Early Period of Training Impacts Out-of-Distribution Generalization

Chen Cecilia Liu & Iryna Gurevych
Ubiquitous Knowledge Processing Lab, Department of Computer Science,
Hessian Center for AI (hessian.AI), Technische Universität Darmstadt

Abstract

Prior research has found that differences in the early period of neural network training significantly impact the performance of in-distribution (ID) tasks. However, trained neural networks are often sensitive to out-of-distribution (OOD) data, making them less reliable in downstream applications. Yet, the impact of the early training period on OOD generalization remains unknown due to its complexity and lack of effective analytical methodologies. In this work, we investigate the relationship between learning dynamics and OOD generalization during the early period of neural network training. We utilize the trace of Fisher Information and sharpness, focusing on gradual unfreezing (i.e., progressively unfreezing parameters during training) as the methodology for investigation. Through a series of empirical experiments, we show that 1) changing the number of trainable parameters at certain times (via gradual unfreezing) has minor impacts on ID results, but significantly improves OOD results; 2) the absolute values of sharpness and trace of Fisher Information at the initial period of training are not indicative for OOD generalization, a higher sharpness may be beneficial for OOD generalization in this period.

1. Introduction

Deep neural networks have achieved impressive results in the tasks they are trained on, but they are often sensitive to distribution shifts (i.e., out-of-distribution, OOD) during inference. As in many applications of deep neural networks, the training and testing data may come from different distributions. Failure to generalize to the OOD setting degrades the models' robustness and reliability.

A plethora of research has found that differences in the early period of training have a significant impact on the in-distribution (ID) performance [1, 8, 10, 24, inter alia] for a wide range of settings. The wide observation of such a period in machine learning applications suggests that the early period of learning is generally important for neural network training [18].

In particular, prior literature identifies that modifications (or interventions) to the optimization process are critical in shaping the early period of training for ID generalization. Training techniques such as weight decay [10], learning rates [15, 24], data augmentation ([10, 23], such as with MixUp, [28]) or adding noise to weights [9] impact learning dynamics early on, and can significantly improve or hamper the final task results depending on the time of application or removal. However, to the best of our knowledge, there haven't been studies on how the early period of training impacts OOD generalization. Such a study could lead to the development of new theories for OOD generalization and new methods for improving model performance in OOD scenarios.

[22] found that gradual unfreezing [13] (i.e., progressively releasing trainable parameters during training at fixed time intervals) impacts the trace of Fisher Information ($\mathtt{tr}(\mathrm{F})$) in the early period of training. However, the work was focused on which parameters to select for better cross-lingual transfer (where cross-lingual transfer is a form of natural OOD generalization). Besides the $\mathtt{tr}(\mathrm{F})$,

other sharpness metrics have also been used to study the generalization of network training, especially after the success of methods such as Sharpness-Aware Minimization (SAM, [7, 20, 29]), however, sharpness is less used in prior work to study the early period of training.

In this work, we conduct empirical investigations into how the early period of training impacts OOD generalization (with a focus on covariate/input shift). We first use gradual unfreezing [13] to intervene in the dynamics of the early period of training from scratch, examining its impact on OOD generalization. Next, we investigate different metrics for studying the early period of training for OOD generalization. To summarize, we show that 1) changing the number of trainable parameters at certain times (via gradual unfreezing) has a minor impact on ID results, but significantly improves OOD results; 2) the absolute values of sharpness and $\texttt{tr}(\mathrm{F})$ at the initial period of training are not indicative for OOD generalization, a higher sharpness may be beneficial for OOD generalization in this period.

2. Preliminaries

2.1. Fisher Information Matrix (FIM)

To investigate the training process, we first look at the Fisher Information [6]. Fisher Information reflects the local curvature and measures the amount of information with respect to network parameters, i.e., how sensitive the network predictions are to the changes in parameters. A larger Fisher Information indicates that a small change in network parameters can change the output significantly, which can be interpreted as a "sharper" loss landscape.

Estimating the full Fisher Information is generally expensive. Prior work shows that the trace of the Fisher Information ($\texttt{tr}(\mathrm{F})$) correlates well with the full Fisher Information when used in real applications to capture signals during the learning process [1, 15, 27, inter alia]. Let x be the inputs and y be the labels of a dataset D. Given a neural network that is parameterized by w, and using the empirical data distribution $\hat{Q}(x)$, the Fisher Information is defined as:

$$\texttt{tr}(\mathrm{F}) = \mathbb{E}_{x\sim\hat{Q}(x)}\, \mathbb{E}_{\hat{y}\sim p_w(\hat{y}|x)}\, ||\nabla_w \log p_w(\hat{y}|x)||^2. \tag{1}$$

2.2. Sharpness

Let $\mathcal{L}_{\mathcal{D}}(w) = \frac{1}{|D|}\sum_{(x,y)\in D} \log p_w(y|x)$ be the loss over training datasets of a neural network parameterized by w, and δ be a small perturbation drawn from a noise distribution, such as a Gaussian distribution $\mathcal{N}(0, \rho^2 diag(c^2))$. The definitions of average and worst-case sharpness are [2, 7, 12, 20]:

$$S^{\rho}_{avg} = \mathbb{E}_{\delta\sim\mathcal{N}(0,\rho^2 diag(c^2))}\, \mathcal{L}_{\mathcal{D}}(w-\delta) - \mathcal{L}_{\mathcal{D}}(w), \tag{2}$$

$$S^{\rho}_{worst} = \max_{\|\delta\odot c^{-1}\|_p \,\leq\, \rho} \mathcal{L}_{\mathcal{D}}(w-\delta) - \mathcal{L}_{\mathcal{D}}(w), \tag{3}$$

where $\odot$ denotes element-wise multiplication.

The sharpness here refers to how rapidly the loss changes with respect to the changes in the model parameters.[1] While both the Fisher Information and sharpness are used for investigating loss landscapes and generalization, they offer different views of the training process.

Both S^{ρ}_{avg} and S^{ρ}_{worst} are studied in prior literature for generalization [5, 16, 20]. While prior work believes that flatter (less sharp) minima in the loss landscape lead to better generalization in

[1] The sharpness can be negative.

neural networks [3, 12, 14, 17], these metrics' attribution to the early period of training and how their early period trends are related to OOD generalization is understudied.

2.3. Gradual Unfreezing

Gradual unfreezing [13] is a simple tuning method that progressively increases the number of trainable parameters (i.e., unfreeze, layer-by-layer) of the neural network from the top to the bottom of the network with a fixed interval of training steps, k (i.e., the unfreezing interval). In this paper, we applied a modified version of gradual unfreezing [22], progressively unfreezing parameter "blocks" top-down early in training. A "block" can range from a single layer to multiple consecutive layers, defined by the standard parameter namespaces in typical implementations in our experiments. See Appendix A for details and the algorithm.

3. Experimental Setup

We perform our experiments on the ResNet-18 network, trained from scratch. For evaluation, we use MNIST [21], CIFAR10 [19] and CIFAR100 [19] for training, and MNIST-C [25], CIFAR10-C [11] and CIFAR-100-C [11] for OOD evaluation (results averaged across different corruption types and severity). We use the Auto-PGD algorithm [4] as implemented in [2] (we refer the readers to the original papers for details). In our experiments, the default learning rate specified in Appendix B is denoted as lr_d and we also experimented with reduced learning rates which are 1/10th of the default, specified as $0.1{*}lr_d$. Other training details are also given in Appendix B.

4. Gradual Unfreezing in the Early Period of Training Can Improve Out-of-Distribution Generalization

Here, we first validate that gradual unfreezing applied to the early period of training in our controlled setting (training from scratch) could also help OOD generalization. By examining three different datasets and two model architectures in Table 1, progressive parameter unfreezing (i.e., gradual unfreezing) does not influence ID results by a large margin (mostly minor degradation, but can also positively impact the ID results). However, gradual unfreezing has a non-negligible positive impact on the OOD results. This observation applies to various learning rates, but empirically, the default (higher) learning rate performs better for both ID and OOD tasks on CIFAR datasets. Here, we provide evidence that gradual unfreezing can improve OOD performance when training from scratch even if it was proposed for transfer learning [13], and validate the usability of gradual unfreezing as an intervention for our study.

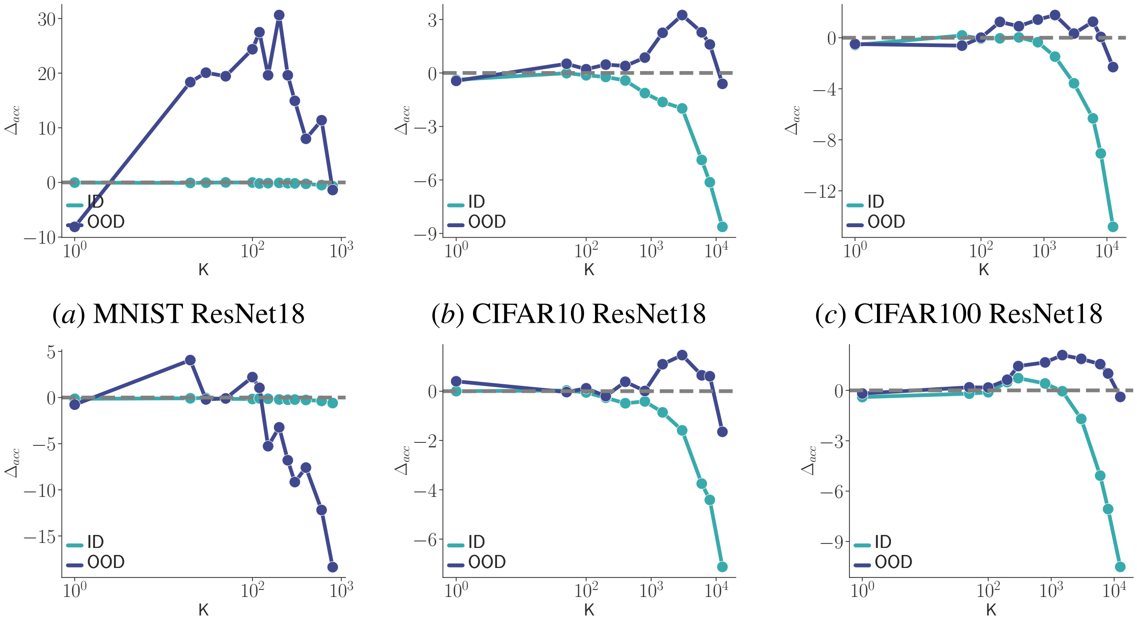
5. Evidence of Impact on Out-of-Distribution Generalization

Using gradual unfreezing, we experimented with different unfreezing steps k (ranging from 1 to equally dividing the total training steps among the number of trainable parameter blocks) to measure its impact on both ID and OOD test results. Indeed (as in Figure 1), it is possible that withholding trainable parameters can influence the OOD generalization as early as after training on a single batch of data. The effect is especially prominent for simpler datasets like MNIST.

Prolonging the unfreezing interval of parameters during training initially results in minimal change in the ID test performance with a larger learning rate, subsequently leading to quick de-

Table 1: Classification results on various datasets and model architectures. RN18 is ResNet-18. The default learning rate (lr_d) cases are the same as described in §3, 0.1*lr_d indicates the learning rate is 1/10th of the default learning rate. **GU:** indicates that gradual unfreezing is applied to the early period of training, here we observe the results that are close to the ID results, but with better OOD results (bolded).

Method	**MNIST RN18** ID / OOD	**CIFAR10 RN18** ID / OOD	**CIFAR100 RN18** ID / OOD
lr_d	99.06/33.36	93.32/72.36	71.07/45.10
lr_d + GU	98.98/**63.99**	93.26/**72.95**	71.03/**46.34**
0.1*lr_d	99.26/58.46	91.66/71.14	69.95/44.59
0.1*lr_d + GU	99.18/**62.51**	91.51/**71.26**	70.67/**46.03**



(*a*) MNIST ResNet18 (*b*) CIFAR10 ResNet18 (*c*) CIFAR100 ResNet18

(*d*) MNIST ResNet18 (*e*) CIFAR10 ResNet18 (*f*) CIFAR100 ResNet18

Figure 1: Change in ID and OOD evaluation results when unfreezing parameters at different times (compared to standard training). The shortest unfreezing interval k is 1 (training with 1 batch of data). Experiments in the first row use lr_d and the second row uses 0.01*lr_d. The x-axis is in the log scale.

terioration of the ID results. The deterioration of ID results over unfreezing intervals aligns with trends observed in the early stages of training using other interventions and complements the prior work [1, 10]. The impact on OOD results serves as evidence that the early period of training can influence OOD generalization.

Gradual unfreezing reveals interesting trends in the OOD evaluation results and highlights the trade-off between ID and OOD generalization. There exists a brief period during which OOD results may improve, prior to a rapid decline in ID performance. As soon as the ID results start to decrease rapidly, the OOD results first increase, then rapidly decrease for CIFAR10/100 with ResNet, but not in MNIST. The experiment results indicate that there may be a range of k in the early period of training with improved OOD results and minimal decay in ID results.

6. Learning Dynamics in the Early Period of Training

Observe in Figure 2 (for CIFAR10, and Figure 3 in the Appendix for other datasets) that by freezing the number of trainable parameters at a time (and gradually unfreezing them), we can induce higher Fisher Information and larger S^{ρ}_{avg}, S^{ρ}_{worst} at the beginning of training compared to the standard training procedure. In general, the longer we withhold parameters, the higher the level of sharpness and $\texttt{tr}(\mathrm{F})$ we can sustain, unfreezing parameters reduces these metrics.

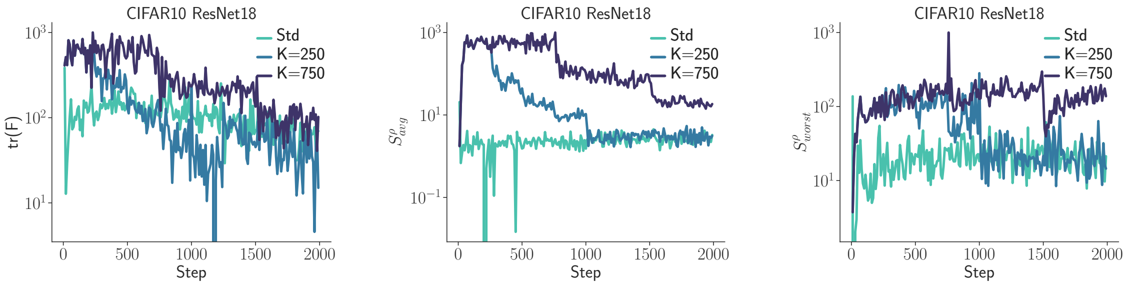
While there are variations between S^{ρ}_{avg}, S^{ρ}_{worst} and $\texttt{tr}(\mathrm{F})$, they are all sensitive to the early period of training and interventions. S^{ρ}_{avg} shows more consistent trends across datasets and architectures compared to the other two metrics. Due to the randomness in estimating S^{ρ}_{avg}, S^{ρ}_{worst}, and $\texttt{tr}(\mathrm{F})$, it is also evident that a single, absolute largest value of these metrics during the early period of training may not be a consistent indication of OOD generalization (or ID generalization, in fact). This indicates that the discussion for a high or low value of S^{ρ}_{avg}, S^{ρ}_{worst}, or $\texttt{tr}(\mathrm{F})$ during the early period of training should focus on relative rather than absolute values of these metrics.

Empirically, our findings differ from prior work on ID generalization (such as [15]) that demonstrated an 'explosion' of $\texttt{tr}(\mathrm{F})$ during the early period of training (due to using a small learning rate) is harmful. Here, a higher $\texttt{tr}(\mathrm{F})$ induced by parameter freezing does not hurt generalization, in both ID and OOD. When considering the trainable parameters as a variable, having initial higher sharpness or $\texttt{tr}(\mathrm{F})$ can be *advantageous up to a certain time frame during training*. Overall, our results suggest that while a lower sharpness or $\texttt{tr}(\mathrm{F})$ during the early period of training may be good for ID generalization, when factoring in the trainable parameters, a lower initial sharpness or $\texttt{tr}(\mathrm{F})$ could lead to worse OOD generalization. This observation applies strictly to the ***very early period of training***, and the eventual reduction of sharpness or $\texttt{tr}(\mathrm{F})$ after the initial period is still desirable (evident in Figure 1, Figure 2, and work like SAM, [7, 20, 26]). This indicates the need to develop new theories for OOD generalization. While sharpness and $\texttt{tr}(\mathrm{F})$ are effective for studying the early period of training, their relative trends are important for OOD generalization (also in ID).

7. Impact of Early Period of Training is a General Phenomenon

Gradual unfreezing is a specific case where high sharpness at the initial learning period could benefit OOD generalization. A critical question is: are there alternative methods for intervening during the early period of training with higher initial sharpness, that also positively impact OOD generalization?

Higher Initial Sharpness via Learning Rate Recall that a higher learning rate typically results in lower sharpness (and a lower learning rate results in higher sharpness, as indicated in [15]). Based on our findings in the previous sections, we hypothesize that using a lower learning rate at the initial period of learning, then switching to a higher learning rate later (high sharpness to low sharpness)



(*a*) CIFAR10 $\mathtt{tr}(\mathrm{F})$ (*b*) CIFAR10 S^{ρ}_{avg} (*c*) CIFAR10 S^{ρ}_{worst}

Figure 2: Unfreezing parameters at different times affects the learning dynamics in the early period of training. The y-axis uses a log scale and is normalized from 0 to 1000 for visualization.

Table 2: Effect of interventions during the early period of training on ID and OOD performance. Sharpness profiles are induced by: changing the learning rate (LR, left in blue), and delaying the application of a Fisher penalty (FP, right in red).

Baseline (LR)	**ID/OOD**	Intervention (LR)	**ID/OOD**	Baseline (FP)	**ID/OOD**	Intervention (FP)	**ID/OOD**
lr_d	93.32/72.36	lr_{l2h}	93.35/**72.89**	w/o FP	84.45/65.61	-	-
0.1*lr_d	91.66/71.14	lr_{h2l}	91.58/71.18	w/ FP	85.39/65.77	w/ FP$_{k=2000}$	85.20/**66.71**

may help OOD generalization (surprisingly, this is a simple form of learning rate warm-up!). To validate, we use CIFAR10 dataset on ResNet18 with two learning rates: we initially use 1/10th of the default learning rate, then increase the learning rate to the default value after k steps (i.e., low-to-high, denoted as lr_{l2h}, and the reverse (high-to-low) lr_{h2l}. Additionally, we experiment with using a Fisher penalty [15] on CIFAR10 and a simple CNN network (see details in the Appendix D) and delay the application of the Fisher penalty in order to induce a high-to-low sharpness profile. Results are in Table 2. Inducing high-to-low sharpness profiles leads to improved OOD results in both cases.

8. Conclusions

In this work, we empirically study the early period of training of neural networks and show that 1) changing the number of trainable parameters at certain times (via gradual unfreezing) has a minor impact on ID results, but can significantly improve OOD results (covariate shift); 2) the absolute values of sharpness and trace of Fisher Information at the initial period of training are not indicative for OOD generalization, but the relative values are.

References

[1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=BkeStsCcKQ.

[2] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 840–902. PMLR, 2023. URL https://proceedings.mlr.press/v202/andriushchenko23a.html.

[3] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22405–22418, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcb41ccdc4363c6848a1d760f26c28a0-Abstract.html.

[4] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

[5] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

[6] Rory A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22:700 – 725, 1925.

[7] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

[8] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/405075699f065e43581f27d67bb68478-Abstract.html.

[9] Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hkl1iRNFwS.

[10] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10677–10687. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/87784eca6b0dea1dff92478fb786b401-Abstract.html.

[11] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[13] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

[14] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 876–885. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/313.pdf.

[15] Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4772–4784. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jastrzebski21a.html.

[16] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

[17] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

[18] Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Aq35gl2c1k.

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[20] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914. PMLR, 2021. URL http://proceedings.mlr.press/v139/kwon21b.html.

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

[22] Chen Cecilia Liu, Jonas Pfeiffer, Ivan Vulic, and Iryna Gurevych. Improving generalization of adapter-based cross-lingual transfer with scheduled unfreezing. *CoRR*, abs/2301.05487, 2023. doi: 10.48550/ARXIV.2301.05487. URL https://doi.org/10.48550/arXiv.2301.05487.

[23] Zixuan Liu, Ziqiao Wang, Hongyu Guo, and Yongyi Mao. Over-training with mixup may hurt generalization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=JmkjrlVE-DG.

[24] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=nzpLWnVAyah.

[25] Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. *CoRR*, abs/1906.02337, 2019. URL http://arxiv.org/abs/1906.02337.

[26] Tom Sherborne, Naomi Saphra, Pradeep Dasigi, and Hao Peng. TRAM: Bridging trust regions and sharpness aware minimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kxebDHZ7b7.

[27] Yi-Lin Sung, Varun Nair, and Colin Raffel. Training neural networks with fixed sparse masks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24193–24205, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cb2653f548f8709598e8b5156738cc51-Abstract.html.

[28] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

[29] Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8156–8165. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Regularizing_Neural_Networks_via_Adversarial_Model_Perturbation_CVPR_2021_paper.html.

Appendix A. Gradual Unfreezing

Following the notations and algorithm in [22], let FORWARD$(*)$ be the standard forward pass, and BACKWARD$(*)$ calculates gradients and performs updates for trainable parameters. The modified gradual unfreezing algorithm is in Algorithm 1.

In our experiments, we partition the blocks by their natural namespaces as follows:

ResNet18: The definition block follows the standard implementation of ResNet, with an input convolution layer and a batch norm group together as the additional block. The model parameters are partitioned into 5 blocks, and a classification head.

Algorithm 1: Gradual Unfreezing

```
Require A model's eventual trainable parameters are partitioned into blocks $j \in \{0, \dots, L-1\}$
  parameterized by $\theta_j$, with a task-specific classification head $C$, and an unfreezing interval $k$. A
  set $\mathcal{S}$ of the indices of parameter blocks to unfreeze.
Initialize $C$, $\theta_j$ for all $j$;
$\mathcal{S} \leftarrow \{C\}$;
$j \leftarrow L-1$;
for $i = 1 \dots N$ do
    Sample a data batch $b \sim D$;
    if $i \mod k == 0$ and $i \leq kL$ then
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{\theta_j\}$;
        $j \leftarrow j - 1$;
    end
    FORWARD$(*)$;
    BACKWARD$(\mathcal{S})$ ;
end
```

Appendix B. Hyperparameters

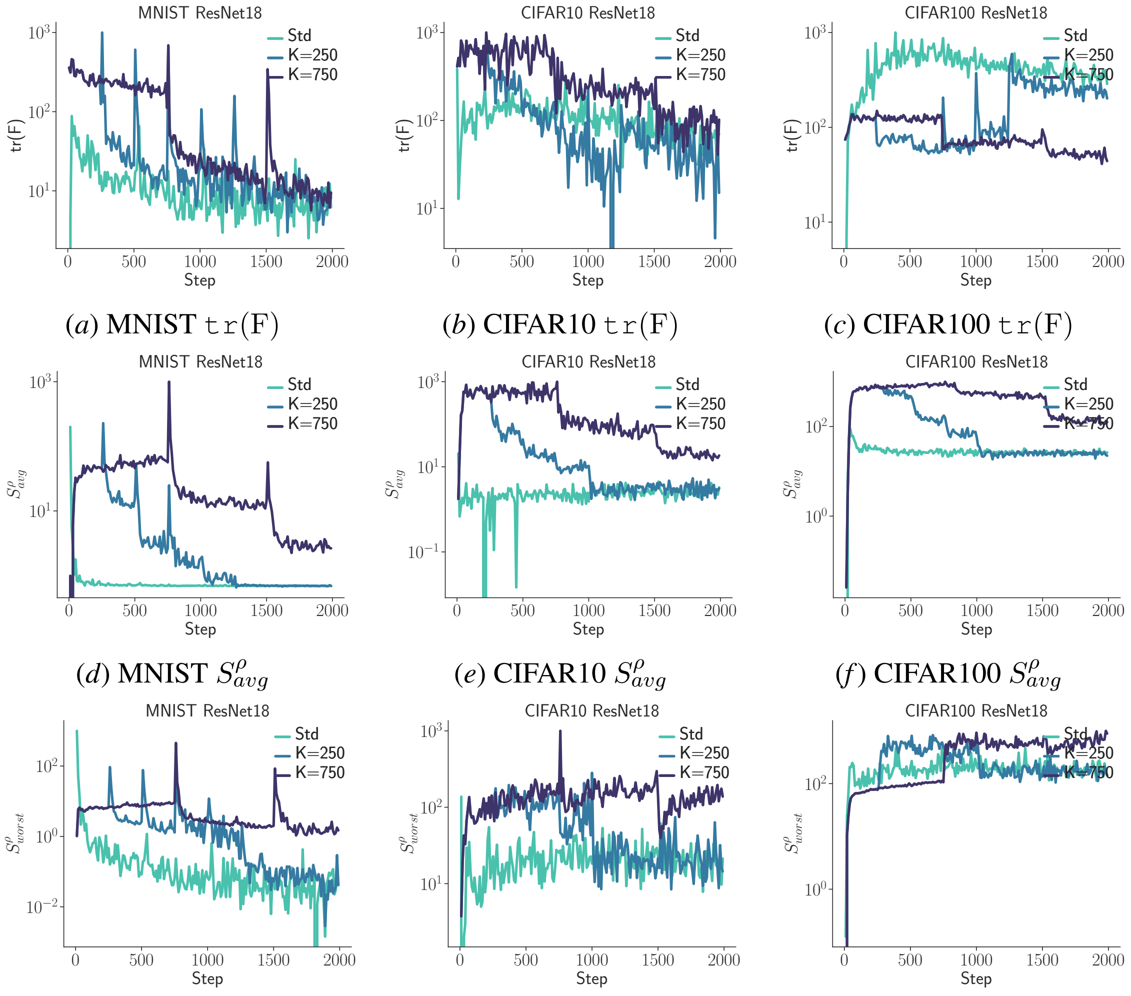
The hyperparameters are listed in Table 3 for all our experiments. We use the SGD optimizer and apply standard augmentations (i.e., random crops and flips) to the CIFAR datasets. We report results over 6 random seeds for MNIST (due to the high variance in OOD results), and we use 4 random seeds for all other experiments. The experiments use a single NVIDIA P100, A6000 or A100 GPU depending on the availability.

For calculating S^{ρ}_{worst} and S^{ρ}_{avg}, we use $L2$ norm and $\rho = 0.01$ with 15 examples. We follow the setup in [2] and use the implementation with 2048 data points from the training data (un-augmented when calculating sharpness metrics) for all experiments. We use a batch size of 256 for calculating all the metrics. The sharpness and `tr`(F) are recorded every 10 batches (steps) for all datasets.

Table 3: Hyperparameters used in all experiments.

	MNIST RN18	CIFAR10 RN18	CIFAR100 RN18
optimizer	AdamW	SGD	SGD
lr scheduler	const.	const.	const.
lr_d	0.01	0.1	0.01
batch size	128	128	128
training epochs	10	200	200
weight decay	0.01	0	0.0005
momentum	0.9	0	0.9

Appendix C. Learning Dynamics



(*a*) MNIST $\mathtt{tr}(\mathrm{F})$ (*b*) CIFAR10 $\mathtt{tr}(\mathrm{F})$ (*c*) CIFAR100 $\mathtt{tr}(\mathrm{F})$

(*d*) MNIST S^{ρ}_{avg} (*e*) CIFAR10 S^{ρ}_{avg} (*f*) CIFAR100 S^{ρ}_{avg}

(*g*) MNIST S^{ρ}_{worst} (*h*) CIFAR10 S^{ρ}_{worst} (*i*) CIFAR100 S^{ρ}_{worst}

Figure 3: Unfreezing parameters at different times affects the learning dynamics in the early period of training (with lr_d). The y-axis is in the log scale and is normalized between 0 and 1000 for visualization.

Appendix D. Fisher Penalty

A small learning rate can induce a higher sharpness and $\mathtt{tr}(\mathrm{F})$ in general, and prior work has shown that regularizing $\mathtt{tr}(\mathrm{F})$ can help with ID generalization [15]. Let $\mathcal{J}$ be the original loss, the total loss with Fisher penalty is in Eqn. 4. Following the simple CNN setting in [15, Appendix I.2], we train a simple 4-layer CNN (with one MaxPooling layer, no dropout) and a final fully connected layer of 128 hidden units on the CIFAR10 dataset with data augmentations. The model is trained for 300 epochs using an SGD optimizer with batch size 128, momentum 0.9, and a learning rate decay of 0.1 after epochs 150 and 225. We use a starting learning rate of 0.001 (a smaller learning rate than the default) and apply the Fisher penalty (FP) with a strength of 0.01 (α) every 10 steps.

$$\mathcal{J}_{total} = \mathcal{J} + \alpha * \texttt{tr}(\mathrm{F})\,. \tag{4}$$