On the Connection between Pre-training Data Diversity and Robustness

Anonymous Authors[1]

Abstract

Our work studies the implications of transfer learning on model behavior beyond accuracy: *how does the pre-training distribution affect the downstream robustness of a fine-tuned model?* We analyze model robustness using the framework proposed by Taori et al. (2020), which demonstrates that in-distribution and out-of-distribution performances are highly correlated along a *robustness linear trend*. We explore various interventions that significantly alter the pre-training distribution, including label space, label semantics, and the pre-training dataset itself. In most cases, changes during pre-training have minimal impact on the original linear trend produced by models pre-trained on the full ImageNet dataset. We demonstrate these findings on pre-training distributions constructed from ImageNet and iNaturalist, and fine-tuning data from the iWildCams-WILDS benchmark.

1. Introduction

There is often a mismatch between data distribution in the real world and the training set that machine learning models are trained on. When this happens, these models can perform in unexpected and undesirable ways (Rosenfeld et al., 2018; Koh et al., 2021). As such, robustness under distribution shift is a fundamental concern for producing reliable ML systems. For example, a self-driving car should be able to generalize to a wide variety of weather scenarios to be considered safe, each of which could represent a distribution shift from what it has seen during training.

However, there are many different ways for test data to be "out-of-distribution". This raises a fundamental question of what a meaningful way to measure robustness would be. In our work, we study natural distribution shifts, named so because these shifts reflect real-world processes. In particular, we focus on evaluating model robustness on the iWildCam dataset (Koh et al., 2021) which uses geo-location of camera traps to form distinct test sets. Following Miller et al. (2021), we measure robustness as the difference in performance on the in-distribution (ID) test set and the out-of-distribution (OOD) test set. This allows us to take advantage of results from Miller et al. (2021), which demonstrate that ID and OOD performances are strongly correlated. We can thus use differences in *robustness trend* to isolate important factors for improving model invariance to distribution shift. This trend has been shown to be consistent across many architectures, sample sizes and optimization procedures, which means we can make conclusions about the importance of a *pre-training distribution* instead of a particular trained model.

In most cases, Miller et al. (2021) observes that pre-training does not affect the finetuning robustness trend. One notable exception, which we study in this work, is the iWildCams-WILDS benchmark, where there is a significant deviation in the linear trend between pre-training on ImageNet and training from scratch (e.g., see the cyan and blue lines in Figure 1). A natural question arises, which motivates us to study the connection between pre-training and downstream robustness: what is it about the pre-training process that causes this striking shift in linear trend?

We tackle this question along three different ablation axes: **(i)** Label diversity of the pre-training distribution, **(ii)** Label semantics of the pre-training object classes, and **(iii)** Similarity between pre-training and fine-tuning distributions. Inspired by Huh et al. (2016), we perform these ablations using the inherent structure within both ImageNet (WordNet) and iNaturalist (biological taxonomy) class labels. To summarize our main findings:

1. We have found that while changes to the pre-training setup to increase the alignment between pre-training and fine-tuning distributions (e.g., using more semantically similar classes) can improve OOD performance, they have minimal impact on the robustness trends.

2. Similar to previous results that study influence of ImageNet pre-trained features on *accuracy* (Huh et al., 2016), we find that *label diversity* or *label granularity* doesn't have a strong impact on downstream robustness.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022. Do not distribute.

3. Finally, there is no robustness benefit from using noisier pre-training data, or a pre-training task that is more in line with the downstream wildlife classification setting.

From our initial findings, we hypothesize that with regards to finetuning, pre-training works on a "threshold basis": Once a certain threshold of pre-training data diversity is reached, the robustness trend shifts to a completely different regime compared to training models from scratch. How to make this shift gradual and how to characterize the limit of this shift are interesting open questions that we plan to continue to investigate. Refer to Section 4 for discussion of possible future directions.

2. Background

The main motivation for our paper comes from previous work by Huh et al. (2016), which investigates various factors that affect the quality of ImageNet pre-trained features for transfer learning on a variety of downstream tasks. In our work, we shift the focus from accuracy to robustness to distribution shift, which has been a long-standing issue in machine learning (Quiñonero-Candela et al., 2008; Szegedy et al., 2013; Biggio & Roli, 2018; Biggio et al., 2013). In particular, we analyze the robustness of pre-trained features to natural distribution shifts, i.e. where the test images could be observed in the real world and are not intentionally perturbed by synthetic corruptions, through the iWildCam-WILDS benchmark (Koh et al., 2021). More details on our experiments can be found in Section 3.1.

Ideally a natural shift between two test distributions should not affect performance, for instance because the shift does not affect the accuracy of humans labelers (Shankar et al., 2020). If model performances on the two test sets are plotted along the x- and y-axis of a scatter plot, then a more robust model would lie closer to the diagonal $y = x$ line. This notion of robustness was captured by Taori et al. (2020) under the term *effective robustness*, which measures the difference between a model's out-of-distribution (OOD) performance and that predicted from having higher in-distribution (ID) performance.

Miller et al. (2021) adopted this effective robustness framework and evaluated hundreds of models on various distribution shift settings. The authors observed that when the training data distribution is fixed, changes to model architecture, training set size, training algorithm, and other model-related factors do *not* change effective robustness in most cases. In other words, any model trained on the same data distribution should lie on the same linear trend that maps ID accuracy to OOD accuracy. This linear trend, and how close it is to the $y = x$ line, is what we also use in our work to compare the quality of the pre-trained features. More notably, Miller et al. (2021) discovered that on iWildCam dataset, models trained from scratch and models that have been pre-trained on ImageNet lie on distinct linear trends, with the latter exhibiting much more robustness. Motivated by this observation, our work seeks to better understand aspects of ImageNet pre-training that contribute to this higher robustness on downstream tasks, and how they translate to other pre-training data sources such as iNaturalist.

Previous work (Andreassen et al., 2021) has also studied effective robustness over the course of fine-tuning and found that the pre-trained models exhibit high effective robustness in the middle of fine-tuning, which eventually decreases as the training proceeds. The paper experimented with ImageNet as one of the pre-training data sources. In our investigation, as a sanity check to remove training epoch as a potential source of bias for the linear fit, we adopt the linear trend of models pre-trained on ImageNet and fine-tuned on iWildCam computed previously by Miller et al. (2021) as the baseline. We then report the residuals (i.e., effective robustness) from comparing actual OOD performance at different epochs to what could be predicted from the corresponding ID F1 score using this baseline. Refer to Appendix Figure 4 for more details. We find that in the context of iWildCam fine-tuning, at each epoch, the residuals from our architectures of choice concentrate and appear symmetric around the $y = 0$ line. This in turn allows us to use fine-tuning epoch as a hyperparamter that we vary in order to obtain models covering a wide range of test performances for the scatter plots.

3. Experiment Results

3.1. Setup

We use ImageNet (Deng et al., 2009) and iNaturalist (Van Horn et al., 2018) as the pre-training datasets, given their hierarchical structures, complexity, and relevance to the downstream task. The downstream task is classification with iWildCam-WILDS dataset (Koh et al., 2021): input is a photo taken by a camera trap, and output is one of 182 different animal species. There are 2 test sets for evaluation: ID test data consists of images taken by the same camera traps as the training set, but on different days from the training and validation (ID) images, while OOD test data contains images taken by a disjoint set of camera traps from training and validation (ID) images. We report the macro F1 scores of the trained networks, following (Koh et al., 2021), since this metric emphasizes performance on rare species, which are critical to the biodiversity monitoring application that iWildCam was intended for.

We then train a range of standard neural network architectures such as ResNet (He et al., 2016), ResNext (Xie et al., 2017), DenseNet (Iandola et al., 2014), AlexNet (Krizhevsky et al., 2012) and MobileNet V3 (Howard et al.,

2019) to obtain data for our linear trends. In our scatter plots, besides varying the architectures, we also vary the extent of training (i.e., epoch) to obtain points with different F1 scores for plotting. Further training details can be found in Appendix B.

In the subsequent scatter plots, we focus on the linear trends that result from our interventions with the pre-training datasets, and exclude explicit markers for the lines that we obtained from previous work (Miller et al., 2021), which include models trained from scratch (blue line) on iWildCam as well as models pre-trained on ImageNet (cyan line).

3.2. Effect of Number of Classes

We start by adapting the question raised in previous work (Huh et al., 2016) to our investigation—how does varying the number of pre-training classes affect the downstream robustness? We follow Huh et al. (2016) and construct *supersets* of classes in ImageNet using the WordNet hierarchy. We use the maximum of the shortest path distance from root of WordNet to a label to compute the maximum depth of the current label set. We contract label nodes along the shortest path to construct superclasses. Specifically, we investigate depths 5, 6, and 7, which result in class counts of 37, 85, and 232 respectively, to provide good coverage across a range of label granularity.

As seen in Figure 1, with depth 7, models are able to reach higher F1 scores. Other than that, we observe that the robustness trend does not vary much with increased label granularity. Overall we find that pre-training with the full 1000 classes provides the most robustness. However, with the same data and about 25% of the labels, robustness only decreases slightly, and further reduction of the label set doesn't harm model robustness. This is similar to the previous observation made by Huh et al. (2016), which shows that increases in number of classes past a certain point have diminishing return on downstream task *accuracy*. Further experiment with label contraction is needed to test the limit of label diversity on downstream robustness.

3.3. Animals versus Objects

We next investigate whether pre-training with classes whose semantics are more aligned with the downstream data would improve robustness. To do so, we experiment with separately pre-training models on ImageNet classes that are subsets of the "object" and "animal" WordNet synsets. This yields 2 broad categories that are roughly similar in sample size, each having around 600K images. In Figure 2, we find that using "animal" classes yields slightly better OOD performance, which is not surprising given that the fine-tuned data distribution, iWildCam, comprises images of animals in the wild. However, it's worth noting that the linear trends of these 2 very different categories of pre-training data still closely follow each other, and are much better than the linear relationship provided by models without any pre-training (blue line).

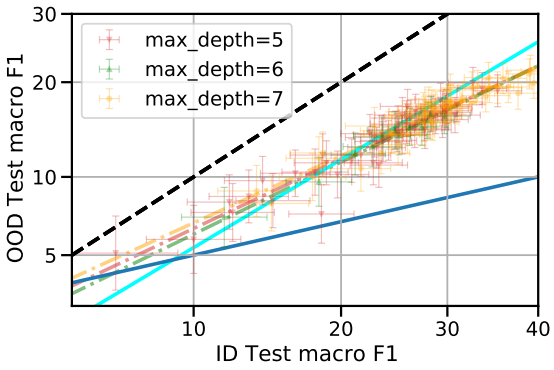

Figure 1. Varying the label sets using the WordNet tree changes the label diversity of the pre-training task (ImageNet). We observe that the robustness on the downstream task decreases compared to using the full 1000 classes. However, further constriction of the label set (i.e., reduces the maximum depth of the WordNet tree where labels are obtained from) has negligible effect on downstream robustness.

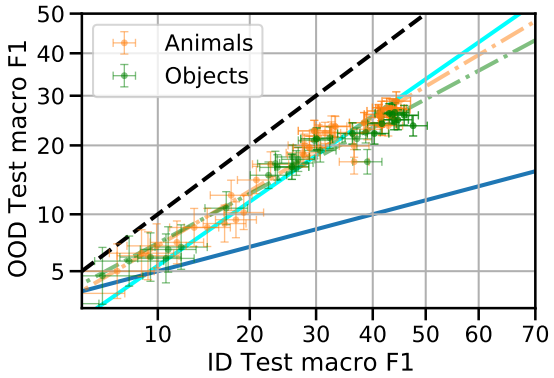

Figure 2. Varying the broad category of classes included in the pre-training data yields similar robustness, with those pre-trained only on "animal" classes still having slightly better OOD performance. Models pre-trained only on "object" classes are still much more robust than models that do not undergo any pre-training.

We hypothesize that this is because some images from "object" classes also contain animals (due to co-occurrences that are not accounted for by ImageNet labels), and that training on a diverse set of classes in general helps the model pick up on useful robustness-inducing variances that in turn lead to similar downstream robustness. Future work could explore Visual Genome dataset (Krishna et al., 2017), which comes with dense annotations of all objects in an image, thus allowing a clearer separation between training inputs with only "objects" and those with only "animals".

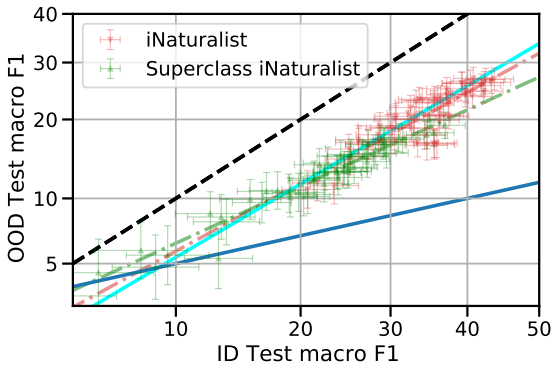

Figure 3. Pre-training on a noisy, long-tailed distribution of natural images (iNaturalist) doesn't change the robustness on downstream task, compared to pre-training on a clean dataset like ImageNet. Similar to Section 3.2, we observe that reducing label diversity by grouping related classes in iNaturalist together makes the fine-tuned models less robust.

3.4. Pre-training with iNaturalist

Moving beyond aligning the semantics of classes in the pre-training and downstream tasks, we ask whether increased similarity betwween the tasks themselves would help with robustness. This leads us to experimenting with iNaturalist (Van Horn et al., 2018).

Compared to ImageNet, iNaturalist exhibits vastly different characteristics (e.g., long-tailed, less clean data, different categories of objects, more domain-specific). Its data collection procedure is also much more similar to iWildCam's. We hypothesize that pre-training on the diverse species represented in iNaturalist will provide a boost on robustness for the animal-in-the-wild classification setting in iWildCam.

However, in Figure 3, we find that iNaturalist (red line) behaves similarly to ImageNet (cyan line) as a pre-training data source. When we reduce the class label space to its phylum, model robustness deteriorates slightly (green line), which is in line with our previous observation from the ImageNet experiments (Section 3.2). This again illustrates that label diversity is important to a certain extent in the pre-training phase.

4. Conclusion

In our work, we have demonstrated that many important factors during pre-training — label diversity, label semantics, and the pre-training dataset itself — do not significantly alter downstream robustness to distributional shifts. This leads to the natural question, what does? There are other aspects of the pre-training distribution that we look forward to exploring next:

- Sample diversity: given the same set of class labels, would sampling data from more fine-grained subgroups (e.g., different dog breeds) help with robustness? We are aware that "image diversity" is itself difficult to define, and other diversity heuristics such as FID score (Heusel et al., 2017) could be used to provide a more complete picture.

- Dataset size: How quickly robustness improvements saturate with more samples per class? A related question is how much pre-training data would be needed to alter the linear trend from standard training.

Our findings so far lead us to posit a "threshold hypothesis": Once a certain threshold of pre-training "diversity" is reached, the robustness trend shifts suddenly and discretely to new behavior, compared to training from scratch. Validating this threshold hypothesis would require more large-scale experimentation.

Another interesting open question is to determine what characteristic of iWildCams-WILDS leads to the difference in linear trend between pre-training and training from scratch. Many other datasets (e.g., fMoW, see (Koh et al., 2021)) do not exhibit this behavior after fine-tuning, so it is important to pinpoint other distribution shifts where this is the case. We propose that finding a unifying quality among such datasets would allow for better interpretation of our current results, and perhaps allow for interesting benchmarks with which to test the quality of pre-training features.

Finally, pre-training on web-crawled datasets has been gaining popularity recently, for producing large-scale models with remarkable performance in zero-shot settings (e.g., CLIP (Radford et al., 2021)). The dataset size reported is often multiple orders of magnitude bigger than ImageNet. With this scale in mind, a pertinent question to ask is how "diversity" of the pre-training distribution would be defined and measured differently given the open vocabulary that comes with web-crawled datasets.

References

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

Biggio, B. and Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

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 *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Huh, M., Agrawal, P., and Efros, A. A. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. Mit Press, 2008.

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*, pp. 8748–8763. PMLR, 2021.

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

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/shankar20c.pdf.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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*, 33:18583–18599, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

A. Effect of Training Stage on Effective Robustness

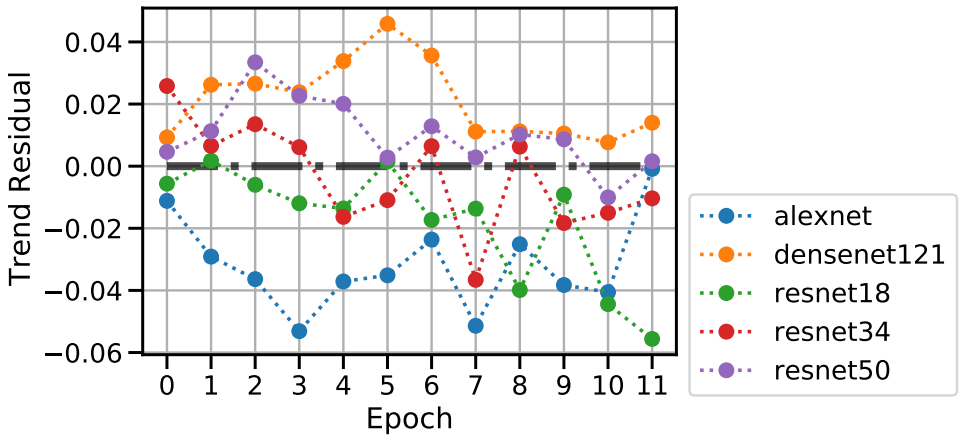

Figure 4. We visualize the residuals of various architectures after fitting a linear trend that predicts OOD accuracy from ID accuracy. All models are pre-trained on the full ImageNet dataset and fine-tuned on iWildCam for 12 epochs. We observe that overall the residuals fluctuate around the $y = 0$ line and residual values for most architectures fluctuate over the course of fine-tuning, except for alexnet, with negative residuals throughout, and densenet121, with positive residuals throughout.

B. Training Details

For the standard deep neural network architectures used in our investigation, their implementation comes from PyTorch's torchvision.models package. We use standard hyperparameters found in the WILDS and PyTorch's official GitHub repositories (for training models with iWildCam-WILDS and ImageNet respectively), which match the hyperparameters provided in (He et al., 2016) and (Koh et al., 2021). ImageNet pre-training happens for 90 epochs in total, and then all models are fine-tuned on iWildCam for 12 epochs.