Robust fine-tuning of zero-shot models

Mitchell Wortsman[*†] **Gabriel Ilharco**[*†] **Jong Wook Kim**[§] **Mike Li**[‡]

Simon Kornblith[◇] **Rebecca Roelofs**[◇] **Raphael Gontijo Lopes**[◇]

Hannaneh Hajishirzi[†○] **Ali Farhadi**[†] **Hongseok Namkoong**[‡] **Ludwig Schmidt**[†△]

Abstract

Large pre-trained models such as CLIP or ALIGN offer consistent accuracy across a range of data distributions when performing zero-shot inference (i.e., without fine-tuning on a specific dataset). Although existing fine-tuning approaches substantially improve accuracy in-distribution, they often reduce out-of-distribution robustness. We address this tension by introducing a simple and effective method for improving robustness: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements out-of-distribution, while preserving high in-distribution accuracy. On ImageNet (in-distribution) and five derived distribution shifts, WiSE-FT improves out-of-distribution accuracy by 4 to 6 percentage points (pp) over prior work while increasing in-distribution accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness improvements (2 to 23 pp) on a diverse set of six further distribution shifts, and in-distribution accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on seven commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.

For the complete version of this paper, refer to `https://arxiv.org/abs/2109.01903`.

1 Introduction

A foundational goal of machine learning is to develop models that work reliably across a broad range of data distributions. Over the past few years, researchers have proposed a variety of challenging out-of-distribution benchmarks on which current algorithmic approaches to enhance robustness yield little to no gains [23, 17]. While these negative results highlight the difficulty of learning robust models, large pre-trained models such as CLIP [20] and ALIGN [12] have recently demonstrated unprecedented robustness to these challenging distribution shifts. The success of CLIP and ALIGN points towards pre-training on large, heterogeneous datasets as a promising direction for increasing robustness. However, an important caveat is that these robustness improvements are largest in the zero-shot setting, i.e., when the model performs inference without fine-tuning on a specific target distribution.

In a concrete application, a zero-shot model can be fine-tuned on extra application-specific data, which often yields large performance gains on the target distribution. However, in the experiments of Radford *et al.* [20], fine-tuning comes at the cost of robustness: across several natural distribution

[*] These authors contributed equally.

[†] University of Washington [‡] Columbia University [◇] Google Research, Brain Team

[§] OpenAI [○] Allen Institute for Artificial Intelligence [△] Toyota Research Institute

Code provided at `https://github.com/mlfoundations/wise-ft`.

Workshop on Distribution Shifts (NeurIPS 2021).

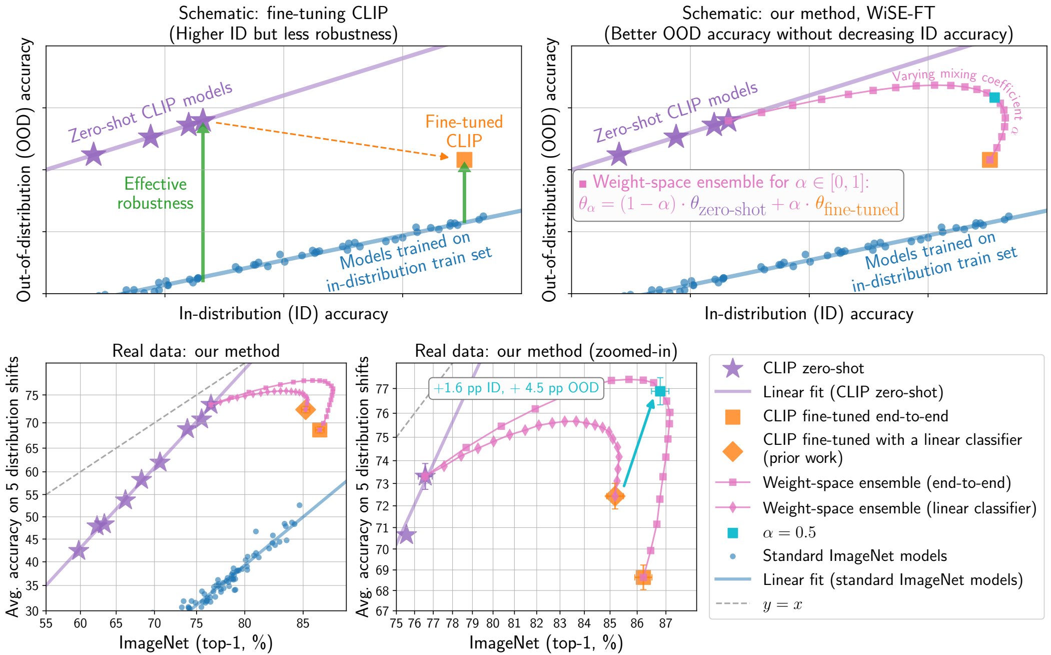

Figure 1: **(Top left)** Zero-shot CLIP models exhibit high effective robustness and moderate in-distribution accuracy, while standard fine-tuning—either end-to-end or with a linear classifier (final layer)—attains higher ID accuracy and less effective robustness. **(Top right)** Our method linearly interpolates between the zero-shot and fine-tuned models with a mixing coefficient $\alpha \in [0, 1]$. **(Bottom)** On five distribution shifts derived from ImageNet (ImageNetV2, ImageNet-R, ImageNet Sketch, ObjectNet, and ImageNet-A), WiSE-FT improves average OOD accuracy relative to both the zero-shot and fine-tuned models while maintaining or improving ID accuracy.

shifts, the out-of-distribution accuracy of their fine-tuned CLIP models is lower than that of the original zero-shot model. This leads to a natural question: *Can zero-shot models be fine-tuned without reducing out-of-distribution accuracy?*

As pre-trained models are becoming a cornerstone of machine learning, techniques for fine-tuning them on downstream applications are increasingly important. Indeed, the question of robustly fine-tuning pre-trained models has recently also been raised as an open problem by several authors [1, 4, 20]. Andreassen *et al.* [1] explored several fine-tuning approaches but found that none yielded models with improved robustness at high accuracy. Furthermore, Taori *et al.* [23] demonstrated that no current algorithmic robustness interventions provide consistent gains across the distribution shifts where zero-shot CLIP excels.

In this paper, we conduct an empirical investigation to understand and improve fine-tuning of zero-shot models from a distributional robustness perspective. First, we measure how different fine-tuning approaches (last-layer vs. end-to-end fine-tuning, hyperparameter changes, etc.) affect the out-of-distribution accuracy of the resulting fine-tuned models. Our empirical analysis uncovers two key issues in the standard fine-tuning process. The robustness of fine-tuned models substantially varies under even small changes in hyperparameters, but the best hyperparameters cannot be inferred from in-distribution accuracy alone. In addition, more aggressive fine-tuning (e.g., using a larger step size) yields larger in-distribution improvements but can also reduce out-of-distribution accuracy by a larger amount.

Motivated by the above concerns, we propose a robust way of fine-tuning zero-shot models that addresses the aforementioned trade-off and achieves the best of both worlds: increased performance out-of-distribution while maintaining or even improving in-distribution accuracy relative to standard fine-tuning. In addition, our method simplifies the choice of hyperparameters in the fine-tuning process.

Our method (Figure 1) has two steps: first, we fine-tune the zero-shot model on application-specific data. Second, we combine the original zero-shot and fine-tuned models by linearly interpolating between their weights, which we refer to as weight-space ensembling. Interpolating models dates back

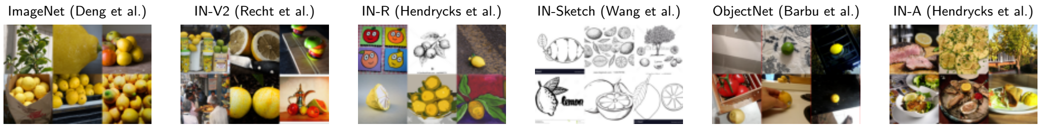

Figure 2: Samples of the class *lemon*, from ImageNet [8] (in-distribution) and the derived out-of-distribution datasets considered in our main experiments: ImageNet-V2 [21], ImageNet-R [10], ImageNet Sketch [24], ObjectNet [2], and ImageNet-A [11].

to early work in convex optimization by Polyak and Juditsky [19]. Here, we empirically study model interpolation for non-convex models from the perspective of distributional robustness. Interestingly, linear interpolation in weight space still succeeds despite the non-linearity in the activation functions of the neural networks.

Weight-space ensembles for fine-tuning (WiSE-FT) substantially improve out-of-distribution accuracy compared to prior work while maintaining high in-distribution performance. Concretely, on ImageNet [8] and five of the natural distribution shifts studied by Radford et al. [20], WiSE-FT applied to standard end-to-end fine-tuning improves out-of-distribution accuracy by 4 to 6 percentage points (pp) over prior work while maintaining or improving the in-distribution accuracy of the fine-tuned model. Relative to the zero-shot model, WiSE-FT improves out-of-distribution accuracy by 1 to 9 pp. Moreover, WiSE-FT improves over a range of alternative approaches such as regularization and evaluating at various points throughout fine-tuning. These robustness gains come at no additional computational cost during fine-tuning or inference.

To understand the robustness gains of WiSE-FT, we first study WiSE-FT when fine-tuning a linear classifier (last layer) as it is amenable to analysis. In this linear case, our procedure is equivalent to ensembling the outputs of two models, and experiments point towards the complementarity of model predictions as a key property. For end-to-end fine-tuning, we connect our observations to earlier work on the phenomenology of deep learning. Neyshabur *et al.* [18] found that end-to-end fine-tuning the same model twice yielded two different solutions that were connected via a linear path in weight space along which error remains low, known as linear mode connectivity [9]. Our observations suggest a similar phenomenon along the path generated by WiSE-FT, but the exact shape of the loss landscape and connection between in- and out-of-distribution error are still an open problem.

In addition to the aforementioned ImageNet distribution shifts, WiSE-FT consistently improves robustness on a diverse set of six further distribution shifts including: (i) geographic shifts in satellite imagery and wildlife recognition (WILDS-FMoW, WILDS-iWildCam) [13, 6, 3], (ii) reproductions of the popular image classification dataset CIFAR-10 with a distribution shift (CIFAR-10.1 and CIFAR-10.2) [21, 16], and (iii) datasets with distribution shift induced by temporal perturbations in videos (ImageNet-Vid-Robust and YTBB-Robust) [22]. Beyond the robustness perspective, WiSE-FT also improves in-distribution performance compared to standard fine-tuning, reducing the relative error rate by 4-49% on a range of seven datasets: ImageNet, CIFAR-10, CIFAR-100 [15], Describable Textures [7], Food-101 [5], SUN397 [25], and Stanford Cars [14]. Even when fine-tuning data is scarce, reflecting many application scenarios, we find that WiSE-FT improves performance.

Overall, WiSE-FT is simple, universally applicable in the problems we studied, and can be implemented in a few lines of code. Hence we encourage its adoption for fine-tuning zero-shot models.

References

[1] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning, 2021. https://arxiv.org/abs/2106.15831.

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

[3] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR) FGVC8 Workshop*, 2021. https://arxiv.org/abs/2105.03494.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models, 2021. https://arxiv.org/abs/2108.07258.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014. https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

[6] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1711.07846.

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. https://arxiv.org/abs/1311.3618.

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

[9] 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://arxiv.org/abs/1912.05671.

[10] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *International Conference on Computer Vision (ICCV)*, 2021. https://arxiv.org/abs/2006.16241.

[11] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. https://arxiv.org/abs/1907.07174.

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2102.05918.

[13] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2012.07421.

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV) Workshops*, 2013. https://ieeexplore.ieee.org/document/6755945.

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[16] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *International Conference on Machine Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning*, 2020. http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

[17] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2107.04649.

[18] 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.

[19] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 1992. https://epubs.siam.org/doi/abs/10.1137/0330046?journalCode=sjcodc.

[20] 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. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2103.00020.

[21] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.10811.

[22] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time?, 2019. https://arxiv.org/abs/1906.02168.

[23] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2007.00644.

[24] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.13549.

[25] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 2016. https://link.springer.com/article/10.1007/s11263-014-0748-y.