Robust fine-tuning of zero-shot models

Mitchell Wortsman*
University of Washington
mitchnw@cs.washington.edu

Gabriel Ilharco*
University of Washington
gamaga@cs.washington.edu

Jong Wook Kim
OpenAI
jongwook@openai.com

Mike Li
Columbia University
mli24@gsb.columbia.edu

Simon Kornblith
Google Research, Brain Team
skornblith@google.com

Rebecca Roelofs
Google Research, Brain Team
rofls@google.com

Raphael Gontijo Lopes
Google Research, Brain Team
iraphael@google.com

Hannaneh Hajishirzi
University of Washington
hannaneh@cs.washington.edu

Ali Farhadi*
University of Washington
ali@cs.washington.edu

Hongseok Namkoong*
Columbia University
namkoong@gsb.columbia.edu

Ludwig Schmidt
University of Washington
schmidt@cs.washington.edu

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 methods substantially improve accuracy on a given target distribution, they often reduce robustness to distribution shifts. We address this tension by introducing a simple and effective method for improving robustness while fine-tuning: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements under distribution shift, while preserving high accuracy on the target distribution. On ImageNet and five derived distribution shifts, WiSE-FT improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while increasing ImageNet accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness gains (2 to 23 pp) on a diverse set of six further distribution shifts, and accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.

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 distribution shifts on which current algorithmic approaches to enhance robustness yield little to no gains [68, 95]. While these negative results highlight the difficulty of learning robust models, large pre-trained models such as CLIP [79], ALIGN [44] and BASIC [75] have recently demonstrated unprecedented robustness to these challenging distribution shifts. The success of these models 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 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*. [79] and Pham *et al*. [75], fine-tuning comes at the cost of robustness: across several natural distribution shifts, the accuracy of their fine-tuned 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 accuracy under distribution shift?*

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 [3, 9, 75, 79]. Andreassen *et al*. [3] explored several fine-tuning approaches but found that none yielded models

*These authors contributed equally. ArXiv version: 2109.01903.

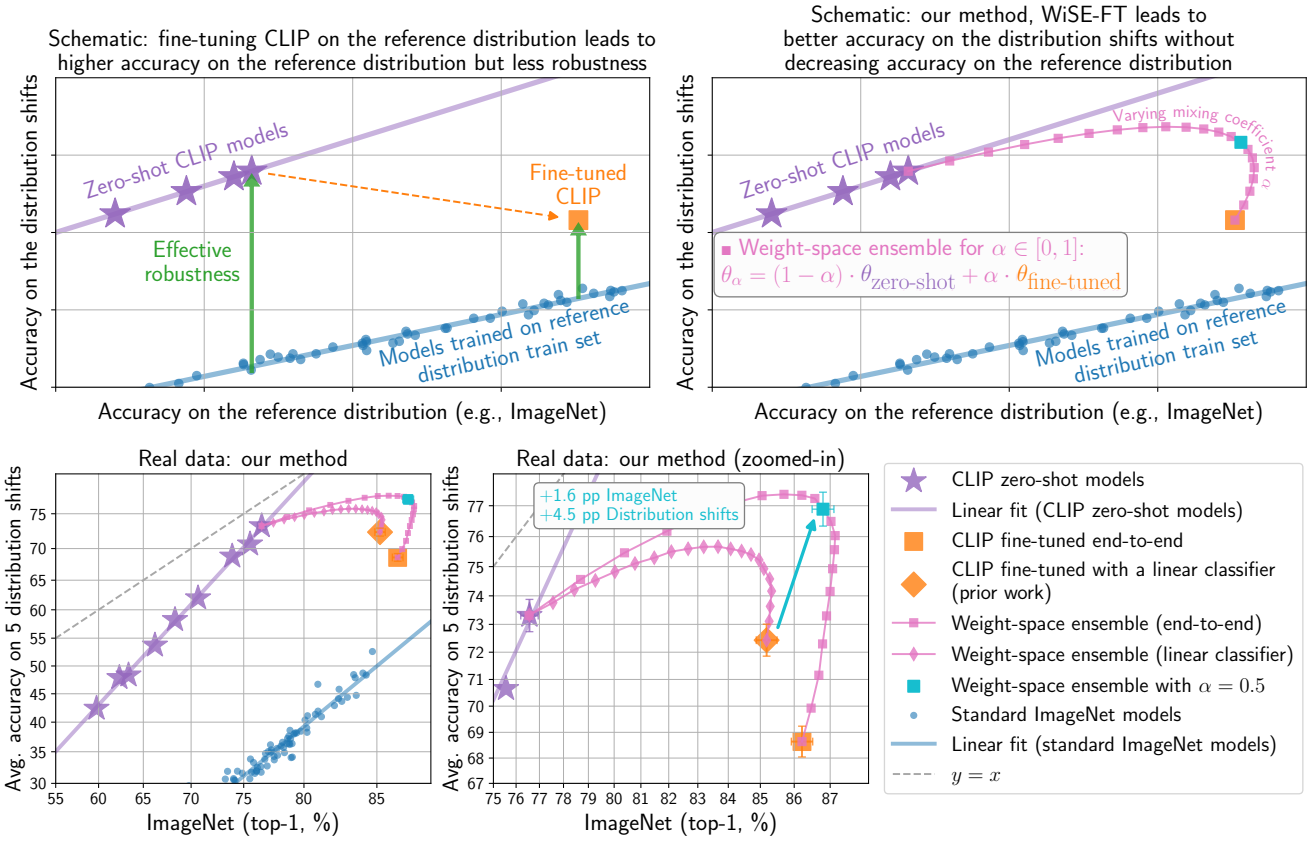

Figure 1. **(Top left)** Zero-shot CLIP models exhibit moderate accuracy on the reference distribution (x-axis, the target for fine-tuning) and high effective robustness (accuracy on the distribution shifts beyond the baseline models). In contrast, standard fine-tuning—either end-to-end or with a linear classifier (final layer)—attains higher accuracy on the reference distribution but 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 accuracy relative to both the zero-shot and fine-tuned models while maintaining or improving accuracy on ImageNet.

with improved robustness at high accuracy. Furthermore, Taorio *et al*. [95] demonstrated that no current algorithmic robustness interventions provide consistent gains across the distribution shifts where zero-shot models excel.

In this paper, we conduct an empirical investigation to understand and improve fine-tuning of zero-shot models from a distributional robustness perspective. We begin by measuring how different fine-tuning approaches (last-layer vs. end-to-end fine-tuning, hyperparameter changes, etc.) affect the accuracy under distribution shift of the resulting fine-tuned models. Our empirical analysis uncovers two key issues in the standard fine-tuning process. First, the robustness of fine-tuned models varies substantially under even small changes in hyperparameters, but the best hyperparameters cannot be inferred from accuracy on the target distribution alone. Second, more aggressive fine-tuning (e.g., using a larger learning rate) yields larger accuracy improvements on the target distribution, but can also reduce accuracy under distribution shift by a large 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 under distribution shift while maintaining or even improving accuracy on the target distribution 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 the target distribution. 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 model parameters is a classical idea in convex optimization dating back decades (e.g., see [76, 82]). 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 accuracy under distribution shift compared to prior work while maintaining high performance on the target distribution. Concretely, on ImageNet [17] and five of the natural distribution shifts studied by Radford *et al*. [79], WiSE-FT applied to standard end-to-end fine-tuning improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while maintaining or improving the ImageNet accuracy of the fine-tuned CLIP model. Relative to the zero-shot model, WiSE-FT improves accuracy under distribution shift 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.

While our investigation centers around CLIP, we observe similar trends for other zero-shot models including ALIGN [44], BASIC [75], and a ViT model pre-trained on JFT [21]. For instance, WiSE-FT improves the ImageNet accuracy of a fine-tuned BASIC-L model by 0.4 pp, while improving average accuracy under distribution shift by 2 to 11 pp.

To understand the robustness gains of WiSE-FT, we first study WiSE-FT when fine-tuning a linear classifier (last layer) as it is more 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*. [71] 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 [25]. Our observations suggest a similar phenomenon along the path generated by WiSE-FT, but the exact shape of the loss landscape and connection between error on the target and shifted distributions are still open problems (analysis in Appendix A).

In addition to the aforementioned ImageNet distribution shifts, WiSE-FT consistently improves robustness on a diverse set of six additional distribution shifts including: (i) geographic shifts in satellite imagery and wildlife recognition (WILDS-FMoW, WILDS-iWildCam) [6, 13, 47], (ii) reproductions of the popular image classification dataset CIFAR-10 with a distribution shift (CIFAR-10.1 and CIFAR-10.2) [60,81], and (iii) datasets with distribution shift induced by temporal perturbations in videos (ImageNet-Vid-Robust and YTBB-Robust) [86]. Beyond the robustness perspective, WiSE-FT also improves accuracy compared to standard fine-tuning, reducing the relative error rate by 4-49% on a range of seven datasets: ImageNet, CIFAR-10, CIFAR-100 [52], Describable Textures [14], Food-101 [10], SUN397 [101], and Stanford Cars [51]. 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.

2. Background and experimental setup

Our experiments compare the performance of zero-shot models, corresponding fine-tuned models, and models produced by WiSE-FT. To measure robustness, we contrast model accuracy on two related but different distributions, a reference distribution $\mathcal{D}_{\text{ref}}$ which is the target for fine-tuning, and shifted distribution $\mathcal{D}_{\text{shift}}$.[1] We assume both distributions have test sets for evaluation, and $\mathcal{D}_{\text{ref}}$ has an associated training set $\mathcal{S}^{\text{tr}}_{\text{ref}}$ which is typically used for training or fine-tuning. The goal for a model is to achieve both high accuracy and consistent performance on the two distributions $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{shift}}$. This is a natural goal as humans often achieve similar accuracy across the distribution shifts in our study [87].

For a model f, we let $\text{Acc}_{\text{ref}}(f)$ and $\text{Acc}_{\text{shift}}(f)$ refer to classification accuracy on the reference and shifted test sets, respectively. We consider k-way image classification, where x_i is an image with corresponding label $y_i \in \{1, ..., k\}$. The outputs of f are k-dimensional vectors of non-normalized class scores.

Distribution shifts. Taori *et al*. [95] categorized distribution shifts into two broad categories: (i) *synthetic*, e.g., ℓ_∞-adversarial examples or artificial changes in image contrast, brightness, etc. [2, 7, 8, 29, 36]; and (ii) *natural*, where samples are not perturbed after acquisition and changes in data distributions arise through naturally occurring variations in lighting, geographic location, crowdsourcing process, image styles, etc. [35, 38, 47, 81, 95]. Following Radford *et al*. [80], our focus here is on natural distribution shifts as they are more representative of the real world when no active adversary is present. Specifically, we present our key results for five natural distribution shifts derived from ImageNet (i.e., $\mathcal{S}^{\text{tr}}_{\text{ref}}$ is ImageNet): (a) ImageNet-V2 (IN-V2) [81], a reproduction of the ImageNet test set with distribution shift (b) ImageNet-R (IN-R) [35], renditions (e.g., sculptures, paintings) for 200 ImageNet classes (c) ImageNet Sketch (IN-Sketch) [98], which contains sketches instead of natural images (d) ObjectNet [4], a test set of objects in various scenes with 113 classes overlapping with ImageNet (e) ImageNet-A (IN-A) [38], a test set of natural images mis-

[1] $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{shift}}$ are sometimes referred to as *in-distribution* (ID) and *out-of-distribution* (OOD). In this work, we include evaluations of zero-shot models, which are *not* trained on data from the reference distribution, so referring to $\mathcal{D}_{\text{ref}}$ would be imprecise. For clarity, we avoid the ID/OOD terminology.

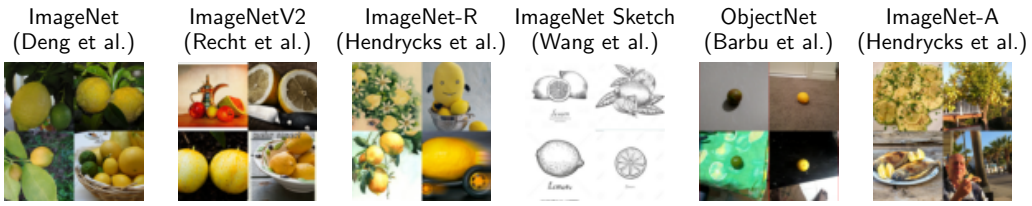

Figure 2. Samples of the class *lemon*, from the reference distribution ImageNet [17] and the derived distribution shifts considered in our main experiments: ImageNet-V2 [81], ImageNet-R [35], ImageNet Sketch [98], ObjectNet [4], and ImageNet-A [38].

classified by a ResNet-50 [34] for 200 ImageNet classes. Figure 2 illustrates the five distribution shifts.

Effective robustness and scatter plots. To compare the robustness of models with different accuracies on the reference distribution, we follow the *effective robustness* framework introduced by Taori *et al*. [95]. Effective robustness quantifies robustness as accuracy *beyond a baseline* trained only on the reference distribution. A useful tool for studying (effective) robustness are scatter plots that illustrate model performance under distribution shift [81, 95]. These scatter plots display accuracy on the reference distribution on the x-axis and accuracy under distribution shift on the y-axis, i.e., a model f is shown as a point $(\mathsf{Acc}_{\text{ref}}(f),\ \mathsf{Acc}_{\text{shift}}(f))$. Figure 1 exemplifies these scatter plots with both schematics and real data. For the distribution shifts we study, accuracy on the reference distribution is a reliable predictor of accuracy under distribution shift [68, 95]. In other words, there exists a function $\beta : [0, 1] \to [0, 1]$ such that $\mathsf{Acc}_{\text{shift}}(f)$ approximately equals $\beta(\mathsf{Acc}_{\text{ref}}(f))$ for models f trained on the train set $\mathcal{S}^{\text{tr}}_{\text{ref}}$. Effective robustness [95] is accuracy beyond this baseline, defined formally as $\rho(f) = \mathsf{Acc}_{\text{shift}}(f) - \beta(\mathsf{Acc}_{\text{ref}}(f))$.

In the corresponding scatter plots, effective robustness is vertical movement above expected accuracy under distribution shift (Figure 1, top). Effective robustness thereby disentangles accuracy changes on the reference distribution from the effect of robustness interventions. When we say that a model is robust to distribution shift, we mean that effective robustness is positive. Taori *et al*. [95] observed that no algorithmic robustness intervention consistently achieves substantial effective robustness across the distribution shifts in Figure 2—the first method to do so was zero-shot CLIP. Empirically, when applying logit (or probit) axis scaling, models trained on the reference distribution approximately lie on a linear trend [68, 95]. As in Taori *et al*. [95], we apply logit axis scaling and show 95% Clopper-Pearson confidence intervals for the accuracies of select points.

Zero-shot models and CLIP. We primarily explore CLIP models [79], although we also investigate other zero-shot models including ALIGN [44], BASIC [75] and a ViT model pre-trained on JFT [21]. Zero-shot models exhibit effective robustness and lie on a qualitatively different linear trend (Figure 1). CLIP-like models are pre-trained using image-caption pairs from the web. Given a set of image-caption pairs $\{(x_1, s_1)..., (x_B, s_B)\}$, CLIP-like models train an image-encoder g and text-encoder h such that the similarity $\langle g(x_i), h(s_i)\rangle$ is maximized relative to unaligned pairs. CLIP-like models perform zero-shot k-way classification given an image x and class names $C = \{c_1, ..., c_k\}$ by matching x with potential captions. For instance, using caption $s_i =$ "a photo of a $\{c_i\}$" for each class i, the zero-shot model predicts the class via $\arg\max_j \langle g(x), h(s_j)\rangle$.[2] Equivalently, one can construct $\mathbf{W}_{\text{zero-shot}} \in \mathbb{R}^{d\times k}$ with columns $h(s_j)$ and compute outputs $f(x) = g(x)^\top \mathbf{W}_{\text{zero-shot}}$. Unless explicitly mentioned, our experiments use the CLIP model `ViT-L/14@336px`, although all CLIP models are displayed in our scatter plots (additional details provided in Appendix F.1).

3. Weight-space ensembles for fine-tuning

This section describes and motivates our proposed method, WiSE-FT, which consists of two simple 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, also referred to as weight-space ensembling. WiSE-FT can be implemented in a few lines of PyTorch, and we provide example code in Appendix C.

The zero-shot model excels under distribution shift while standard fine-tuning achieves high accuracy on the reference distribution. Our motivation is to combine these two models into one that achieves the best of both worlds. Weight-space ensembles are a natural choice as they ensemble without extra computational cost. Moreover, previous work has suggested that interpolation in weight space may improve performance when models share part of their optimization trajectory [42, 71].

Step 1: Standard fine-tuning. As in Section 2, we let $\mathcal{S}^{\text{tr}}_{\text{ref}}$ denote the dataset used for fine-tuning and g denote the image encoder used by CLIP. We are now explicit in writing $g(x, \mathbf{V}_{\text{enc}})$ where x is an input image and

[2] For improved accuracy, the embedding of a few candidate captions are averaged, e.g., $s_i^{(1)}$="a *photo* of a $\{c_i\}$" and $s_i^{(2)}$="a *picture* of a $\{c_i\}$" (referred to as prompt ensembling [79]).

$\mathbf{V}_{\text{enc}}$ are the parameters of the encoder g. Standard fine-tuning considers the model $f(x, \theta) = g\left(x, \mathbf{V}_{\text{enc}}\right)^\top \mathbf{W}_{\text{classifier}}$ where $\mathbf{W}_{\text{classifier}} \in \mathbb{R}^{d \times k}$ is the classification head and $\theta = [\mathbf{V}_{\text{enc}}, \mathbf{W}_{\text{classifier}}]$ are the parameters of f. We then solve $\arg\min_\theta \left\{ \sum_{(x_i, y_i) \in \mathcal{S}^{\text{tr}}_{\text{ref}}} \ell(f(x_i, \theta),\ y_i) + \lambda R(\theta) \right\}$ where ℓ is the cross-entropy loss and R is a regularization term (e.g., weight decay). We consider the two most common variants of fine-tuning: end-to-end, where all values of θ are modified, and fine-tuning only a linear classifier, where $\mathbf{V}_{\text{enc}}$ is fixed at the value learned during pre-training. Appendices F.2 and F.3 provide additional details.

Step 2: Weight-space ensembling. For a *mixing coefficient* $\alpha \in [0, 1]$, we consider the *weight-space ensemble* between the zero-shot model with parameters θ_0 and the model obtained via standard fine-tuning with parameters θ_1. The predictions of the weight-space ensemble wse are given by

$$\mathsf{wse}(x, \alpha) = f(x, (1 - \alpha) \cdot \theta_0 + \alpha \cdot \theta_1)\,, \qquad (1)$$

i.e., we use the element-wise weighted average of the zero-shot and fined-tuned parameters. When fine-tuning only the linear classifier, weight-space ensembling is equivalent to the traditional output-space ensemble [11, 20, 26] $(1 - \alpha) \cdot f(x, \theta_0) + \alpha \cdot f(x, \theta_1)$ since Equation 1 decomposes as $(1 - \alpha) \cdot g(x, \mathbf{V}_{\text{enc}})^\top \mathbf{W}_{\text{zero-shot}} + \alpha \cdot g(x, \mathbf{V}_{\text{enc}})^\top \mathbf{W}_{\text{classifier}}$.

As neural networks are non-linear with respect to their parameters, ensembling all layers—as we do when end-to-end fine-tuning—typically fails, achieving no better accuracy than a randomly initialized neural network [25]. However, as similarly observed by previous work where part of the optimization trajectory is shared [25, 42, 71], we find that the zero-shot and fine-tuned models are connected by a linear path in weight-space along which accuracy remains high (explored further in Section A.2).

Remarkably, as we show in Section 4, WiSE-FT improves accuracy under distribution shift while maintaining high performance on the reference distribution relative to fine-tuned models. These improvements come without any additional computational cost as a single set of weights is used.

4. Results

This section presents our key experimental findings. First, we show that WiSE-FT boosts the accuracy of a fine-tuned CLIP model on five ImageNet distribution shifts studied by Radford *et al*. [79], while maintaining or improving ImageNet accuracy. Next, we present additional experiments, including more distribution shifts, the effect of hyperparameters, accuracy improvements on the reference distribution, and experiments in the low-data regime. Finally, we demonstrate that our findings are more broadly applicable by exploring WiSE-FT for BASIC [75], ALIGN [44], and a ViT-H/14 [21] model pre-trained on JFT-300M [91].

Main results: ImageNet and associated distribution shifts. As illustrated in Figure 1, when the mixing coefficient α varies from 0 to 1, $\mathsf{wse}(\cdot, \alpha)$ is able to simultaneously improve accuracy on both the reference and shifted distributions. A breakdown for each dataset is shown in Appendix E.1. Table 1 presents our main results on ImageNet and five derived distribution shifts. WiSE-FT (end-to-end, α=0.5) outperforms numerous strong models in both average accuracy under distribution shift and the average accuracy on the reference and shifted distributions. While future work may lead to more sophisticated strategies for choosing the mixing coefficient α, α=0.5 yields close to optimal performance across a range of experiments. Hence, we recommend α=0.5 when no domain knowledge is available. Appendix D further explores the effect of α. Moreover, results for 12 other backbones are shown in Appendix E.

Robustness on additional distribution shifts. Beyond the five distribution shifts derived from ImageNet, WiSE-FT consistently improves robustness on a diverse set of further distributions shifts including geographic shifts in satellite imagery and wildlife recognition (WILDS-FMoW [13, 47], WILDS-iWildCam [6, 47]), reproductions of the popular image classification dataset CIFAR-10 [52] with a distribution shift (CIFAR-10.1 [81] and CIFAR-10.2 [60]), and datasets with distribution shift induced by temporal perturbations in videos (ImageNet-Vid-Robust and YTBB-Robust [87]). Concretely, WiSE-FT (α=0.5) improves performance under distribution shift by 3.5, 6.2, 1.7, 2.1, 9.0 and 23.2 pp relative to the fine-tuned solution while decreasing performance on the reference distribution by at most 0.3 pp (accuracy on the reference distribution often improves). In contrast to the ImageNet distribution shifts, the zero-shot model initially achieves less than 30% accuracy on the WILDS distribution shifts, and WiSE-FT provides improvements regardless. Appendix E.2 (Figure 9 and Table 6) includes more detailed results.

Hyperparameter variation and alternatives. As illustrated by Figure 3, moderate changes in standard hyperparameters such as the learning rate or the number of epochs can substantially affect performance under distribution shift. Moreover, these performance differences cannot be detected reliably from model performance on reference data alone. For instance, while training for 10 epochs with learning rate $3 \cdot 10^{-5}$ and $3 \cdot 10^{-6}$ lead to a small accuracy difference on ImageNet (0.3 pp), accuracy under distribution shift varies by as much as 8 pp.

Furthermore, tuning hyperparameters on ImageNet data can also reduce robustness. For instance, while moving from small to moderate learning rates (10^{-7} to $3 \cdot 10^{-5}$) improves

	IN (reference)	Distribution shifts IN-V2	IN-R	IN-Sketch	ObjectNet*	IN-A	Avg shifts	Avg ref., shifts
CLIP `ViT-L/14@336px`								
Zero-shot [79]	76.2	70.1	88.9	60.2	70.0	77.2	73.3	74.8
Fine-tuned LC [79]	85.4	75.9	84.2	57.4	66.2	75.3	71.8	78.6
Zero-shot (PyTorch)	76.6	70.5	89.0	60.9	69.1	77.7	73.4	75.0
Fine-tuned LC (ours)	85.2	75.8	85.3	58.7	67.2	76.1	72.6	78.9
Fine-tuned E2E (ours)	86.2	76.8	79.8	57.9	63.3	65.4	68.6	77.4
WiSE-FT (ours)								
LC, α=0.5	83.7	76.3	89.6	63.0	70.7	79.7	75.9	79.8
LC, optimal α	85.3	76.9	89.8	63.0	70.7	79.7	75.9	80.2
E2E, α=0.5	86.8	**79.5**	89.4	64.7	71.1	79.9	76.9	81.8
E2E, optimal α	**87.1**	**79.5**	**90.3**	**65.0**	**72.1**	**81.0**	**77.4**	**81.9**

Table 1. Accuracy of various methods on ImageNet and derived distribution shifts for CLIP `ViT-L/14@336px` [79]. E2E: end-to-end; LC: linear classifier. *Avg shifts* displays the mean performance among the five distribution shifts, while *Avg reference, shifts* shows the average of ImageNet (reference) and Avg shifts. For optimal α, we choose the single mixing coefficient that maximizes the column. Results for additional models are provided in Appendix E.7.

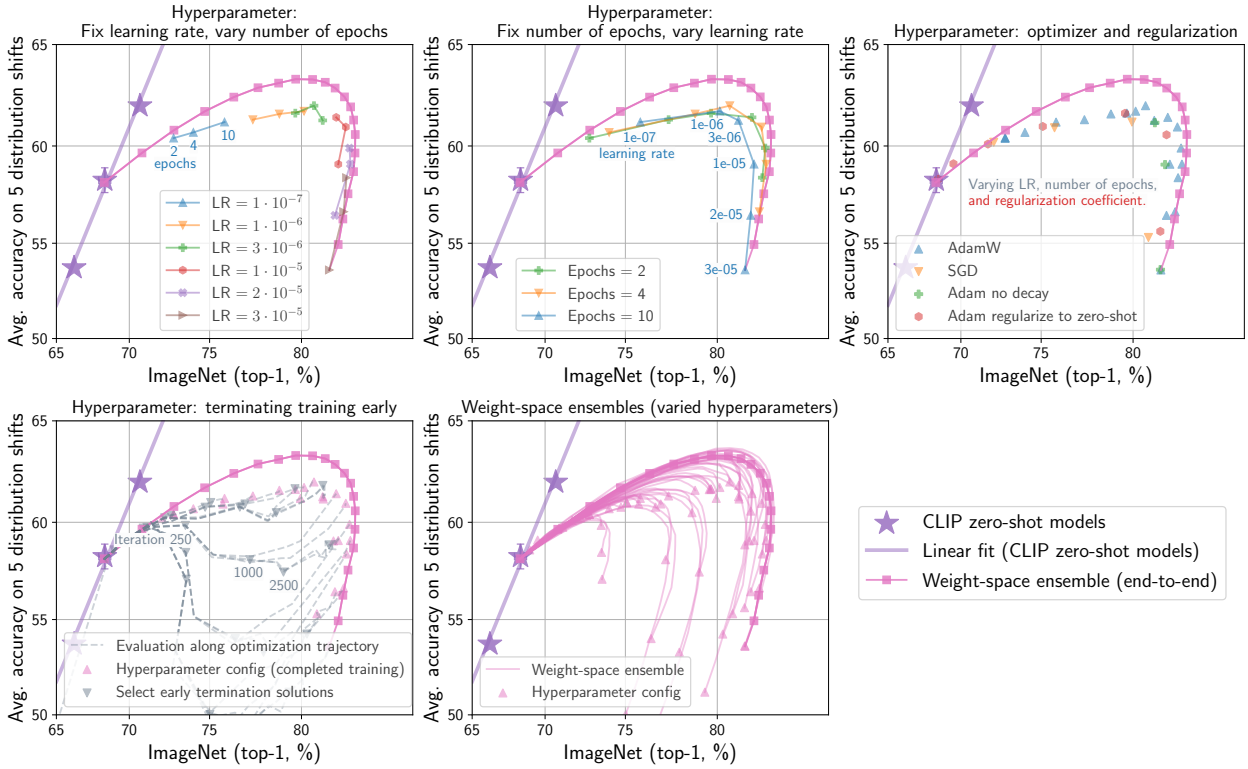

Figure 3. The robustness of fine-tuned models varies substantially under even small changes in hyperparameters. Applying WiSE-FT addresses this brittleness and can remove the trade-off between accuracy on the reference and shifted distributions. Results shown for CLIP `ViT-B/16` fine-tuned with cosine-annealing learning rate schedule and all models in the top left and top middle plots are fine-tuned with AdamW [59]. Moreover, *regularize to zero-shot* appends the regularizer $\lambda\|\theta - \theta_0\|_2^2$ to the fine-tuning objective, where θ_0 are the parameters of the zero-shot model.

performance on ImageNet by 5 pp, it also deteriorates accuracy under distribution shift by 8 pp.

WiSE-FT addresses this brittleness of hyperparameter tuning: even when using a learning rate $3 \cdot 10^{-5}$ where standard fine-tuning leads to low robustness, applying WiSE-FT removes the trade-off between accuracy on the reference and shifted distributions. The models which can be achieved by varying α are as good or better than those achievable by other hyperparameter configurations. Then, instead of searching over a wide range of hyperparameters, only α needs to be

considered. Moreover, evaluating different values of α does not require training new models.

There is no hyperparameter in Figure 3 which can be varied to match or exceed the optimal curve produced by WiSE-FT. In our experiments, this frontier is reached only through methods that average model weights, either using WiSE-FT or with a more sophisticated averaging scheme: keeping an exponential moving average of all model iterates (EMA, [93]). Comparisons with EMA are detailed in Appendix E.3.2.

Additional comparisons are also presented in Appendix E.3, including distillation, additional regularization, and CoOp [110]. Finally, Appendix E.4 recreates Figure 3 with stronger data augmentation and finds similar trends.

Accuracy gains on reference distributions. Beyond robustness to distribution shift, Table 2 demonstrates that WiSE-FT also improves accuracy after fine-tuning on seven datasets. When fine-tuning end-to-end on ImageNet, CIFAR-10, CIFAR-100, Describable Textures, Food-101, SUN397, and Stanford Cars, WiSE-FT reduces relative error by 4 to 49%. Even though standard fine-tuning directly optimizes for high accuracy on the reference distribution, WiSE-FT achieves better performance. Appendix E.5 includes more details, including explorations in the low-data regime.

Beyond CLIP. Figure 4 illustrates that WiSE-FT is generally applicable to zero-shot models beyond CLIP, and beyond models pre-trained contrastively with image-text pairs. First, we interpolate between the weights of the zero-shot and fine-tuned BASIC-L model [75], finding that α=0.5 improves average accuracy on five distribution shifts derived from ImageNet by over 7 pp while improving ImageNet accuracy by 0.4 pp relative to the fine-tuned BASIC-L model (a per-dataset breakdown is provided in Figure 23 and Table 12 of the Appendix). As in Pham *et al*. [75], the model is fine-tuned using a contrastive loss and half of the ImageNet training data. WiSE-FT provides improvements on both reference and shifted distributions, despite these experimental differences.

Next, we consider the application of WiSE-FT to a ViT-H/14 model [21] pre-trained on JFT-300M [91], where the zero-shot classifier is constructed by manually identifying a class correspondence (details provided in Section E.7.2). WiSE-FT improves performance under distribution shift over both the zero-shot and fine-tuned models. When α=0.8, WiSE-FT outperforms the fine-tuned model by 2.2 pp on distribution shifts, while maintaining ImageNet performance within 0.2 pp of the fine-tuned model. This result demonstrates that WiSE-FT can be successfully applied even to models which do not use contrastive image-text pre-training.

Finally, we apply WiSE-FT to the ALIGN model of Jia *et al*. [44], which is similar to CLIP but is pre-trained with a different dataset, finding similar trends.

[*] Although this table considers ImageNet class names, ObjectNet provides alternative class names which can improve the performance of zero-shot CLIP by 2.3 percentage points (Appendix F.4).

5. Related work

Concurrent and subsequent related work is in Appendix B.

Robustness. Understanding how models perform under distribution shift remains an important goal, as real world models may encounter data from new environments [78, 96]. Previous work has studied model behavior under synthetic [2, 23, 29, 36, 63, 97] and natural distribution shift [4, 35, 38, 47, 98]. Interventions used for synthetic shifts do not typically provide robustness to many natural distribution shifts [95]. In contrast, accuracy on the reference distribution is often a reliable predictor for accuracy under distribution shift [67, 68, 92, 95, 104]. On the other hand, D'Amour *et al*. [16] show that accuracy under certain distribution shifts cannot be reliably inferred from accuracy on the reference distribution. We observe a similar phenomenon when fine-tuning with different hyperparameters (Section 4, Figure 3).

Pre-training and transfer learning. Pre-training on large amounts of data is a powerful technique for building high-performing machine learning systems [12, 21, 48, 80, 88, 105]. One increasingly popular class of vision models are those pre-trained with auxiliary language supervision, which can be used for zero-shot inference [18, 44, 75, 79, 84, 107, 109]. When pre-trained models are adapted to a specific distribution through standard fine-tuning, effective robustness deteriorates at convergence [3]. In natural language processing, previous work proposed stable fine-tuning methods that incur computational overhead [45, 111], alleviating problems such as representational collapse [1]. More generally, a variety of methods have attempted to mitigate catastrophic forgetting [65]. Kirkpatrick *et al*. [46]; Zenke *et al*. [106] explored weighted quadratic regularization for sequential learning. Xuhong*et al*. [103] showed that, for fine-tuning, the simple quadratic regularization explored in Section 4 performs best, while Lubana *et al*. [61] explored the connection between quadratic regularization and interpolation. Andreassen *et al*. [3] found that many approaches from continual learning do not provide robustness to multiple natural distribution shifts. Finally, Li *et al*. [57] investigate the effect of fine-tuning hyperparameters on performance.

Traditional (output-space) ensembles. Traditional ensemble methods, which we refer to as output-space ensembles, combine the predictions (outputs) of many classifiers [5, 11, 20, 26, 27, 56]. Typically, output-space ensembles outperform individual classifiers and provide uncertainty es-

	ImageNet	CIFAR10	CIFAR100	Cars	DTD	SUN397	Food101
Standard fine-tuning	86.2	98.6	92.2	91.6	81.9	80.7	94.4
WiSE-FT (α=0.5)	86.8 (+0.6)	99.3 (+0.7)	93.3 (+1.1)	93.3 (+1.7)	84.6 (+2.8)	83.2 (+2.5)	96.1 (+1.6)
WiSE-FT (opt. α)	87.1 (+0.9)	99.5 (+0.8)	93.4 (+1.2)	93.6 (+2.0)	85.2 (+3.3)	83.3 (+2.6)	96.2 (+1.8)

Table 2. Beyond robustness, WiSE-FT can improve accuracy after fine-tuning on several datasets.

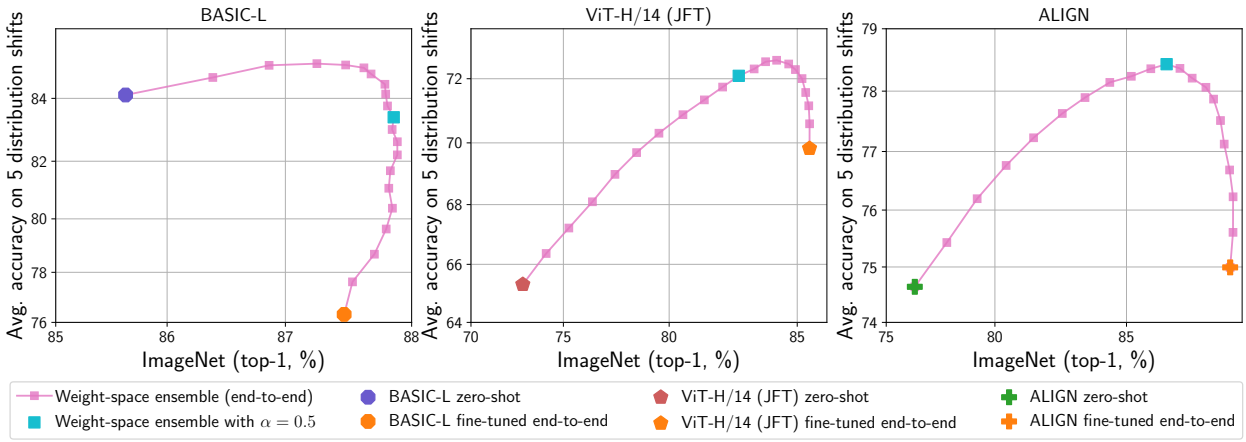

Figure 4. WiSE-FT applied to BASIC-L [75], a ViT-H/14 [21] model pre-trained on JFT-300M [91] and ALIGN [44].

timates under distribution shift that are more callibrated than baselines [56, 73, 90]. In contrast to these works, we consider the ensemble of two models which have observed different data. Output-space ensembles require more computational resources as they require a separate pass through each model. Compared to an ensemble of 15 models trained on the same dataset, Mustafa *et al*. [70] find an improvement of 0.8–1.6 pp under distribution shift (on ImageNetV2, ImageNet-R, ObjectNet, and ImageNet-A) by ensembling a similar number of models pre-trained on different datasets. In contrast, we see an improvement of 2–15 pp from ensembling two models. Moreover, as we ensemble in weight-space, no extra compute is required compared to a single model.

Weight-space ensembles. Weight-space ensembles linearly interpolate between the weights of different models [25, 32, 62, 93]. For example, Izmailov *et al*. [42] average checkpoints saved throughout training for improved performance. Indeed, averaging the weights along the training trajectory is a central method in optimization [72, 77, 82]. For instance, Zhang *et al*. [108] propose optimizing with a set of fast and slow weights, where every k steps, these two sets of weights are averaged and a new trajectory begins. Here, we revisit these techniques from a distributional robustness perspective and consider the weight-space ensemble of models which have observed different data.

6. Limitations, impact, and conclusion

Limitations. While we expect our findings to be more broadly applicable to other domains such as natural language processing, our investigation here is limited to image classification. Exploring fine-tuning for object detection and natural language processing are interesting directions for future work. Moreover, although the interpolation parameter setting α=0.5 provides good overall performance, we leave the question of finding the optimal α for specific target distributions to future work.

Impact. Radford *et al*. [79] and Brown *et al*. [12] extensively discuss the broader impact of large zero-shot models and identify potential causes of harm including model biases and potential malicious uses such as surveillance systems. WiSE-FT is a fine-tuning method that builds on such models, and thus may perpetuate their negative impact.

Conclusion. We view WiSE-FT as a first step towards more sophisticated fine-tuning schemes and anticipate that future work will continue to leverage the robustness of zero-shot models for building more reliable neural networks.

Acknowledgements

We thank Anders Andreassen, Tim Dettmers, Jesse Dodge, Katie Everett, Samir Gadre, Ari Holtzman, Sewon Min, Mohammad Norouzi, Nam Pho, Ben Poole, Sarah Pratt, Alec Radford, Jon Shlens, and Rohan Taori for helpful discussions and draft feedback, Hyak at UW for computing support, and Basil Mustafa for providing an earlier version of the mapping between JFT and ImageNet classes. This work is in part supported by NSF IIS 1652052, IIS 17303166, DARPA N66001-19-2-4031, DARPA W911NF-15-1-0543 and gifts from Allen Institute for Artificial Intelligence.

References

[1] Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations (ICLR)*, 2021. https://openreview.net/forum?id=OQ08SN70M1V. 7

[2] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1811.11553. 3, 7

[3] 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. 1, 7

[4] 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. 3, 4, 7

[5] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 1999. https://link.springer.com/article/10.1023/A:1007515423169. 7

[6] 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. 3, 5, 21, 22

[7] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 2013. https://arxiv.org/abs/1708.06131. 3

[8] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018. https://arxiv.org/abs/1712.03141. 3

[9] 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. 1

[10] 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/. 3, 25, 26, 27

[11] Leo Breiman. Bagging predictors. *Machine learning*, 1996. https://link.springer.com/article/10.1007/BF00058655. 5, 7

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2005.14165. 7, 8

[13] 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. 3, 5, 21, 22

[14] 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. 3, 25, 26, 27

[15] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.02918. 17

[16] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning, 2020. https://arxiv.org/abs/2011.03395. 7

[17] 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. 3, 4, 27

[18] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. https://arxiv.org/abs/2006.06666. 7

[19] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout, 2017. https://arxiv.org/abs/1708.04552. 17

[20] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 2000. https://link.springer.com/chapter/10.1007/3-540-45014-9_1. 5, 7

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2010.11929. 3, 4, 5, 7, 8, 16, 30, 33

[22] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1712.02779. 17

[23] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi

Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1707.08945. 7

[24] 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 (NeurIPS)*, 2020. https://arxiv.org/abs/2010.15110. 44

[25] 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. 3, 5, 8, 15

[26] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997. https://www.sciencedirect.com/science/article/pii/S002200009791504X. 5, 7

[27] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*. Springer series in statistics New York, 2001. 7

[28] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1811.12231. 17

[29] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. https://arxiv.org/abs/1808.08750. 3, 7

[30] Raphael Gontijo-Lopes, Yann Dauphin, and Ekin D Cubuk. No one representation to rule them all: Overlapping features of training methods, 2021. https://arxiv.org/abs/2007.01434. 14

[31] Raphael Gontijo-Lopes, Yann Dauphin, and Ekin D. Cubuk. No one representation to rule them all: overlapping features of training methods, 2021. https://arxiv.org/abs/2110.12899. 16

[32] Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations (ICLR)*, 2014. https://arxiv.org/abs/1412.6544. 8

[33] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017. https://arxiv.org/abs/1706.04599. 14

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.03385. 4

[35] 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. 3, 4, 7

[36] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019. https://arxiv.org/abs/1903.12261. 3, 7

[37] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020. https://arxiv.org/abs/1912.02781. 17

[38] 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. 3, 4, 7

[39] John Hewitt, Xiang Lisa Li, Sang Michael Xie, Benjamin Newman, and Percy Liang. Ensembles and cocktails: Robust finetuning for natural language generation. In *NeurIPS 2021 Workshop on Distribution Shifts*, 2021. https://openreview.net/forum?id=qXucB21w1C3. 15, 16

[40] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NeurIPS) Deep Learning Workshop*, 2015. https://arxiv.org/abs/1503.02531. 24

[41] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 1998. https://ieeexplore.ieee.org/document/709601. 40

[42] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018. https://arxiv.org/abs/1803.05407. 4, 5, 8, 15

[43] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. https://arxiv.org/abs/1806.07572. 44

[44] 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. 1, 3, 4, 5, 7, 8, 16, 30, 31

[45] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models

through principled regularized optimization. In *Association for Computational Linguistics (ACL)*, 2019. https://arxiv.org/abs/1911.03437. 7

[46] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences (PNAS)*, 2017. https://arxiv.org/abs/1612.00796. 7

[47] 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. 3, 5, 7, 21, 22

[48] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision (ECCV)*, 2020. https://arxiv.org/abs/1912.11370. 7

[49] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1905.00414. 14, 41

[50] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1805.08974. 25, 26

[51] 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. 3, 25, 26, 27

[52] 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. 3, 5, 21, 25, 26, 27

[53] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning distorts pretrained features and underperforms out-of-distribution, 2021. https://openreview.net/forum?id=UYneFzXSJWh. 15

[54] Ananya Kumar, Aditi Raghunathan, Tengyu Ma, and Percy Liang. Calibrated ensembles: A simple way to mitigate ID-OOD accuracy tradeoffs. In *NeurIPS 2021 Workshop on Distribution Shifts*, 2021. https://openreview.net/forum?id=dmDE-9e9F_x. 15

[55] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 2003. https://doi.org/10.1023/A:1022859003006. 14

[56] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. https://arxiv.org/abs/1612.01474. 7, 8

[57] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2020. https://arxiv.org/abs/2002.11770. 7

[58] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2016. https://arxiv.org/abs/1608.03983. 23, 39

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. https://openreview.net/forum?id=Bkg6RiCqY7. 6, 23, 39

[60] 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. 3, 5, 21, 22

[61] Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert P. Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation, 2021. https://arxiv.org/abs/2102.02805. 7

[62] James Lucas, Juhan Bae, Michael R Zhang, Stanislav Fort, Richard Zemel, and Roger Grosse. Analyzing monotonic linear interpolation in neural network loss landscapes. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2104.11044. 8

[63] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2017. https://arxiv.org/abs/1706.06083. 7, 17

[64] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging, 2021. https://arxiv.org/abs/2111.09832. 16

[65] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 1989. https://www.sciencedirect.com/science/article/pii/S0079742108605368. 7

[66] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 2012. 40

[67] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning (ICML)*, 2020. https://arxiv.org/abs/2004.14444. 7

[68] 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. 1, 4, 7, 17

[69] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.02629. 23, 39

[70] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Deep ensembles for low-data transfer learning, 2020. https://arxiv.org/abs/2010.06866. 8

[71] 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. 3, 4, 5, 15

[72] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018. https://arxiv.org/abs/1803.02999. 8

[73] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.02530. 8

[74] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1912.01703. 23, 39

[75] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2021. https://arxiv.org/abs/2111.10050. 1, 3, 4, 5, 7, 8, 16, 30, 34, 35

[76] Boris Teodorovich Polyak. New method of stochastic approximation type. *Automation and remote control*, 1990. 2

[77] 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. 8

[78] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009. 7

[79] 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. 1, 3, 4, 5, 6, 7, 8, 24, 25, 30, 39

[80] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. https://openai.com/blog/better-language-models/. 3, 7

[81] 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. 3, 4, 5, 21, 22

[82] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process, 1988. https://ecommons.cornell.edu/handle/1813/8664. 2, 8

[83] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.04584. 17

[84] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision (ECCV)*, 2020. https://arxiv.org/abs/2008.01392. 7

[85] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1904.12843. 17

[86] 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. 3, 21, 22

[87] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020. http://proceedings.mlr.press/v119/shankar20c/shankar20c.pdf. 3, 5

[88] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014. https://arxiv.org/abs/1403.6382. 7

[89] David B Skalak et al. The sources of increased accuracy for two proposed boosting algorithms. In *American Association for Artificial Intelligence (AAAI), Integrating Multiple Learned Models Workshop*, 1996. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.2269&rep=rep1&type=pdf. 40

[90] Asa Cooper Stickland and Iain Murray. Diverse ensembles improve calibration. In *International Conference on Machine Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning*, 2020. https://arxiv.org/abs/2007.04206. 8

[91] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in

deep learning era. In *International Conference on Computer Vision (ICCV)*, 2017. https://arxiv.org/abs/1707.02968. 5, 7, 8, 33

[92] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://arxiv.org/abs/1912.04838. 7

[93] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.00567. 7, 8, 23

[94] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. https://proceedings.mlr.press/v97/tan19a/tan19a.pdf. 38

[95] 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. 1, 2, 3, 4, 7, 17, 18

[96] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. https://people.csail.mit.edu/torralba/publications/datasets_cvpr11.pdf. 7

[97] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2017. https://arxiv.org/abs/1705.07204. 7

[98] 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. 3, 4, 7

[99] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 23, 25, 26, 39

[100] Mitchell Wortsman, Maxwell C Horton, Carlos Guestrin, Ali Farhadi, and Mohammad Rastegari. Learning neural network subspaces. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2102.10472. 15

[101] 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. 3, 25, 26, 27

[102] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://arxiv.org/abs/1911.04252. 38

[103] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning (ICML)*, 2018. https://arxiv.org/abs/1802.01483. 7

[104] Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.10498. 7

[105] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification, 2019. https://arxiv.org/abs/1905.00546. 7

[106] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*, 2017. https://arxiv.org/abs/1703.04200. 7

[107] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2021. https://arxiv.org/abs/2111.07991. 7

[108] Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1907.08610. 8

[109] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text, 2020. https://arxiv.org/abs/2010.00747. 7

[110] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models, 2021. https://arxiv.org/abs/2109.01134. 7, 16, 24, 25

[111] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2020. https://arxiv.org/abs/1909.11764. 7