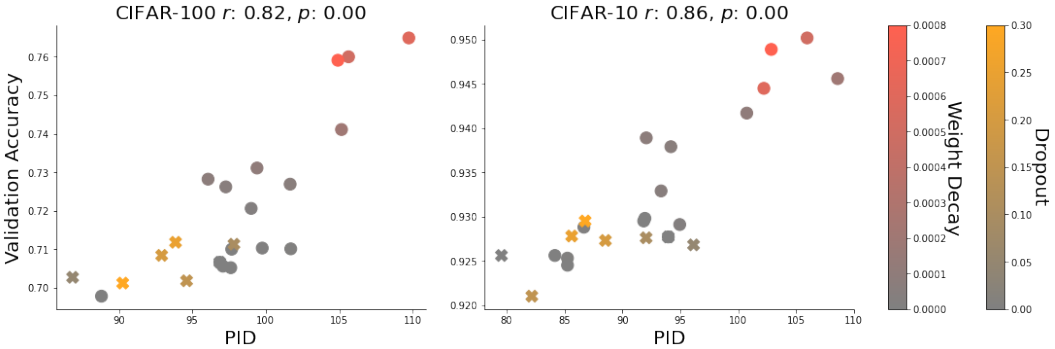

Figure 6: The correlation between PID and validation accuracy for datasets CIFAR-100 (left) and CIFAR-10 (right) for various dropout and weight decay values.

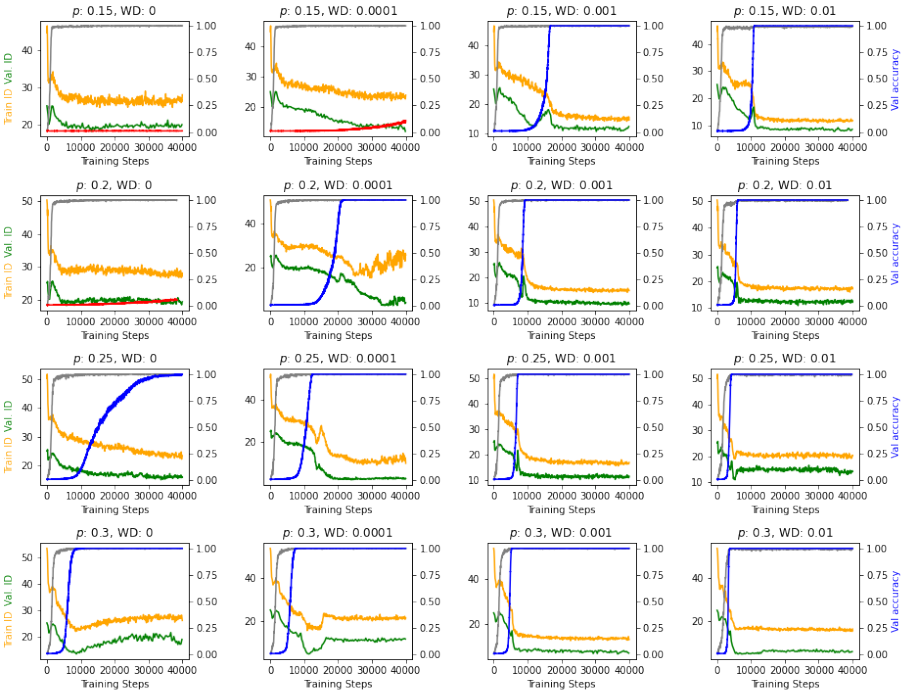

Figure 7: The values of train accuracy, validation accuracy, and LLID of transformer models throughout training on a modular division task. Train accuracy very quickly reaches 100% for all models and training LLID follows the same trends as the LLID of the validation set activations.