In Search of the *Successful* Interpolation: On the Role of *Sharpness* in CLIP Generalization

Alireza Abdollahpoorrostam
Department of Computer Science & Communication Systems
EPFL
Switzerland, Lausanne
alireza.abdollahpoorrostam@epfl.ch

Abstract

Zero-shot models like CLIP are often fine-tuned on a target dataset to improve its accuracy further, but this can compromise out-of-distribution (OOD) robustness. Robust Fine-Tuning (RFT) [Wortsman et al., 2022c], which interpolates between the *zero-shot* and *fine-tuned* models, has been proposed to address this issue. However, understanding when RFT actually improves OOD error remains limited. In this work, we empirically investigate the robustness of RFT in CLIP models, with a focus on the *sharpness* of the CLIP model during interpolation. First, we demonstrate that while sharpness may not serve as a reliable indicator for predicting the generalization of modern architectures like CLIP on OOD data, this challenges the conventional belief in the generalization benefits of flat minima in foundation models. However, by examining the role of the *straggler layer* phenomenon, we show that, unlike overall sharpness, the *layer-wise* sharpness of *straggler* layers can reliably capture the generalization performance of interpolated CLIP models on OOD data. Our extensive experiments reveal that *layer-wise* sharpness correlates with generalization in OOD accuracy for RFT. Furthermore, we demonstrate that by inducing sparsity in the *straggler* layers, we can mitigate the *failure mode* phenomenon in RFT. To the best of our knowledge, this is the first work to study the role of sharpness in the *success* of interpolation in the weight space of CLIP foundation models. Our code is available at https://github.com/alirezaabdollahpour/CLIP_Mode_Connectivity.

1 Introduction

Understanding the behavior of large machine learning models like CLIP [Radford et al., 2021] on OOD tasks is important for their safe deployment. Analyzing their behavior on a path between the initial and the final parameters has been proposed as a simple yet insightful approach this. However, prior works [Vlaar and Frankle, 2022, Lucas et al., 2021, Neyshabur et al., 2020, Draxler et al., 2018, Entezari et al., 2022, Chatterji et al., 2020] has primarily focused on CNN models for this analysis and whether such analysis extends to other kinds of architecture has not been thoroughly explored. On the other hand, several works have shown that while foundation models like CLIP exhibit outstanding zero-shot OOD performance, this can be further improved if they are fine-tuned on the relevant target domain. However, this improvement comes at the cost of reduced performance on domains that it is not trained on. To solve this problem, inspired by the above-mentioned works on interpolation in CNNs, Wortsman et al. [2022b] showed that on the path connecting the *zero-shot* model and the final *fine-tuned* model, there exists a model with better OOD performance and proposed an algorithm called *Robust Fine Tuning* (RFT) to find this parameter. However, RFT does not always succeed in achieving large improvement in OOD accuracy compared to the *zero-shot*

AdvML-Frontiers24: The 3nd Workshop on New Frontiers in Adversarial Machine Learning@NeurIPS24, Vancouver, CA.

model, and very little understanding exists of when the improvement is large and when it isn't. In this work, we aim to address this lack of knowledge. Inspired by earlier work on the interpolation between two CNN models, we first provide extensive experimental results to examine the correlation between the weight space geometry and CLIP's capability to generalize on OOD tasks. We aim to address the following question:

> *How does sharpness on OOD samples relate to CLIP generalization?*

Second, we investigate the role of the specific layer's sharpness on CLIP's OOD generalization. In particular, we ask the following question:

> *What occurs within a layer during interpolation that leads to a failure mode? By measuring the sharpness of that layer during interpolation, can we predict its impact on generalization?*

Robust Fine-Tuning (`RFT`) method has two steps: first, they fine-tune the *zero-shot* model on the target distribution. Second, they combine the original *zero-shot* and fine-tuned models by linearly interpolating between their weights, coined as weight-space ensembling. Nevertheless, the connection between linear interpolation and OOD generalization for CLIP has not been thoroughly investigated. The question of why the linear interpolation between *zero-shot* and fine-tuned CLIP models succeeds in OOD tasks, and the conditions under which the linear path between two CLIP models indicates robust generalization performance on OOD tasks, remains an unresolved problem. The recent advancements in the understanding of loss landscapes in CNNs and their connection to generalization through linear paths have prompted Abdolahpourrostam et al. [2024] to revisit these findings within the context of foundation models like CLIP. Abdolahpourrostam et al. [2024] aims to bridge the gap between the assumptions made about linear interpolation and loss landscape geometry in CNNs and the generalization capabilities of CLIP. Their study seeks to identify the conditions under which linear interpolation can be *successfully* applied between two CLIP models, with particular attention to the roles of data augmentation and learning rate magnitude during the fine-tuning process.

On the role of sharpness: There is a body of literature suggesting that flatter minima may have better generalization properties [Xing et al., 2018, Zhou et al., 2021, Cha et al., 2021, Park and Kim, 2022, Lyu et al., 2023, Andriushchenko et al., 2023] for standard or OOD data. However, the definitions of sharpness commonly used in the field do not align effectively with the concept of generalization, as discussed [Kaur et al., 2023] this can be primarily due to the model's lack of invariance under reparametrizations that not change the model [Dinh et al., 2017, Granziol, 2020, Zhang et al., 2021, Andriushchenko et al., 2023]. The utilization of adaptive sharpness seems to hold more potential as it effectively resolves the reparametrization problem and has been demonstrated to exhibit a stronger empirical correlation with generalization. [Kwon et al., 2021, Andriushchenko et al., 2023]. Furthermore, SAM demonstrates notable utility in emerging architectures such as vision transformers [Chen et al., 2022, Andriushchenko et al., 2023]. In addition, although transfer learning has become the prevailing method for vision problems, the consequences of sharpness in this context have not been thoroughly investigated. Furthermore, the correlation between sharpness and OOD generalization has not been extensively examined. These rising innovations highlight the necessity to reevaluate the significance of sharpness in these new environments.

1.1 Background on Interpolation and Notations

Loss barrier. For loss landscapes, *barriers* refer to regions of increased loss encountered along the interpolation path between two sets of model parameters.

We examine a CLIP architecture that is parametrized by $\boldsymbol{\theta}$ and is fine-tuned on a task represented by a training set S_{train} and a test set S_{test}. In the following, as we are interested in the generalization of CLIP on OOD tasks, we consider OOD loss and accuracy and write $\mathcal{L}(\boldsymbol{\theta}), \mathcal{A}(\boldsymbol{\theta})$ for $\mathcal{L}(\boldsymbol{\theta}, S_{\text{OOD}}), \mathcal{A}(\boldsymbol{\theta}, S_{\text{OOD}})$. Assume that we have fixed two different different sets of weights $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Let $\mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \mathcal{L}(\alpha\boldsymbol{\theta}_0 + (1-\alpha)\boldsymbol{\theta}_1)$ and $\mathcal{A}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \mathcal{A}(\alpha\boldsymbol{\theta}_0 + (1-\alpha)\boldsymbol{\theta}_1)$ for $\alpha \in [0, 1]$ be the loss and accuracy, respectively, of the CLIP network created by linearly interpolating between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Then, building upon the Frankle et al. [2020] definition for linear interpolation instability, we define it for CLIP on OOD as the following notion.

Definition 1. *The difference between the supremum of the loss for any interpolation* $\sup_\alpha \mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ *and the average loss of the endpoints* $\frac{1}{2}(\mathcal{L}(\boldsymbol{\theta}_0) + \mathcal{L}(\boldsymbol{\theta}_1))$ *is called the linear interpolation instability for the CLIP on OOD.*

Recall that *zero-shot* CLIP performs better on OOD tasks compared to the fine-tuned version of CLIP. Within the same settings of [Wortsman et al., 2022b,a], we are interested in exploring the linear path between *zero-shot* CLIP and fine-tuned CLIP. Therefore, we set $\boldsymbol{\theta}_0$ as *zero-shot* model.

Two parametrizations $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ have a **barrier** between them if the linear interpolation instability for ***sufficiently*** large δ, there exists an $\alpha \in [0, 1]$ such that:

$$\sup_{\alpha} \mathcal{L}_{\alpha}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}) - \mathcal{L}(\boldsymbol{\theta}_0; S_{\text{OOD}}) \geq \delta > 0 \tag{1}$$

The value of δ can be empirically determined for each OOD task. Similarly, we state that linear interpolation or the `RFT` algorithm can achieve ***high gain accuracy*** if there exists an $\alpha \in [0, 1]$ such that:

$$\sup_{\alpha} \mathcal{A}_{\alpha}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}) - \mathcal{A}(\boldsymbol{\theta}_0; S_{\text{OOD}}) \geq \xi > 0 \tag{2}$$

where ξ is ***sufficiently*** large.

Also, we define a linear path as having a *gain* if the *supremum* in Eq. 2 exists with $(\xi > 0)$. It is important to mention that a path is considered a ***failure mode*** if the *supremum* in Eq. 2 does not exist. Figure 1 illustrates scenarios in which several distinct *fine-tuned* CLIP models experience either *failure mode* or *high gain accuracy* outcomes during the interpolation (`RFT`).

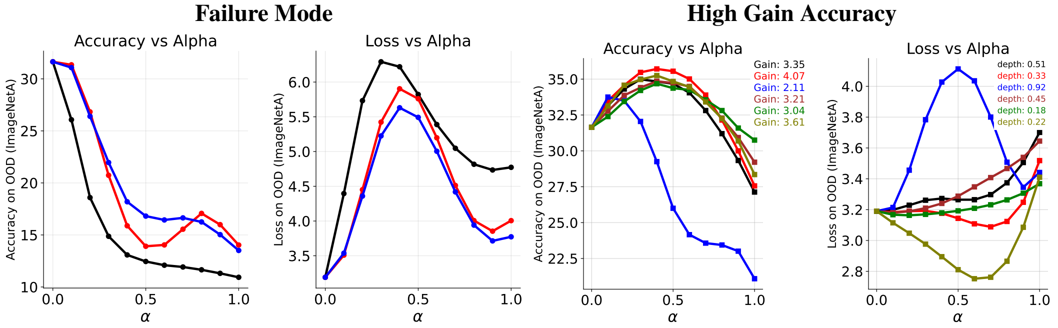


Figure 1: For 9 distinct fine-tuned CLIP models (each color shows different CLIP models) on ImageNet [Deng et al., 2009], this plot demonstrates the accuracy and loss on ImageNet-A [Hendrycks et al., 2021] as an OOD task. For each model, we show the maximum accuracy gain achieved along a corresponding interpolation path. In the loss plot, we show **depth** as the largest barrier on the interpolation path starting from the *zero-shot* model.

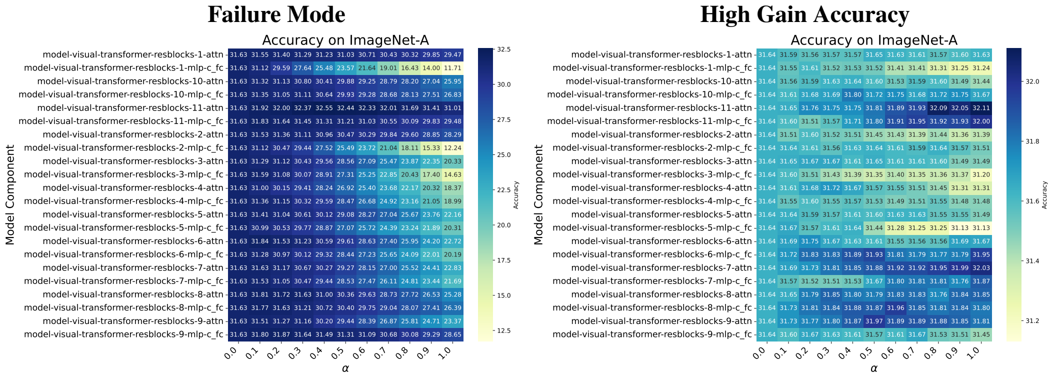
Layer-wise interpolation. In the following, we analogously define a layer-wise notion of instability. Let $\mathcal{M}$ be structured in L layers $\{\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)}\}$. In our experiments, we consider both weights and bias as one set of parameters describing a layer. Let us fix a layer $\boldsymbol{W}^{(i)}$. Consider a parametrization that is defined by α, $\mathcal{W}_1$ and $\mathcal{W}_2$ as $\{\boldsymbol{W}_j^{(1)}, \boldsymbol{W}_j^{(2)}, \ldots, \alpha\boldsymbol{W}_1^{(i)} + (1-\alpha)\boldsymbol{W}_2^{(i)}, \ldots, \boldsymbol{W}_j^{(L)}\}$ where j can be selected to be 1 or 2.

Definition 1.1. *(Layer-wise linear interpolation instability) The difference between supremum of the loss on the line* $sup_{\alpha}\mathcal{L}_{\alpha,i}(\mathcal{W}_1, \mathcal{W}_2)$ *corresponding to layer* $\boldsymbol{W}^{(i)}$ *and average loss of the original models* $\frac{1}{2}(\mathcal{L}(\mathcal{W}_1) + \mathcal{L}(\mathcal{W}_2))$ *is the* ***layer-wise linear interpolation instability*** *for the given architecture* $\mathcal{M}$ *and selected layer (A similar approach can be employed to analyze this phenomenon by evaluating the accuracy on OOD data.).*

Definition 1.2. *(Straggler layer) If a layer demonstrates layer-wise interpolation instability, it is referred to as a straggler layer.*

We are particularly interested in layers where linear interpolation leads to a ***failure mode*** in terms of accuracy on OOD data. In other words, if a layer exhibits ***layer-wise interpolation instability***, it manifests this ***failure mode*** phenomenon.

Note: Since we utilize the weights of the *zero-shot* CLIP model, denoted as $\mathcal{W}_1$ ($\boldsymbol{\theta}_0$), and the weights $\mathcal{W}_2$ from the *fine-tuned* CLIP model ($\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_{\text{FT}}$), we assign the *zero-shot* CLIP weights to all layers except the target layer i. This approach allows us to specifically analyze the performance of layer i in the *fine-tuned* CLIP model.



Failure Mode

High Gain Accuracy

Figure 2: **Layer-wise interpolation on ImageNet-A as OOD.** For two distinct fine-tuned CLIP models one exhibiting *failure mode* and the other *high gain accuracy* in regular interpolation (RFT), we conduct a *layer-wise* interpolation alongside each layer with the *zero-shot* CLIP model.

2 Adaptive Sharpness and its Invariances

In this section, we begin by providing background on adaptive sharpness and then discuss its invariance properties in modern architectures. We categorize the sharpness of a model into two distinct categories. First, we establish a connection between *general* sharpness and the generalization performance of the interpolated CLIP model. Second, we introduce the concept of *layer-wise* sharpness and, by utilizing the relationship between *straggler* layers and *layer-wise* sharpness, we experimentally demonstrate how the *layer-wise* sharpness of *straggler* layers can capture the generalization performance of interpolated CLIP models.

2.1 Background on Sharpness

Sharpness definitions. Similar to [Andriushchenko et al., 2023], we denote the loss on a set of *OOD* points $\mathcal{S}$ as $L_\mathcal{S}(\boldsymbol{w}) = \frac{1}{|S|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}} \ell_{\boldsymbol{xy}}(\boldsymbol{w})$, where $\ell_{\boldsymbol{xy}}(\boldsymbol{w}) \in \mathbb{R}_+$ represents some loss function (e.g., cross-entropy) on the pair $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{S}$ computed with the network weights $\boldsymbol{w}$. For arbitrary $\boldsymbol{w} \in \mathbb{R}^p$ (i.e., not necessarily a minimum), we define the *average-case* and *adaptive average-case* sharpness with radius ρ and with respect to a vector $\boldsymbol{c} \in \mathbb{R}^p$ as:

$$S^\rho_{avg}(\boldsymbol{w},\boldsymbol{c}) \triangleq \mathbb{E}_{\substack{\mathcal{S}\sim P_m \\ \boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))}} L_\mathcal{S}(\boldsymbol{w}+\boldsymbol{\delta}) - L_\mathcal{S}(\boldsymbol{w}) \tag{3}$$

where $\odot/^{-1}$ denotes elementwise multiplication/inversion and P_m is the data distribution that returns m pairs $(\boldsymbol{x},\boldsymbol{y})$. Using $\boldsymbol{c} = |\boldsymbol{w}|$ leads to *elementwise* adaptive sharpness [Kwon et al., 2021, Andriushchenko et al., 2023] and makes the sharpness invariant under multiplicative reparametrizations. For a thrice differentiable loss $L(\boldsymbol{w})$, the average-case elementwise adaptive sharpness can be computed as (see Andriushchenko et al. [2023] or App. A for proof):

$$S^\rho_{avg}(\boldsymbol{w},|\boldsymbol{w}|) = \mathbb{E}_{\mathcal{S}\sim P_m} \frac{\rho^2}{2}\text{tr}(\nabla^2 L_\mathcal{S}(\boldsymbol{w}) \odot |\boldsymbol{w}||\boldsymbol{w}|^\top) + O(\rho^3)$$

We should also mention that the first-order term cancels out completely. In order for better clarity, we will use the term *general sharpness*. In the upcoming sections, we will examine the connection between the sharpness of interpolated CLIP models and their generalization performance on OOD data. Next, we present our concept of *layer-wise sharpness*, which entails quantifying the sharpness of ***one specific layer*** within the CLIP model during interpolation.

3 Sharpness vs. Generalization

The current understanding of the relationship between sharpness and generalization is primarily based on experiments with non-residual convolutional networks and small datasets such as CIFAR-10 and SVHN [Jiang et al., 2019]. Andriushchenko et al. [2023] were the first to study the correlation between *general* sharpness and generalization in transformer-based modern architectures, such

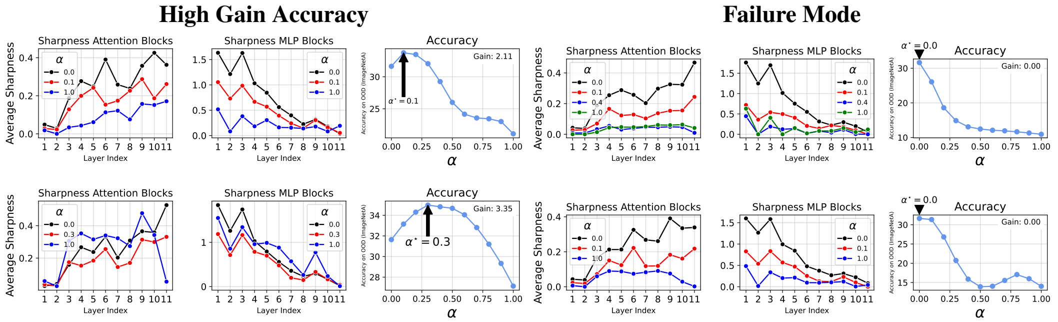


Figure 4: We present an analysis of the *layer-wise* sharpness across four distinct CLIP models, comprising two *failure mode* models and two *high gain accuracy* models, demonstrating the sharpness characteristics of each individual layer.

as fine-tuned CLIP models. Their findings revealed that there is no *strong* correlation between *general* sharpness and generalization on OOD data. Building on their observations, we investigate the correlation between *general* sharpness and interpolation. Additionally, we introduce the concept of *layer-wise* sharpness and demonstrate how, unlike *general* sharpness, it can effectively capture the generalization performance during interpolation in weight space between *zero-shot* and *fine-tuned* CLIP models.

In Fig. 3, we demonstrate that *general sharpness* ***fails*** to directly capture the generalization of interpolated CLIP models on OOD data. Contrary to our expectations, CLIP models fine-tuned on ImageNet indicate that flatter solutions consistently generalize worse on OOD data. This evidence suggests that the commonly held belief in the generalization benefits of flat minima does not hold true in modern settings. This result corroborates the findings of Andriushchenko et al. [2023], specifically for *fine-tuned* CLIP models on OOD data.

3.1 Layer-wise Sharpness

In this part, we introduce the concept of *layer-wise sharpness*, where we perturb the weight space of the target layer in the *fine-tuned* CLIP model during interpolation. Subsequently, we perform interpolation between this newly perturbed *fine-tuned* CLIP model and the *zero-shot* CLIP model. Notably, we do not conduct *layer-wise* interpolation; instead, we apply the previously described RFT algorithm. Informally speaking, we want answer to this question:

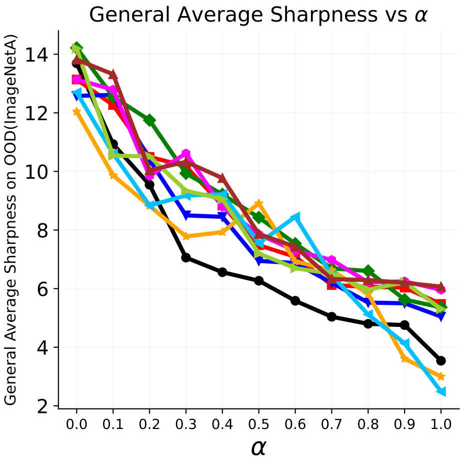


Figure 3: For 9 distinct fine-tuned CLIP models (each color shows different CLIP models) on ImageNet, this plot demonstrates the *general adaptive average sharpness* with $\rho = 1.0$ and 20 iterations on ImageNet-A as an OOD task.

> ***Question***: *What occurs within a layer during interpolation that leads to layer-wise interpolation instability or a failure mode? By measuring the sharpness of that layer during interpolation, can we predict its impact on generalization?*

Furthermore, we empirically investigate what occurs immediately after $\alpha^\star$ in *high gain accuracy* models. As shown in Fig. 1, for these models, we consistently observe a point along the interpolation path where the interpolated model reaches *maximum* accuracy. Beyond this point, a decline in performance begins. Figure 4 demonstrates that in models with *high gain accuracy* (second row), the optimal $\alpha^\star$ corresponds to a point where the interpolated model achieves *maximum* generalization performance. However, within this model, there is at *least one layer* where the *layer-wise* sharpness is *nearly zero*. On the other hand, for *failure mode* models, it is already known that there is no point along the interpolated path where the OOD accuracy surpasses that of the starting and ending points. Consequently, $\alpha^\star$ is exactly at the starting point (the *zero-shot* point). For *failure mode* models, it can be observed that there is at least one layer where the *layer-wise* sharpness is nearly zero. In

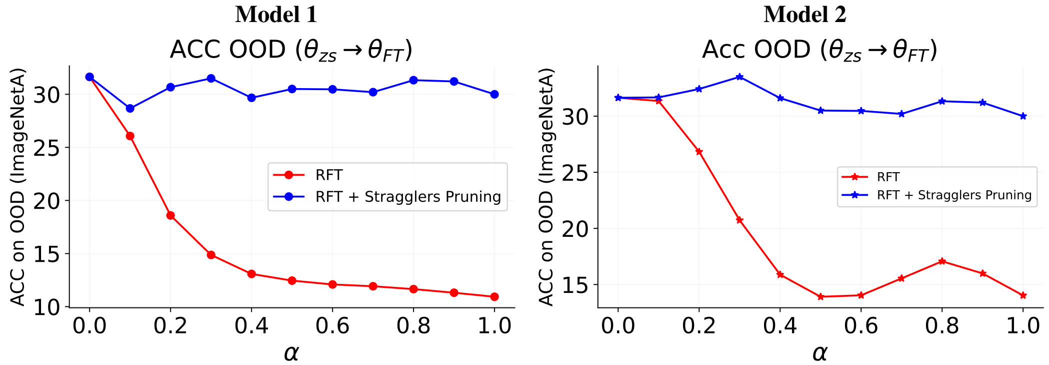


Figure 5: **Straggler layer pruning.** For two distinct *fine-tuned* CLIP models that exhibit *failure mode* during interpolation using the RFT algorithm, we demonstrate that pruning the *straggler* layers of the *fine-tuned* model prevents a collapse in performance.

fact, fine-tuned *failure mode* models inherently possess a *straggler layer*. In the following section, we evaluate our *layer-wise* sharpness and *straggler* layer in a different direction. We introduce a straightforward algorithm based on the *layer-wise* sharpness of the *straggler* layers.

On the role of Sparsity for Generalization and RFT.

Algorithm 1 Pytorch Pseudocode for Straggler Layer Pruning

```
Require: Model $\mathcal{M}$ structured in $L$ layers $\{\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L)}\}$, zero-shot model $\boldsymbol{\theta}_{\texttt{zero-shot}}$.
for $i = 1$ to $L$ do
    if Adaptive Average Sharpness($\boldsymbol{W}^{(i)}$, $\rho$) $\simeq 0$ then
        mask $\leftarrow$ torch.bernoulli(torch.full_like( $\mathcal{M}[\boldsymbol{W}^{(i)}]$, 0.5)).bool()
        $\mathcal{M}[\boldsymbol{W}^{(i)}]$[mask] $\leftarrow 0$
    end if
end for
$\boldsymbol{\theta}_\alpha$ = interpolation($\boldsymbol{\theta}_{\texttt{zero-shot}}$, $\mathcal{M}$)
return $\boldsymbol{\theta}_\alpha$
```

Our objective is to establish a connection between the *layer-wise* sharpness of *straggler* layers and the generalization performance of the interpolated model. While the primary aim of this work is not to introduce a new algorithm that surpasses conventional interpolation methods, we focus on elucidating the importance of the *layer-wise* sharpness phenomenon. First, through five iterations, we identify the *straggler* layers of the fine-tuned CLIP model. Subsequently, we randomly adjust the weights of the identified layers. Specifically, before initiating the interpolation, we make the *straggler* layers ***sparse***. In Algorithm 1, we summarize our algorithm.

4 Conclusion and Future works

In conclusion, our study underscores the critical role of interpolation (RFT) in enhancing the generalization capabilities of CLIP models for OOD tasks. We demonstrate that by putting specific layers in CLIP models under the microscope, referred to as *straggler layers*, and employing the concept of *layer-wise sharpness* as opposed to the traditional notion of *general sharpness*, we can effectively assess the generalization performance of these interpolated models on OOD data. Our findings indicate that if a *fine-tuned* CLIP model contains at least one layer where the *layer-wise sharpness* is nearly zero, it triggers a *failure mode* phenomenon. Furthermore, for interpolated CLIP models that achieve *high gain accuracy* along the interpolation path, a decline in OOD performance begins when, at the point of maximum OOD accuracy ($\alpha^\star$), there exists a layer with nearly zero *layer-wise* sharpness. This specific layer is identified as the *straggler layer*. Importantly, this study is the first to explore the generalization and interpretability of CLIP models, through the lenses of mode connectivity, interpolation and *sharpness*. Our findings provide novel insights into the behavior of these models and their potential for robust application across diverse tasks.

References

Alireza Abdolahpourrostam, Amartya Sanyal, and Seyed-Mohsen Moosavi-Dezfooli. Unveiling CLIP dynamics: Linear mode connectivity and generalization. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://openreview.net/forum?id=DFRAmfsuow. 2

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization, 2023. URL https://arxiv.org/abs/2302.07011. 2, 4, 5, 9

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima, 2021. URL https://arxiv.org/abs/2102.08604. 2

N. S. Chatterji, B. Neyshabur, and H. Sedghi. The intriguing role of module criticality in the generalization of deep networks. *ICLR*, 2020. 1

Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations, 2022. URL https://arxiv.org/abs/2106.01548. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. https://ieeexplore.ieee.org/abstract/document/5206848. 3

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets, 2017. URL https://arxiv.org/abs/1703.04933. 2

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, 2018. https://arxiv.org/abs/1803.00885. 1

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022. https://arxiv.org/abs/2110.06296. 1

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020. https://proceedings.mlr.press/v119/frankle20a.html. 2

Diego Granziol. Flatness is a false friend, 2020. URL https://arxiv.org/abs/2006.09091. 2

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, 2021. 3

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them, 2019. URL https://arxiv.org/abs/1912.02178. 4

Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. On the maximum hessian eigenvalue and generalization, 2023. URL https://arxiv.org/abs/2206.10654. 2

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks, 2021. URL https://arxiv.org/abs/2102.11600. 2, 4

James Lucas, Juhan Bae, Michael R Zhang, Stanislav Fort, Richard Zemel, and Roger Grosse. Analyzing monotonic linear interpolation in neural network loss landscapes, 2021. https://arxiv.org/abs/2104.11044. 1

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction, 2023. URL https://arxiv.org/abs/2206.07085. 2

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2008.11687. 1

Namuk Park and Songkuk Kim. How do vision transformers work?, 2022. URL https://arxiv.org/abs/2202.06709. 2

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

Tiffany J Vlaar and Jonathan Frankle. What can linear interpolation of neural network loss landscapes tell us? In *International Conference on Machine Learning*, pages 22325–22341. PMLR, 2022. 1

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022a. https://arxiv.org/abs/2203.05482. 3

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2022b. 1, 3

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022c. https://arxiv.org/abs/2109.01903. 1

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd, 2018. URL https://arxiv.org/abs/1802.08770. 2

Shuofeng Zhang, Isaac Reid, Guillermo Valle Pérez, and Ard Louis. Why flatness does and does not correlate with generalization for deep neural networks, 2021. URL https://arxiv.org/abs/2103.06219. 2

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and Weinan E. Towards theoretically understanding why sgd generalizes better than adam in deep learning, 2021. URL https://arxiv.org/abs/2010.05627. 2

A Appendix

Following to [Andriushchenko et al., 2023], let $L_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{|\mathcal{S}|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}} \ell_{\boldsymbol{xy}}(\boldsymbol{w})$ be the loss on a set of points $\mathcal{S}$. For arbitrary weights $\boldsymbol{w}$ (i.e., not necessarily a minimum), then the *average-case sharpness* is defined as:

$$S^{\rho}_{avg,p}(\boldsymbol{w},\boldsymbol{c}) \triangleq \mathop{\mathbb{E}}_{\substack{\mathcal{S}\sim P_m \\ \boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))}} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})$$

where $\odot/^{-1}$ denotes elementwise multiplication/inversion and P_m is the data distribution that returns m pairs $(\boldsymbol{x},\boldsymbol{y})$.

If $\boldsymbol{c} = |\boldsymbol{w}|$ then the perturbation set is $\left\|\delta \odot |\boldsymbol{w}|^{-1}\right\|_p \leq \rho$. Assume a new variable $\boldsymbol{\gamma} = \boldsymbol{\delta} \odot |\boldsymbol{w}|^{-1}$ and perform a Taylor expansion around w:

$$L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) = L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\gamma}\odot|\boldsymbol{w}|) = L_{\mathcal{S}}(\boldsymbol{w}) + \langle \nabla L_{\mathcal{S}}(\boldsymbol{w}), |\boldsymbol{w}|\odot\boldsymbol{\gamma}\rangle + \frac{1}{2}\left\langle \boldsymbol{\gamma}\odot|\boldsymbol{w}|, \nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\right\rangle + O(\|\boldsymbol{\gamma}\|_p^3),$$

where $\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})$ denotes the Hessian of $L_{\mathcal{S}}$ at $\boldsymbol{w}$.

Proposition A.1. *(Andriushchenko et al. [2023]), Let $L_{\mathcal{S}} \in C^3(\mathbb{R}^s)$, S be a finite sample of points $(x_i, y_i)_{i=1}^n$ and let P_m denote the uniform distribution over subsamples of size $m \leq n$ from S. Then*

$$\lim_{\rho\to 0} \frac{2}{\rho^2} S^{\rho}_{avg}(\boldsymbol{w}, |\boldsymbol{w}|) = \mathbb{E}_{\mathcal{S}\sim P_m}\left[tr(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w}) \odot |\boldsymbol{w}||\boldsymbol{w}|^{\top})\right] + O(\rho)$$

Proof. Let us consider the loss without the subcript for clarity. Then we consider

$$\mathbb{E}_{\boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))}\, L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})$$

When plugging in the Taylor expansion of the loss, we see that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})\\
&\qquad= \mathbb{E}_{\boldsymbol{\gamma}\in\mathcal{N}(0,\rho^2\boldsymbol{I})}\left[\langle \nabla L_{\mathcal{S}}(\boldsymbol{w}), |\boldsymbol{w}|\odot\boldsymbol{\gamma}\rangle + \frac{1}{2}\left\langle \boldsymbol{\gamma}\odot|\boldsymbol{w}|, \nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\right\rangle + O(\|\boldsymbol{\gamma}\|_2^3)\right]\\
&\qquad= \frac{1}{2}\,\mathbb{E}_{\boldsymbol{\gamma}\in\mathcal{N}(0,\rho^2\boldsymbol{I})}\left[\left\langle \boldsymbol{\gamma}\odot|\boldsymbol{w}|, \nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\right\rangle\right] + O(\rho^3)\\
&\qquad= \frac{1}{2}\,\mathbb{E}_{\boldsymbol{\gamma}\in\mathcal{N}(0,\rho^2\boldsymbol{I})}\left[\left\langle \boldsymbol{\gamma}, (\nabla^2 L_{\mathcal{S}}(\boldsymbol{w}) \odot |\boldsymbol{w}||\boldsymbol{w}|^T)\boldsymbol{\gamma}\right\rangle\right] + O(\rho^3)\\
&\qquad= \frac{\rho^2}{2}\mathrm{tr}(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w}) \odot |\boldsymbol{w}||\boldsymbol{w}|^{\top}) + O(\rho^3)
\end{aligned}$$

□