Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

John Miller [1] Rohan Taori [2] Aditi Raghunathan [2] Shiori Sagawa [2] Pang Wei Koh [2] Vaishaal Shankar [1]
Percy Liang [2] Yair Carmon [3] Ludwig Schmidt [4]

Abstract

For machine learning systems to be reliable, we must understand their performance in unseen, out-of-distribution environments. In this paper, we empirically show that out-of-distribution performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts. Specifically, we demonstrate strong correlations between in-distribution and out-of-distribution performance on variants of CIFAR-10 & ImageNet, a synthetic pose estimation task derived from YCB objects, FMoW-WILDS satellite imagery classification, and wildlife classification in iWildCam-WILDS. The correlation holds across model architectures, hyperparameters, training set size, and training duration, and is more precise than what is expected from existing domain adaptation theory. To complete the picture, we also investigate cases where the correlation is weaker, for instance some synthetic distribution shifts from CIFAR-10-C and the tissue classification dataset Camelyon17-WILDS. Finally, we provide a candidate theory based on a Gaussian data model that shows how changes in the data covariance arising from distribution shift can affect the observed correlations.

1. Introduction

Machine learning models often need to generalize from training data to new environments. A kitchen robot should work reliably in different homes, autonomous vehicles should drive reliably in different cities, and analysis software for satellite imagery should still perform well next year. The standard paradigm to measure generalization is to evaluate a model on a single test set drawn from the same distribution as the training set. But this paradigm provides only a narrow *in-distribution* performance guarantee: a small test error certifies future performance on new samples from exactly the same distribution as the training set. In many scenarios, it is hard or impossible to train a model on precisely the distribution it will be applied to. Hence a model will inevitably encounter *out-of-distribution* data on which its performance could vary widely compared to in-distribution performance. Understanding the performance of models beyond the training distribution therefore raises the following fundamental question: how does out-of-distribution performance relate to in-distribution performance?

Classical theory for generalization across different distributions provides a partial answer (Mansour et al., 2009; Ben-David et al., 2010). For a model f trained on a distribution D, known guarantees typically relate the in-distribution test accuracy on D to the out-of-distribution test accuracy on a new distribution D' via inequalities of the form

$$|\mathrm{acc}_D(f) - \mathrm{acc}_{D'}(f)| \;\leqslant\; d(D, D')$$

where d is a distance between the distributions D and D' such as the total variation distance. Qualitatively, these bounds suggest that out-of-distribution accuracy may vary widely as a function of in-distribution accuracy unless the distribution distance d is small and the accuracies are therefore close (see Figure 1 (top-left) for an illustration). More recently, empirical studies have shown that in some settings, models with similar in-distribution performance can indeed have different out-of-distribution performance (McCoy et al., 2019; Zhou et al., 2020; D'Amour et al., 2020).

In contrast to the aforementioned results, recent dataset reconstructions of the popular CIFAR-10, ImageNet, MNIST, and SQuAD benchmarks showed a much more regular pattern (Recht et al., 2019; Miller et al., 2020; Yadav & Bottou, 2019; Lu et al., 2020). The reconstructions closely followed the original dataset creation processes to assemble new test sets, but small differences were still enough to cause substantial changes in the resulting model accuracies. Nevertheless, the new out-of-distribution accuracies are almost perfectly

[1] Department of Computer Science, UC Berkeley, CA, USA [2] Department of Computer Science, Stanford University, Stanford, CA, USA [3] School of Computer Science, Tel Aviv University, Tel Aviv, Israel [4] Toyota Research Institute, Cambridge, MA, USA. Correspondence to: John Miller <miller_john@berkeley.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

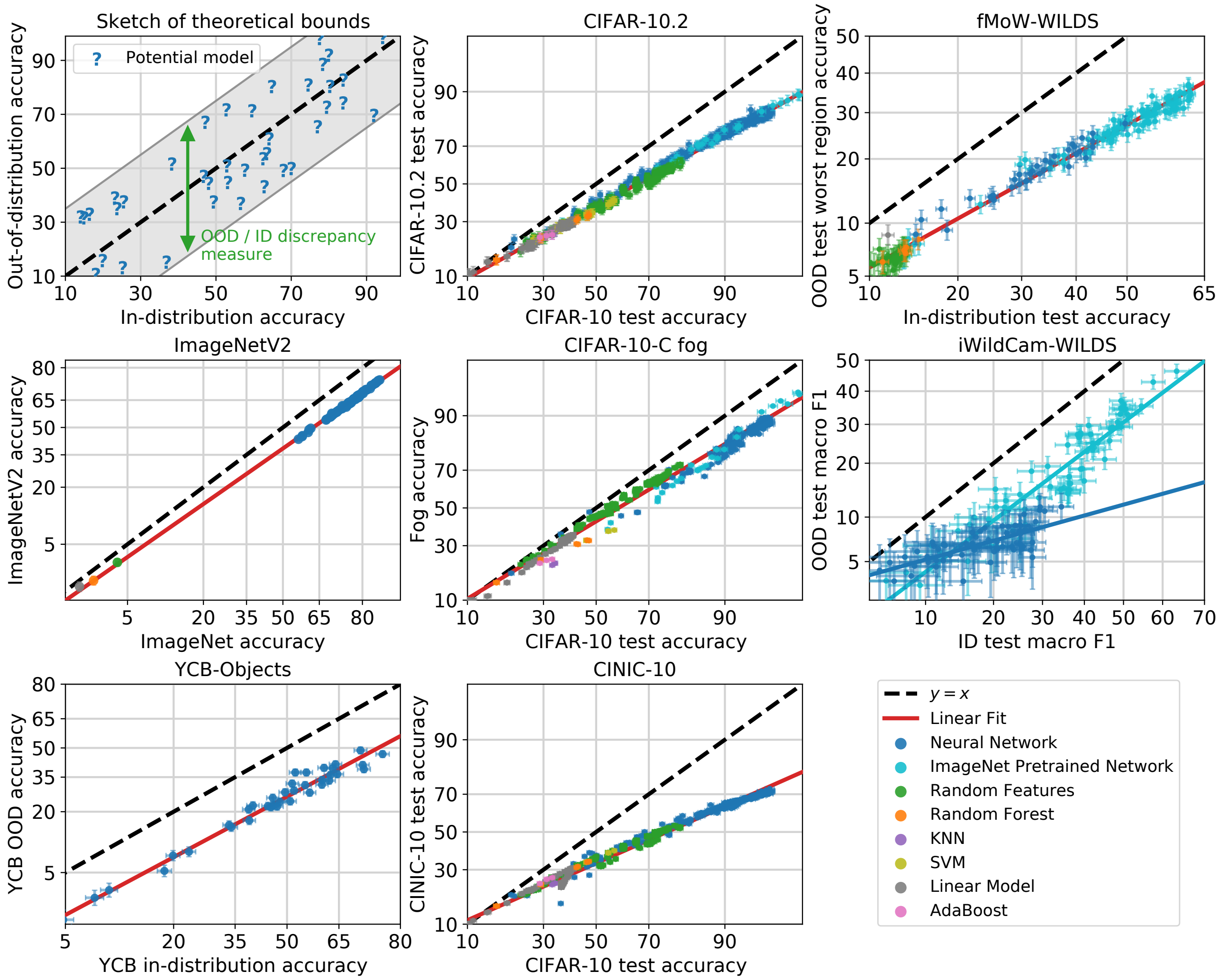

Figure 1. Out-of-distribution accuracies vs. in-distribution accuracies for a wide range of models, datasets, and distribution shifts. **Top left:** A sketch of the current bounds from domain adaptation theory. These bounds depend on distributional distances between in-distribution and out-of-distribution data, and they are loose in that they limit the deviation away from the y = x diagonal but do not prescribe a specific trend within these wide bounds (see Section 7). **Remaining panels:** In contrast, we show that for a wide range of models and datasets, there is a precise linear trend between out-of-distribution accuracy and in-distribution accuracy. Unlike what we might expect from theory, the linear trend does not follow the $y = x$ diagonal. The different panels represent different pairs of in-distribution and out-of-distribution datasets. Within each panel, we plot the performances of many different models, with different model architectures and hyperparameters. These datasets capture a variety of distribution shifts from dataset reproduction (CIFAR-10.2, ImageNet-V2); a real-world spatiotemporal distribution shift on satellite imagery (FMoW-WILDS); using a different benchmark test dataset (CINIC-10); synthetic perturbations (CIFAR-10-C and YCB-Objects); and a real-world geographic shift in wildlife monitoring (iWildCam-WILDS). Interestingly, for iWildCam-WILDS, models pretrained on ImageNet follow a different linear trend than models trained from scratch in-distribution, and we plot a separate trend line for ImageNet pretrained models in the iWildCam-WILDS panel. We explore this phenomenon more in Section 5.

linearly correlated with the original in-distribution accuracies for a range of deep neural networks. Importantly, this correlation holds *despite the substantial gap between in-distribution and out-of-distribution accuracies* (see Figure 1 (top-middle) for an example). However, it is currently unclear how widely these linear trends apply since they have been mainly observed for dataset reproductions and common variations of convolutional neural networks.

In this paper, we conduct a broad empirical investigation to characterize when precise linear trends such as in Figure 1 (top-middle) may be expected, and when out-of-distribution performance is less predictable as in Figure 1 (top-left). Concretely, we make the following contributions:

- We show that precise linear trends occur on several datasets and associated distribution shifts (see Figure 1). Going beyond the dataset reproductions in earlier work, we find linear trends on
  - popular image classification benchmarks (CIFAR-10 (Krizhevsky, 2009), CIFAR-10.1 (Recht et al., 2019), CIFAR-10.2 (Lu et al., 2020), CIFAR-10-C (Hendrycks & Dietterich, 2018), CINIC-10 (Darlow et al., 2018), STL-10 (Coates et al., 2011), ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019)),
  - a pose estimation testbed based on YCB-Objects (Calli et al., 2015),
  - and two distribution shifts derived from concrete applications of image classification: satellite imagery and wildlife photos via the FMoW-WILDS and iWildCam-WILDS variants from WILDS (Christie et al., 2018; Beery et al., 2020; Koh et al., 2020).
- We show that the linear trends hold for many models ranging from state-of-the-art methods such as convolutional neural networks, visual transformers, and self-supervised models, to classical methods like logistic regression, nearest neighbors, and kernel machines. Importantly, we find that classical methods follow the same linear trend as more recent deep learning architectures. Moreover, we demonstrate that varying model or training hyperparameters, training set size, and training duration all result in models that follow the same linear trend.
- We also identify three settings in which the linear trends do *not* occur or are less regular: some of the synthetic distribution shifts in CIFAR-10-C (e.g., Gaussian noise), the Camelyon17-WILDS shift of tissue slides from different hospitals, and a version of the aforementioned iWildCam-WILDS wildlife classification problem with a different in-distribution train-test split (Beery et al., 2020). We analyze these cases in detail via additional experiments to pinpoint possible causes of the linear trends.
- Pre-training a model on a larger and more diverse dataset offers a possibility to increase robustness. Hence we evaluate a range of models pre-trained on other datasets to study the impact of pre-training on the linear trends. Interestingly, even pre-trained models sometimes follow the same linear trends as models trained only on the in-distribution training set. Two examples are ImageNet pre-trained models evaluated on CIFAR-10 and FMoW-WILDS. In other cases (e.g., iWildCam-WILDS), pre-training yields clearly different relationships between in-distribution and out-of-distribution accuracies.
- As a starting point for theory development, we provide a candidate theory based on a simple Gaussian data model. Despite its simplicity, this data model correctly identifies the covariance structure of the distribution shift as one property affecting the performance correlation on the Gaussian noise corruption from CIFAR-10-C.

Overall, our results show a striking linear correlation between the in-distribution and out-of-distribution performance of many contemporary ML models on multiple distribution shift benchmarks. This raises the intriguing possibility that, despite their different creation mechanisms, a diverse range of distribution shifts may share common phenomena. In particular, improving in-distribution performance reliably improves out-of-distribution performance as well. However, it is unclear whether improving in-distribution performance is the only way, or even the best way, to improve out-of-distribution performance. More research is needed to understand the extent of the linear trends observed in this work and whether robustness interventions can improve over the baseline given by empirical risk minimization. We hope that our work serves as a step towards a better understanding of how distribution shifts affect model performance and how we can train models that perform robustly out-of-distribution.

2. Experimental setup

In each of our main experiments, we compare performance on two data distributions. The first is the training distribution D, which we refer to as “in-distribution” (ID). Unless noted otherwise, all models are trained only on samples from D (the main exception is pre-training on a different distribution). We compute ID performance via a held-out test set sampled from D. The second distribution is the “out-of-distribution” (OOD) distribution D' that we also evaluate the models on. For a loss function ℓ (e.g., error or accuracy), we denote the loss of model f on distribution D with $\ell_D(f) = \mathbb{E}_{x,y\sim D}\left[\ell(f(x), y)\right]$.

Experimental procedure. The goal of our paper is to understand the relationship between $\ell_D(f)$ and $\ell_{D'}(f)$ for a wide range of models f (convolutional neural networks, kernel machines, etc.) and pairs of distributions D, D' (e.g., CIFAR-10 and the CIFAR-10.2 reproduction). Hence for each pair D, D', our core experiment follows three steps:

1. Train a set of models $\{f_1, f_2, \ldots\}$ on samples drawn from D. Apart from the shared training distribution, the models are trained independently with different training set sizes, model architectures, random seeds, optimization algorithms, etc.
2. Evaluate the trained models f_i on two test sets drawn from D and D', respectively.
3. Display the models f_i in a scatter plot with each model’s two test accuracies on the two axes to inspect the resulting correlation.

An important aspect of our scatter plots is that we apply a non-linear transformation to each axis. Since we work with loss functions bounded in $[0, 1]$, we apply an axis scaling that maps $[0, 1]$ to $[-\infty, +\infty]$ via the probit transform. The probit transform is the inverse of the cumulative density function (CDF) of the standard Gaussian distribution, i.e., $l_{\text{transformed}} = \Phi^{-1}(l)$. Transformations like the probit or closely related logit transform are often used in statistics since a quantity bounded in $[0, 1]$ can only show linear trends for a bounded range. The linear trends we observe in our correlation plots are substantially more precise with the probit (or logit) axis scaling. Unless noted otherwise, each point in a scatter plot is a single model (not averaged over random seeds) and we show each point with 95% Clopper-Pearson confidence intervals for the accuracies.

We assembled a unified testbed that is shared across experiments and includes a multitude of models ranging from classical methods like nearest neighbors, kernel machines, and random forests to a variety of high-performance convolutional neural networks. Our experiments involved more than 3,000 trained models and 100,000 test set evaluations of these models and their training checkpoints. Due to the size of these experiments, we defer a detailed description of the testbed used to Appendix A.

3. The linear trend phenomenon

In this section, we show precise linear trends between in-distribution and out-of-distribution performance occur across a diverse set of models, data domains, and distribution shifts. Moreover, the linear trends holds not just across variations in models and model architectures, but also across variation in model or training hyperparameters, training dataset size, and training duration.

3.1. Distribution shifts with linear trends

We find linear trends for models in our testbed trained on five different datasets—CIFAR-10, ImageNet, FMoW-WILDS, iWildCam-WILDS, and YCB-Objects—and evaluated on distribution shifts that fall into four broad categories.

Dataset reproduction shifts. Dataset reproductions involve collecting a new test set by closely matching the creation process of the original. Distribution shift arises as a result of subtle differences in the dataset construction pipelines. Recent examples of dataset reproductions are the CIFAR-10.1 and ImageNet-V2 test sets from Recht et al. (2019), who observed linear trends for deep models on these shifts. In Figure 1, we extend this result and show both deep *and classical* models trained on CIFAR-10 and evaluated on CIFAR-10.2 (Lu et al., 2020) follow a linear trend. In Appendix B, we further show linear trends occur for deep and classical CIFAR-10 models evaluated on CIFAR-10.1 and for ImageNet models evaluated on ImageNet-V2.

Distribution shifts between machine learning benchmarks. We also consider distribution shifts between distinct benchmarks which are drawn from different data sources, but which use a compatible set of labels. For instance, both CIFAR-10 and CINIC-10 (Darlow et al., 2018) use the same set of labels, but CIFAR-10 is drawn from TinyImages (Torralba et al., 2008) and CINIC-10 is drawn from ImageNet (Deng et al., 2009) images. We show CIFAR-10 models exhibit linear trends when evaluated on CINIC-10 (Figure 1) or on STL-10 (Coates et al., 2011) (Appendix B).

Synthetic perturbations. Synthetic distribution shifts arise from applying a perturbation, such as adding Gaussian noise, to existing test examples. CIFAR-10-C (Hendrycks & Dietterich, 2018) applies 19 different synthetic perturbations to the CIFAR-10 test set. For many of these perturbations, we observe linear trends for CIFAR-10 trained models, e.g. the `Fog` shift in Figure 1. However, there are several exceptions, most notably adding isotropic Gaussian noise. We give further examples of linear trends on synthetic CIFAR-10-C shifts in Appendix B, and we more thoroughly discuss non-examples of linear trends in Section 4. In Figure 1, we also show that pose-estimation models trained on rendered images of YCB-Objects (Calli et al., 2015) follow a linear trend when evaluated on a images rendered with perturbed lighting and texture conditions.

Distribution shifts in the wild. We also find linear trends on two of the real-world distribution shifts from the WILDS benchmark (Koh et al., 2020): FMoW-WILDS and iWildCam-WILDS. FMoW-WILDS is a satellite image classification task derived from Christie et al. (2018) where in-distribution data is taken from regions (e.g., the Americas, Africa, Europe) across the Earth between 2002 and 2013, the out-of-distribution test-set is sampled from each region during 2016 to 2018, and models are evaluated by their accuracy on the worst-performing region. In Figure 1, we show models trained on FMoW-WILDS exhibit linear trends when evaluated out-of-distribution under both of these temporal and subpopulation distribution shifts.

iWildCam-WILDS is an image dataset of animal photos taken by camera traps deployed in multiple locations around the world (Koh et al., 2020; Beery et al., 2020). It is a multi-class classification task, where the goal is to identify the animal species (if any) within each photo. The held-out test set comprises photos taken by camera traps that are not seen in the training set, and the distribution shift arises because different camera traps vary markedly in terms of angle, lighting, and background. In Figure 1, we show models trained on iWildCam-WILDS also exhibit linear trends when evaluated OOD across different camera traps.

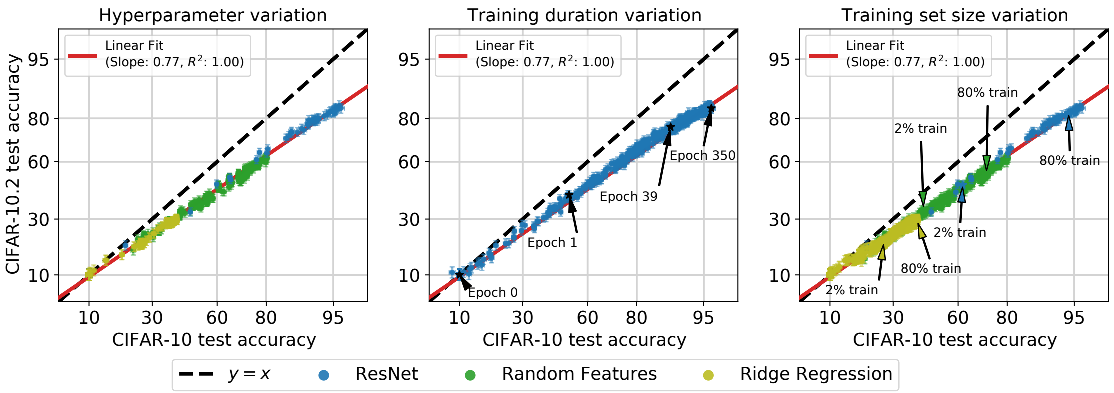

Figure 2. The linear trend between ID and OOD accuracy is invariant to changes in model hyperparameters, the number of training steps, and training set size. In each panel, we compare models with the linear fit from Figure 1. **Left:** For each model family, we vary model-size, regularization, and optimization hyperparameters. **Middle:** We evaluate each network after every epoch of training. **Right:** We train models on randomly sampled subsets of the training data, ranging from 1% to 80% of the CIFAR-10 training set size. In each setting, variation in hyperparameters, training duration, or training set size moves models along the trend line, but does not affect the linear fit.

3.2. Variations in model hyperparameters, training duration, and training dataset size

The linear trends we observe hold not just across different models, but also across variation in model and optimization hyperparameters, training dataset size, and training duration.

In Figure 2, we train and evaluate both classical and neural models on CIFAR-10 and CIFAR-10.2 while systematically varying (1) model hyperparameters, (2) training duration, and (3) training dataset size. When varying hyperparameters controlling the model size, regularization, and the optimization algorithm, the model families continue to follow the same trend line ($R^2 = 0.99$). We also find models lie on the same linear trend line *throughout training* ($R^2 = 0.99$). Finally, we observe models on trained on random subsets of CIFAR-10 lie on the same linear trend line as models trained on the full CIFAR-10 training set, despite their corresponding drop in in-distribution accuracy ($R^2 = 0.99$). In each case, hyperparameter tuning, early stopping, or changing the amount of i.i.d. training data moves models along the trend line, but does not alter the linear fit.

While we focus here on CIFAR-10 models evaluated on CIFAR-10.2, in Appendix B, we conduct an identical set of experiments for CINIC-10, CIFAR-10-C `Fog`, YCB-Objects, and FMoW-WILDS. We find the same invariance to hyperparameter, dataset size, and training duration shown in Figure 2 also holds for these diverse collection of datasets.

4. Distribution shifts with weaker correlations

We now investigate distribution shifts with a weaker correlation between ID and OOD performance than the examples presented in the previous section. We will discuss the Camelyon17-WILDS tissue classification dataset and specific image corruptions from CIFAR-10-C. Further discussion of a version of the iWildCam-WILDS wildlife classification dataset with a different in-distribution train-test split can be found in Appendix C.4.

4.1. Camelyon17-WILDS

Camelyon17-WILDS (Bandi et al., 2018; Koh et al., 2020) is an image dataset of metastasized breast cancer tissue samples collected from different hospitals. It is a binary image classification task where each example is a tissue patch. The corresponding label is whether the patch contains any tumor tissue. The held-out OOD test set contains tissue samples from a hospital not seen in the training set. The distribution shift largely arises from differences in staining and imaging protocols across hospitals.

In Figure 3, we plot the results of training different ImageNet models and random features models from scratch across a variety of random seeds. There is significant variation in OOD performance. For example, the models with 95% ID accuracy have OOD accuracies that range from about 50% (random chance) to 95%. This high degree of variability holds even after averaging each model over ten independent training runs (see Appendix C.1).

Appendix C.1 also contains additional analyses exploring the potential sources of OOD performance variation, including ImageNet pretraining, data augmentation, and similarity between test examples. Specifically, we observe that ImageNet pretraining does not increase the ID-OOD correlation, while strong data augmentation significantly reduces, but does not eliminate, the OOD variation. Another potential reason for the variation is the similarity between images from the same slide / hospital, as similar examples have been shown to result in analogous phenomena in natural

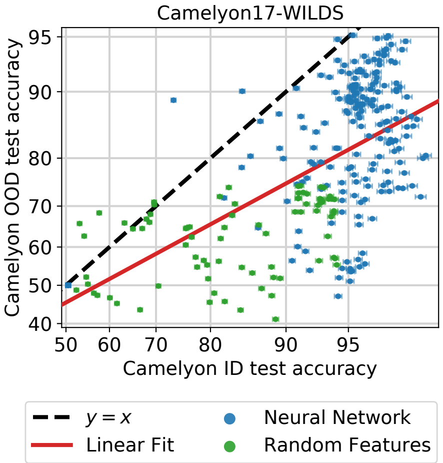

Figure 3. A range of neural network and random feature models trained on Camelyon17-WILDS and evaluated on the ID and OOD test sets. OOD accuracy is highly variable across the spectrum of ID accuracies, and there is no precise linear trend.

language processing (Zhou et al., 2020). We explore this hypothesis in a synthetic CIFAR-10 setting, where we simulate increasing the similarity between examples by taking a small seed set of examples and then using data augmentations to create multiple similar versions. We find that in this CIFAR-10 setting, shrinking the effective test set size in this way increases OOD variation to a substantially greater extent than shrinking the effective training set size.

4.2. CIFAR-10-Corrupted

CIFAR-10-C (Hendrycks & Dietterich, 2018) corrupts CIFAR-10 test images with various image perturbations. The choice of corruption can have a significant impact on the correlation between ID and OOD accuracy. Interestingly, the mathematically simple corruption with Gaussian noise is one of the corruptions with worst ID-OOD correlation. Appendix C.2 details experiments for each corruption.

In Appendix C.3, we also investigate how the relationship between the ID and OOD data covariances impacts the linear trend. We find the linear fit is substantially better when the ID and OOD covariances match up to a scaling factor, which is consistent with the theoretical model we propose and discuss in Section 6.

5. The effect of pretrained models

In this section we expand our scope to methods that leverage models pretrained on a third auxiliary distribution different from the ones we refer to in-distribution (ID) and out-of-distribution (OOD). Fine-tuning pretrained models on the task-specific (ID) training set is a central technique in modern machine learning (Donahue et al., 2014; Razavian et al., 2014; Kornblith et al., 2019; Peters et al., 2018; Devlin et al., 2018), and zero-shot prediction (using the pretrained model directly without any task-specific training) is showing increasing promise as well (Brown et al., 2020; Radford et al., 2021). Therefore, it is important to understand how the use of pretrained models affects the robustness of models to OOD data, and whether fine-tuning and zero-shot inference differ in that respect.

The dependence of the pretrained model on auxiliary data makes the ID/OOD distinction more subtle. Previously, “ID” simply referred to the distribution of the training set, while OOD referred to an alternative distribution not seen in training. In this section, the training set includes the auxiliary data as well, but we still refer to the *task-specific* training set distributions as ID. This means, for example, that when fine-tuning an ImageNet model on the CIFAR-10 training set, we still refer to accuracy on the CIFAR-10 test set as ID accuracy. In other words, the “ID” distributions we refer to in this section are precisely the “ID” distributions of the previous sections (displayed on the x-axes in our scatter plots), but the presence of auxiliary training data alters the meaning of the term.

With the effect of auxiliary data on the meaning of “ID” in mind, it is reasonable to expect that ID/OOD linear trends observed when training purely on ID data will change or break down when pretrained models are used. In this section we test this hypothesis empirically and reveal a more nuanced reality: the task and the use of the pretrained model matter, and sometimes models pre-trained on seemingly broader distributions still follow the same linear trend as the models trained purely on in-distribution data. We first present our findings for fine-tuning pretrained ImageNet models and subsequently discuss results for zero-shot prediction. See Appendix D for more experimental details.

Fine-tuning pretrained models on ID data. Figure 4 plots OOD performance vs. ID performance for models trained from-scratch (purely on ID data) and fine-tuned models whose initialization was pretrained on ImageNet. Across the board, pretrained models attain better performance on both the ID and OOD test sets. However, fine-tuning affects ID-OOD correlations differently across tasks. In particular, for CIFAR-10 reproductions and for FMoW-WILDS, fine-tuning produces results that lie on the same ID-OOD trend as purely ID-trained models (Figure 4 left and center). On the other hand, a similar fine-tuning procedure yields models with a different ID-OOD relationship on iWildCam-WILDS than models trained from scratch on this dataset. Moreover, the weight decay used for fine-tuning seems to also affect the linear trend (Figure 4 right).

One conjecture is that the qualitatively different behavior of fine-tuning on iWildCam-WILDS is related to the fact that ImageNet is a more diverse dataset that may encode

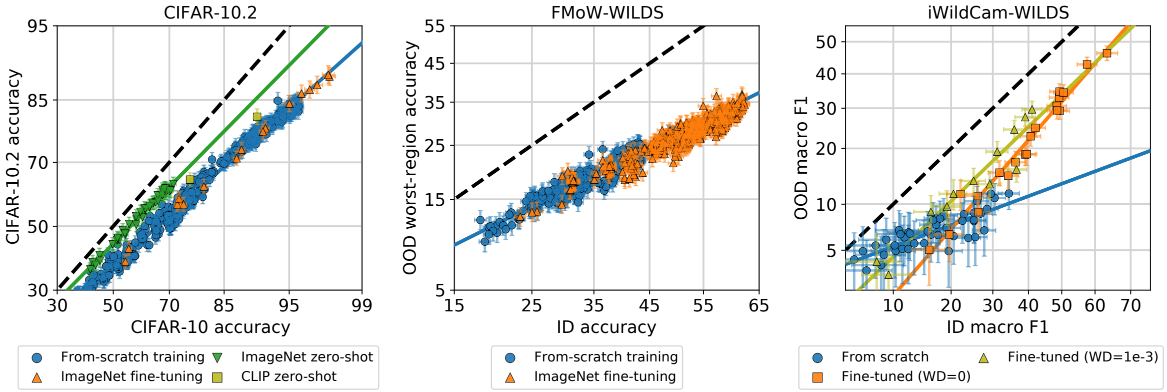

Figure 4. The effect of pre-training with additional data on CIFAR-10.2 (left), FMoW-WILDS (middle), and iWildCam-WILDS (right). On CIFAR-10.2 and FMoW-WILDS, fine-tuning pretrained models moves the models along the predicted ID-OOD line. However, on CIFAR-10.2, zero-shot prediction using pretrained models deviates from this line. On iWildCam-WILDS, fine-tuning pretrained models changes the ID-OOD relationship observed for models trained from scratch. Moreover, the weight decay hyperparameter affects the ID-OOD relationship in fine-tuned models.

robustness-inducing invariances that are not represented in the iWildCam-WILDS ID training set. For instance, both ImageNet and iWildCam-WILDS contain high-resolution images of natural scenes, but the camera perspectives in iWildCam-WILDS may be more limited compared to ImageNet. Hence ImageNet classifiers may be more invariant to viewpoint, which may aid generalization to previously unseen camera viewpoints in the OOD test set of iWildCam-WILDS. On the other hand, the satellite images in FMoW-WILDS are all taken from an overhead viewpoint, so learning invariance to camera viewpoints from ImageNet might not be as beneficial. Investigating this and related conjectures (e.g., invariances such as lighting, object pose, and background) is an interesting direction for future work.

Zero-shot prediction on pretrained models. A common explanation for OOD performance drop is that training on the ID training set biases the model toward patterns that are more predictive on the ID test set than on its OOD counterpart. With that explanation in mind, the fact that fine-tuned models maintain the same ID/OOD linear trend as from-scratch models is surprising: once could reasonably expect that an initialization determined independently of either ID or OOD data would produce models that are less biased toward the former. Indeed, in the extreme scenario that no fine-tuning takes place, the model should have no bias toward either distribution, and we therefore expect to see a different ID/OOD trend.

The CIFAR-10 allows us directly test this expectation directly by performing zero-shot inference on models pretrained on ImageNet: since the CIFAR-10 classes form a subset of the ImageNet classes, we simply feed (resized) CIFAR-10 images to these models, and limit the prediction to the relevant class subset. The resulting classifiers have no preference for either the ID or OOD test set because they depend on neither distribution. We plot the zero-shot prediction results in Figure 4 (left) and observe that, as expected, they deviate from the basic linear trend. Moreover, they form a different linear trend closer—but not identical—to $x = y$. The fact that the zero-shot linear trend is closer to $x = y$ supports the hypothesis that the performance drop partially stems from bias in ID training. However, the fact that this trend is still below $x = y$ suggests that the drop is also partially due to CIFAR-10 reproductions being harder than CIFAR-10 for current methods (interestingly, humans show similar performance on both test sets (Recht et al., 2019; Miller et al., 2020; Shankar et al., 2020)). These finding agree with prior work (Lu et al., 2020).

As another test of zero-shot inference, we apply two publically-available CLIP models on CIFAR-10 by creating last-layer weights out of natural language descriptions of the classes (Radford et al., 2021). As Figure 4 (left) shows, these models are slightly above the basic ID/OOD linear trend, but below the trend of zero-shot inference with ImageNet models.

Additional experiments. In Appendix D we describe additional experiments with pretrained models. To explore a middle ground between zero-shot prediction and full-model fine-tuning, we consider a linear probe on CLIP for both CIFAR-10 and FMoW-WILDS. For CIFAR-10, we also consider models trained on a task-relevant subset of ImageNet classes (Darlow et al., 2018) and models trained in a semi-supervised fashion using unlabeled data from 80 Million Tiny Images (Torralba et al., 2008; Carmon et al., 2019; Augustin & Hein, 2020). Generally, we find that, compared to zero-shot prediction, these techniques deviate less from the basic linear trend. We also report results on additional OOD settings, namely CIFAR-10.1 and different region subsets for FMoW-WILDS, and reach similar conclusions.

6. Theoretical models for linear fits

In this section we propose and analyze a simple theoretical model that distills several of the empirical phenomena from the previous sections. Our goal here is *not* to obtain a general model that encompasses complicated real distributions such as the images in CIFAR-10. Instead, our focus is on finding a simple model that is still rich enough to exhibit some of the same phenomena as real data distributions.

6.1. A simple Gaussian distribution shift setting

We consider a simple binary classification problem where the label y is distributed uniformly on $\{-1, 1\}$ both in the original distribution D and shifted distribution D'. Conditional on y, we consider D such that $\boldsymbol{x} \in \mathbb{R}^d$ is an isotropic Gaussian, i.e.,

$$\boldsymbol{x} \,|\, y \;\sim\; \mathcal{N}(\boldsymbol{\mu} \cdot y;\, \sigma^2 I_{d\times d}),$$

for mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and variance $\sigma^2 > 0$.

We model the distribution shift as a change in σ and $\boldsymbol{\mu}$. Specifically, we assume that the shifted distribution D' corresponds to shifted parameters

$$\boldsymbol{\mu}' = \alpha \cdot \boldsymbol{\mu} + \beta \cdot \boldsymbol{\Delta} \quad \text{and} \quad \sigma' = \gamma \cdot \sigma \qquad (1)$$

where $\alpha, \beta, \gamma > 0$ are fixed scalars and $\boldsymbol{\Delta}$ is *uniformly distributed* on the sphere in $\mathbb{R}^d$. Note that in our setting D' is a random object determined by the draw of $\boldsymbol{\Delta}$.

Within the setup describe above, we focus on linear classifiers of the form $\boldsymbol{x} \mapsto \operatorname{sign}(\boldsymbol{\theta}^\top \boldsymbol{x})$. The following theorem states that, as long as $\boldsymbol{\theta}$ depends only on the training data and is *thereby independent of the random shift direction* $\boldsymbol{\Delta}$, the probit-transformed accuracies on D and D' have a near-linear relationship with slope α/γ. (Recall that the probit transfrom is the inverse of the standard Normal cdf $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t$). The deviation from linearity is of order $d^{-1/2}$ and vanishes in high dimension.

Theorem 1. *In the setting described above where $\boldsymbol{\Delta}$ is independent of $\boldsymbol{\theta}$, let $\delta \in (0, 1)$. With probability at least $1 - \delta$, we have*

$$\left| \Phi^{-1}(\mathrm{acc}_{D'}(\boldsymbol{\theta})) - \frac{\alpha}{\gamma}\, \Phi^{-1}(\mathrm{acc}_{D}(\boldsymbol{\theta})) \right| \leqslant \frac{\beta}{\gamma\sigma} \sqrt{\frac{2\log {}^{2}/_{\delta}}{d}} \,.$$

The theorem is a direct consequence of the concentration of measure; see proof in Appendix E.1.

We illustrate Theorem 1 by simulating its setup and training different linear classifiers by varying the loss function and regularization. Figure 5 shows good agreement between the performance of linear classifiers and the theoretically-predicted linear trend. Furthermore, conventional nonlinear classifiers (nearest neighbors and random forests) also satisfy the same linear relationship, which does not directly follow from our theory. Nevertheless, if the decision boundary of the nonlinear becomes nearly linear in our setting a similar theoretical analysis might be applicable. Our simple Gaussian setup thus illustrates how linear trends can arise across a wide range of models.

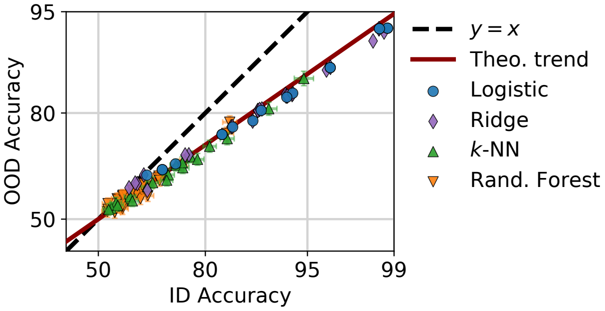

Figure 5. Illustration of the theoretical distribution shift model in Section 6.1 with $d = 10^5$, $\alpha = 0.7$, $\beta = 0.5$ and $\gamma = 1$ (see Appendix E.3 for details). The accuracies for linear models (logistic and ridge regression) agree with the prediction of Theorem 1. Moreover, nonlinear models (nearest neighbors and random features) exhibit the same probit trend we prove for linear classifiers.

6.2. Modeling departures from the linear trend

In the previous section, we identified a simple Gaussian setting that showed linear fits across a large range of models. Now we discuss small changes to the setting that break linear trends and draw parallels to the empirical observations on complex datasets presented in this paper. In Appendix E.2, we discuss each of these modifications in further detail.

Adversarial distribution shifts. Previously, the direction $\boldsymbol{\Delta}$ which determines the distribution shift as defined above in eq. (1), was chosen independent of the tested models $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$. However, when $\boldsymbol{\Delta}$ is instead chosen by an adversary with knowledge of the tested models, the ID-OOD relationship can be highly non-linear. This is reminiscent of adversarial robustness notions where models with comparable in-distribution accuracies can have widely differing adversarial accuracies depending on the training method.

Pretraining data. Additional training data from a *different* distribution available for pretraining could contain information about the shift $\boldsymbol{\Delta}$. In this case, the pretrained models are not necessarily independent of $\boldsymbol{\Delta}$ and these models could lie above the linear fit of classifiers without pretraining. See Section 5 for a discussion of when such behavior arises in practice.

Shift in covariance. Previously, we assumed that $\boldsymbol{x} \mid y$ is always an isotropic Gaussian. Instead consider a setting where the original distribution is of the form $\boldsymbol{x}|y \sim \mathcal{N}(\boldsymbol{\mu}y; \Sigma)$ where Σ is not scalar (i.e., has distinct eigenvalues). Then, the linear trend breaks down even when the distribution shift is simple additive white Gaussian noise corresponding to $\boldsymbol{x}|y \sim \mathcal{N}(\boldsymbol{\mu}y; \Sigma + (\sigma')^2 I_{d\times d})$. For example, ridge regularization turns out to be an effective robustness intervention in this setting. However, if the shifted distribution is of the form $\boldsymbol{x}|y \sim \mathcal{N}(\boldsymbol{\mu}y; \gamma\Sigma)$ for some scalar $\gamma > 0$, it is straightforward to see that a linear trend holds.

These theoretical observations suggest that a covariance change in ID/OOD the distribution shift could be a possible explanation for some departures from the linear trends such as additive Gaussian noise corruptions in CIFAR-10-C. To test this hypothesis, we created a new distribution shift by corrupting CIFAR-10 with noise sampled from the same covariance as the original CIFAR-10 distribution. As discussed in Section 4.2, we find that the correlation between ID and OOD accuracy is substantially higher with the covariance-matched noise than with isotropic Gaussian noise with similar magnitude.

While the theoretical setting we study in this work is much simpler than real-world distributions, the analysis sheds some light on when to expect linear trends and what leads to departures. Ideally, a theory would precisely explain what differentiates CIFAR-10.2, CINIC-10, and the CIFAR-10-C-Fog shift (see Figure 1) where we see linear trends from simply adding Gaussian noise to the images as in CIFAR-10-C-Gaussian where we do not observe linear trends. A possible direction may be to characterize shifts by their generation process, and we leave this to future work.

7. Related work

Due to the large body of research on distribution shifts, domain adaptation, and reliable machine learning, we only summarize the most directly related work here. Appendix F contains a more detailed discussion of related work.

Domain generalization theory. Prior work has theoretically characterized the performance of classifiers under distribution shift. Ben-David et al. (2006) provided the first VC-dimension-based generalization bound. They bound the difference between a classifier's error on the source distribution (D) and target distribution (D') via a classifier-induced divergence measure. Mansour et al. (2009) extended this work to more general loss functions and provided sharper generalization bounds via Rademacher complexity. These results have been generalized to include multiple sources (Blitzer et al., 2007; Hoffman et al., 2018; Mansour et al., 2008). The philosophy underlying these works is that robust models should aim to minimize the induced divergence measure and thus guarantee similar OOD and ID performance.

The linear trends we observe in this paper are not captured by such analyses. As illustrated in Figure 1 (left), the bounds described above can only state that OOD performance is highly predictable from ID performance if they are equal (i.e., when the gray region is tight around the $x = y$ line). In contrast, we observe that OOD performance is *both* highly predictable from ID performance and significantly different from it. Our Gaussian model in Section 6.1 demonstrates how the linear trend phenomenon can come about in a simple setting. However, unlike the above-mentioned domain generalization bounds, it is limited to particular distributions and the hypothesis class of linear classifiers.

Mania & Sra (2020) proposed a condition that implies an approximately linear trend between ID and OOD accuracy, and empirically checked their condition in dataset reproduction settings. The condition is related to model similarity, and requires the probability of certain multiple-model error events to remain invariant under distribution shift. It is unclear whether their condition can predict a priori whether a distribution shift will show a linear trend, and the predicted linearity does not improve under probit accuracy scaling.

Empirical observations of linear trends. Precise linear trends between in-distribution and out-of-distribution generalization were first discovered in the context of dataset reproduction experiments. Recht et al. (2018; 2019); Yadav & Bottou (2019); Miller et al. (2020) constructed new test sets for CIFAR-10 (Krizhevsky, 2009), ImageNet (Deng et al., 2009; Russakovsky et al., 2015), MNIST (LeCun et al., 1998), and SQuAD (Rajpurkar et al., 2016) and found linear trends similar to those in Figure 1.

However, these studies were limited in their scope, as they just focused on dataset reproductions. While Taori et al. (2020) later showed that linear trends still occur for ImageNet models on datasets like ObjectNet, Vid-Robust, and YTBB-Robust (Barbu et al., 2019; Shankar et al., 2019), all of their experiments were limited to ImageNet-like tasks. We significantly broaden the scope of the linear trend phenomenon by including a range of additional distribution shifts such as CINIC-10, STL-10, FMoW-WILDS, and iWildCam-WILDS, as well as identifying negative examples like Camelyon17-WILDS and some CIFAR-10-C shifts. In addition, we also include a pose estimation task with YCB-Objects. The results show that linear trends not only occur in academic benchmarks but also in distribution shifts coming from applications "in the wild." We also show that linear trends hold across different learning approaches, training durations, and hyperparameters.

Kornblith et al. (2019) study linear fits in the context of transfer learning and train or fine-tune models on the distribution corresponding to the y-axis in our setting. On a variety of image classification tasks, they show a model's ImageNet test accuracy linearly correlates with the model's accuracy on the new task after fine-tuning. The similar between their results and those in this work suggest that they may both be part of a broader phenomenon of predictable generalization in machine learning.

Acknowledgements

We would like to thank Sara Beery, Gabriel Ilharco, Moritz Hardt, Daniel Levy, Mike Li, Henrik Marklund, Horia Mania, Yishay Mansour, Hongseok Namkoong, Ben Recht, Max Simchowitz, Kunal Talwar, Nilesh Tripuraneni, Mitchell Wortsman, Michael Xie, and Bin Yu for helpful discussions while working on this paper. We also thank Benjamin Burchfiel, Eric Cousineau, Kunimatsu Hashimoto, Russ Tendrake, and Vickie Ye for assistance and guidance in setting up the YCB-Objects testbed.

This work was funded by an Open Philanthropy Project Award. JM was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814. RT was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1656518. AR was supported by the Google PhD Fellowship and the Open Philanthropy AI Fellowship. SS was supported by the Herbert Kunzel Stanford Graduate Fellowship. YC was supported by Len Blavatnik and the Blavatnik Family foundation, and the Yandex Machine Learning Initiative for Machine Learning.

References

Augustin, M. and Hein, M. Out-distribution aware self-training in an open world setting, 2020. URL https://arxiv.org/abs/2012.12372.

Ball, K. An elementary introduction to modern convex geometry. *Flavors of geometry*, 1997. http://library.msri.org/books/Book31/files/ball.pdf.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 2018. https://ieeexplore.ieee.org/document/8447230.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://objectnet.dev/.

Beery, S., Cole, E., and Gjoka, A. The iWildCam 2020 competition dataset. In *Fine-Grained Visual Categorization Workshop at the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/2004.10340.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems (NeurIPS)*, 2006. https://papers.nips.cc/paper/2006/hash/b1b0432ceafb0ce714426e9114852ac7-Abstract.html.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 2010. https://link.springer.com/article/10.1007/s10994-009-5152-4.

Biggio, B. and Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018. https://arxiv.org/abs/1712.03141.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD)*, 2013. https://arxiv.org/abs/1708.06131.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2021. URL http://www.blender.org.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. https://papers.nips.cc/paper/2007/hash/42e77b63637ab381e8be5f8318cc28a2-Abstract.html.

Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. https://opencv.org/.

Breiman, L. Random forests. *Machine learning*, 2001. https://link.springer.com/article/10.1023/A:1010933404324.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2005.14165.

Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., and Dollar, A. M. Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols, 2015. URL https://arxiv.org/abs/1502.03143.

Carmon, Y., Raghunathan, A., Schmidt, L., and Duchi, J. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.13736.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2006.09882.

Chaurasia, A. and Culurciello, E. LinkNet: exploiting encoder representations for efficient semantic segmentation. In *Visual Communications and Image Processing (VCIP)*, 2017. https://arxiv.org/abs/1707.03718.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2006.10029.

Chen, X. and He, K. Exploring simple siamese representation learning, 2020. https://arxiv.org/abs/2011.10566.

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., and Feng, J. Dual path networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. https://arxiv.org/abs/1707.01629.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1610.02357.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1711.07846.

Coates, A. and Ng, A. Y. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*. Springer, 2012. https://www-cs.stanford.edu/~acoates/papers/coatesng_nntot2012.pdf.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. http://proceedings.mlr.press/v15/coates11a.html.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. RandAugment: practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://papers.nips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning, 2020. URL https://arxiv.org/abs/2011.03395.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. CINIC-10 is not ImageNet or CIFAR-10, 2018. URL https://arxiv.org/abs/1810.03505.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. http://arxiv.org/abs/1810.04805.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., Gelly, S., Houlsby, N., Zhai, X., and Lucic, M. On robustness and transferability of convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. https://arxiv.org/abs/2007.08558.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, 2014. https://arxiv.org/abs/1310.1531.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 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.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning (ICML)*, 2013. http://proceedings.mlr.press/v28/germain13.html.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A new PAC-Bayesian perspective on domain adaptation. In *International conference on machine learning (ICML)*, 2016. https://arxiv.org/abs/1506.04573.

Gowal, S., Qin, C., Uesato, J., Mann, T., and Kohli, P. Uncovering the limits of adversarial training against norm-bounded adversarial examples. https://arxiv.org/abs/2010.03593, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2007.01434.

Hashimoto, K., Ta, D.-N., Cousineau, E., and Tedrake, R. Kosnet: A unified keypoint, orientation and scale network for probabilistic 6d pose estimation. http://groups.csail.mit.edu/robotics-center/public_papers/Hashimoto20.pdf, 2020.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. Multi-class AdaBoost. *Statistics and its Interface*, 2009. http://ww.web.stanford.edu/~hastie/Papers/SII-2-3-A8-Zhu.pdf.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. https://arxiv.org/abs/1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016b. https://arxiv.org/abs/1603.05027.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1903.12261.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2020. https://arxiv.org/abs/2006.16241.

Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., and Navab, N. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision*, 2012. https://link.springer.com/chapter/10.1007/978-3-642-37331-2_42.

Hoffman, J., Mohri, M., and Zhang, N. Algorithms and theory for multiple-source adaptation. In *International Conference on Neural Information Processing Systems (ICML)*, 2018. https://arxiv.org/abs/1805.08727.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Conference on Computer Vsion and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1709.01507.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1608.06993.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size, 2016. https://arxiv.org/abs/1602.07360.

Kendall, A., Grimes, M., and Cipolla, R. PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In *International Conference on Computer Vision (ICCV)*, 2015. https://arxiv.org/abs/1505.07427.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2012.07421.

Kornblith, S., Shlens, J., and Le, Q. V. Do better ImageNet models transfer better? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1805.08974.

Krizhevsky, A. Learning multiple layers of features from tiny images, 2009. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. http://yann.lecun.com/exdb/mnist/, 1998.

Lepetit, V. and Fua, P. Monocular model-based 3D tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 2005. https://ieeexplore.ieee.org/document/8187270.

Lepetit, V., Moreno-Noguer, F., and Fua, P. EPnP: An accurate o(n) solution to the PnP problem. *International Journal of Computer Vision*, 2009. https://link.s pringer.com/article/10.1007/s11263-008-0152-6.

Li, X. and Bilmes, J. A bayesian divergence prior for classiffier adaptation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007. http://proceedings.mlr.press/v2/li07a.html.

Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., Yuille, A., Huang, J., and Murphy, K. Progressive neural architecture search. In *European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1712.00559.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. https://arxiv.org/abs/1411.4038.

Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., and Schmidt, L. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In *European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1807.11164.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1706.06083.

Mania, H. and Sra, S. Why do classifier accuracies show linear trends under distribution shift?, 2020. https://arxiv.org/abs/2012.15483.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation with multiple sources. *Advances in neural information processing systems (NeurIPS)*, 2008. https://papers.nips.cc/paper/2008/hash/0e65972dce68dad4d52d063967f0a705-Abstract.html.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory (COLT)*, 2009. https://arxiv.org/abs/0902.3430.

McCoy, R. T., Min, J., and Linzen, T. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2019. https://arxiv.org/abs/1911.02969.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning (ICML)*, 2020. https://arxiv.org/abs/2004.14444.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010. https://ieeexplore.ieee.org/document/5288526.

Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. G., and Daniilidis, K. 6-dof object pose from semantic keypoints. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2011–2018. IEEE, 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011. https://www.jmlr.org/papers/v12/pedregosa11a.html.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. https://arxiv.org/abs/1802.05365.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Rad, M. and Lepetit, V. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3828–3836, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2103.00020.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. Designing network design spaces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://arxiv.org/abs/2003.13678.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. https://arxiv.org/abs/1606.05250.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014. https://arxiv.org/abs/1403.6382.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do CIFAR-10 classifiers generalize to CIFAR-10? https://arxiv.org/abs/1806.00451, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.10811.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. https://arxiv.org/abs/1409.0575.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1801.04381.

Santurkar, S., Tsipras, D., and Madry, A. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2008.04859.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. https://arxiv.org/abs/1804.11285.

Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. Do image classifiers generalize across time?, 2019. https://arxiv.org/abs/1906.02168.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020. http://proceedings.mlr.press/v119/shankar20c.html.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. 2015. https://arxiv.org/abs/1409.1556.

Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. https://arxiv.org/abs/1312.6199.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Conference on Computer vision and Pattern Recognition (CVPR)*, 2015. https://arxiv.org/abs/1409.4842.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.00567.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. https://arxiv.org/abs/1602.07261.

Tan, M. and Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1905.11946.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2007.00644.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and van der Laak, J. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million Tiny Images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. https://people.csail.mit.edu/torralba/publications/80millionImages.pdf.

Torralba, A., Efros, A. A., et al. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. https://ieeexplore.ieee.org/document/5995347.

Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., and Birchfield, S. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning (CoRL)*, 2018. https://arxiv.org/abs/1809.10790.

Wang, M. and Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018. https://arxiv.org/abs/1802.03601.

Xiang, Y., Schmidt, T., Narayanan, V., and Fox, D. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. 2017. https://arxiv.org/abs/1711.00199.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1611.05431.

Yadav, C. and Bottou, L. Cold case: The lost MNIST digits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. http://arxiv.org/abs/1905.10498.

Zhang, X., Li, Z., Loy, C. C., and Lin, D. Polynet: A pursuit of structural diversity in very deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1611.05725.

Zhang, X., Zhou, X., Lin, M., and Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1707.01083.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1612.01105.

Zhou, X., Nie, Y., Tan, H., and Bansal, M. The curse of performance instability in analysis datasets: Consequences, source, and suggestions, 2020. URL https://arxiv.org/abs/2004.13606.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Springer, 2018.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1707.07012.