ASSESSING MODEL OUT-OF-DISTRIBUTION GENERALIZATION WITH SOFTMAX PREDICTION PROBABILITY BASELINES AND A CORRELATION METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the use of Softmax prediction probabilities to assess model generalization under distribution shift. Specifically, given an out-of-distribution (OOD) test set and a pool of classifiers, we aim to develop a prediction probability-based measure which has a monotonic relationship with OOD generalization performance. We first show existing uncertainty measures (*e.g.*, entropy and maximum Softmax prediction probability) are fairly useful of predicting generalization in some OOD scenarios. We then move ahead with proposing a new measure, Softmax Correlation (SoftmaxCorr). To obtain the SoftmaxCorr score for a classifier, we compute the class-class correlation matrix from all the Softmax vectors in a test set, and then its cosine similarity with an identity matrix. We show that the class-class correlation matrix reveals significant knowledge about the confusion matrix: its high similarity with the identity matrix means predictions have low confusion (uncertainty) and evenly cover all classes, and vice versa. Across three setups including ImageNet, CIFAR-10, and WILDS, we show that SoftmaxCorr is well predictive of model accuracy on both in-distribution and OOD datasets.

1 INTRODUCTION

Understanding the generalization of deep neural networks is an essential problem in deep learning. There is substantial interest in predicting ID generalization gap via complexity measures (Jiang et al., 2020a; Neyshabur et al., 2015; Bartlett et al., 2017; Keskar et al., 2017; Nagarajan & Kolter, 2019; Neyshabur et al., 2017; Chuang et al., 2021; Jiang et al., 2020b; Smith & Le, 2018; Arora et al., 2018; Dziugaite & Roy, 2017; Dinh et al., 2017). Although significant, developing measures to characterize OOD generalization remains under-explored. In fact, the test environment in the real world often undergoes distribution shift due to factors like sample bias and non-stationarity. Ignoring the potential distribution shift can lead to serious safety concerns in self-driving cars (Kuutti et al., 2020) and histopathology (Bandi et al., 2018), *etc*.

Softmax prediction probability is found to be useful to analyze the test environment (Hendrycks & Gimpel, 2016; Guillory et al., 2021; Deng et al., 2022a; Liang et al., 2018; Garg et al., 2022). For example, Hendrycks & Gimpel (2016); Liang et al. (2018) utilize maximum Softmax prediction probability to identify samples from open-set classes. This gives us a hint: models' prediction probabilities may be informative to reflect their OOD performance. Therefore, we are interested in measures based on prediction probability and conduct large-scale correlation study to explore whether they are useful to characterize generalization of various models under distribution shift. Concretely, given various deep models, we aim to study and develop prediction probability-based measures which monotonically relate to model generalization.

To this end, we construct a catalog of empirical prediction probability-based measures and create a wide range of experimental setups. We collect 502 different classification models ranging from standard convolutional neural network to Vision Transformers. We cover 19 ID and OOD datasets with various types of distribution shift, such as ImageNet-V2 (Recht et al., 2019) with dataset reproduction shift and stylized-ImageNet (Geirhos et al., 2019) with style shift.

Based on experimental results, we observe that empirical uncertainty measures based on prediction probabilities (*e.g.*, entropy) are useful in characterizing OOD generalization to some extent. How-

ever, they are limited in leveraging class-wise relationship encoded in prediction probabilities. We thus further propose Softmax correlation (SoftmaxCorr), an effective prediction probability-based measure describing for each classifier to what extent classes correlate with each other. Specifically, for each classifier we compute a class correlation matrix from all prediction vectors in a test set. Then, we calculate its cosine similarity with an identity matrix to evaluate whether this classifier makes diverse and certain predictions. We show that class-class correlation effectively uncovers knowledge of confusion matrix, thus better reflecting overall accuracy on OOD test set. The broad correlation study shows the efficacy of SoftmaxCorr. Our contributions are summarized below.

- We observe that Softmax prediction probability-based measures generally give good baseline indicators of the OOD accuracy of various classification models.
- Furthering this finding, we propose SoftmaxCorr which assesses model generalization by explicitly calculating class-class correlation, a new angle to leverage the prediction probability. A wide range of experiment suggests the effectiveness of this method.

2 Related Work

Predicting generalization in deep learning studies the ID generalization gap (*i.e.*, the difference between training and test accuracy) of deep neural networks. Representative methods develop complexity measures based on model parameters and training set (Jiang et al., 2020a;b; Neyshabur et al., 2015; Keskar et al., 2017; Bartlett et al., 2017; Neyshabur et al., 2018; Liang et al., 2019; Chuang et al., 2021; Smith & Le, 2018; Arora et al., 2018; Dziugaite & Roy, 2017; Dinh et al., 2017), such as distance of training weights from initialization (Nagarajan & Kolter, 2019), the product of norms of weights across layers (Neyshabur et al., 2017) and the change of model accuracy with respect to different perturbation levels in training data (Schiff et al., 2021). These methods assume that training and test data come from the same distribution and do not incorporate characteristics of test data, so we can unlikely make reasonable predictions on a different distribution. To mitigate this limitation, we investigate the model generalization under distribution shift by developing measures that reflect models' generalization property on OOD datasets.

OOD generalization. Machine learning models should generalize from training distribution to OOD datasets (Djolonga et al., 2021; Koh et al., 2021). To study the problem of OOD generalization, several benchmarks are proposed (Hendrycks & Dietterich, 2019; Koh et al., 2021; Gulrajani & Lopez-Paz, 2021), such as corruption benchmark (Hendrycks & Dietterich, 2019) and domain generalization testbed (Gulrajani & Lopez-Paz, 2021). Moreover, several methods are proposed to improve model OOD generalization (Volpi et al., 2018; Zhao et al., 2020; Sagawa et al., 2020; Liu et al., 2021a; Mansilla et al., 2021; Shi et al., 2021), such as adversarial domain augmentation (Volpi et al., 2018; Qiao & Peng, 2021) and inter-domain gradient matching (Shi et al., 2021).

There are few works study the characterization of model OOD generalization. Ben-David et al. (2006; 2010) provide upper bounds of OOD generalization error for domain adaptation. Some works further bound the OOD generalization error based on the divergence between the two distributions (Acuna et al., 2021; Zhang et al., 2019; Tachet des Combes et al., 2020). However, as suggested by Miller et al. (2021), when the shift becomes larger, the above bounds on OOD performance become looser. Moreover, Vedantam et al. (2021) report that the adapting theory from domain adaptation is limited in predicting OOD generalization. In this work, we aim to assess model generalization under distribution shift by OOD measures.

Out-of-distribution detection aims to identify test samples that deviate from the training distribution (Hendrycks & Gimpel, 2016; Yang et al., 2021). Specifically, it detects test data that do not belong to any of the classes modeled in training distribution. Many methods are developed based on model outputs (Hendrycks & Gimpel, 2016; Liang et al., 2018; Hendrycks et al., 2022; Liu et al., 2020; Wang et al., 2022). For example, Hendrycks & Gimpel (2016) use maximum prediction probability to identify OOD samples. Hendrycks et al. (2022) further show that, without being normalized by Softmax function, maximal model logit is also a strong baseline for OOD detection. Liu et al. (2020) calculate an energy score based on Softmax probability, which is regarded as a simple yet effective replacement for maximum Softmax prediction probability. This work does not aim to detect OOD data, but to explore Softmax-based measures to assess model generalization.

3 TASK FORMULATION

Task definition. We consider a K-way classification task, and let $\mathcal{Y} = \{1, ..., K\}$ denote the label space and $\mathcal{X} \in \mathbb{R}^d$ denote the input space. We are given a labeled training set $\mathcal{D}^S := \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ that contains N_s data i.i.d drawn from a source distribution P_S, and an OOD test set $\mathcal{D}^T := \{(x_i, y_i)\}_{i=1}^{N}$ that contains N data i.i.d drawn from another distribution P_T ($P_S \neq P_T$). We train M neural network classifiers $\{\phi_m\}_{m=1}^{M}$ on $\mathcal{D}^S$. Given a sample $(\boldsymbol{x}, y)$ from $\mathcal{D}^T$, the classifier $\phi_m : \mathcal{X} \to \Delta^K$ generates Softmax prediction probabilities for $\boldsymbol{x}$ on K classes, where Δ^K denote $K-1$ dimensional unit simplex. By testing on $\mathcal{D}^T$, ϕ_m yields a prediction matrix $\boldsymbol{P} \in \mathbb{R}^{N \times K}$, whose each row represents prediction probabilities of a test data. Specifically, the prediction matrix satisfies $\sum_{j=1}^{k} P_{i,j} = 1 \ \forall i \in 1 \dots N$ and $P_{i,j} \geq 0 \ \forall i \in 1 \dots N, j \in 1 \dots K$, where $P_{i,j}$ indicates the probability that i-th sample is predicted to the j-th class.

The dataset has an evaluation metric (*e.g.*, accuracy) to obtain ground-truth generalization G_m of classifier ϕ_m. The goal is to design a measure to calculate a score S_m for each classifier ϕ_m without access to data annotations. The calculated scores $\{S_m\}_{m=1}^{M}$ ideally should well correlate with $\{G_m\}_{m=1}^{M}$, so that we can assess the OOD generalization of models based on the scores.

Evaluation metrics. We use Spearman's Rank Correlation coefficient ρ (Kendall, 1948) to measure monotonicity between calculated scores and model generalization. We additionally use the weighted variant of Kendall's rank correlation τ_w, which is shown to be useful when selecting the best ranked item is of interest (You et al., 2021). Both coefficients range from $[-1, 1]$. A value closer to -1 or 1 indicates a strong negative or positive correlation, respectively, and 0 means no correlation. Similar to Miller et al. (2021); Baek et al. (2022), we apply the same probit scale to both accuracy and SoftmaxCorr in our experiment for better linear fit.

4 SOFTMAX PREDICTION PROBABILITY-BASED OOD MEASURES

This work aims to design an OOD measure monotonically related to model generalization under distribution shift. We first discuss the motivation of developing OOD measure based on Softmax prediction probability. Then, we describe some commonly used uncertainty measures. Lastly, we introduce our method which considers more useful information than existing measures.

4.1 WHY USE SOFTMAX PREDICTION PROBABILITY?

Deep neural networks tend to make over-confident predictions (Ovadia et al., 2019; Minderer et al., 2021; Guo et al., 2017; Hein et al., 2019), which at the first look makes it less reliable to use Softmax Prediction as an uncertainty measure on the test data. However, existing works show that it is indeed informative in analyzing test environments. For example, Hendrycks & Gimpel (2016) show that the maximum Softmax prediction probability (MaxPred) of correct samples tends to be higher than that of incorrect or OOD samples. Guillory et al. (2021); Garg et al. (2022) suggest it is useful to estimate accuracy of a trained classifier by calculating its MaxPred on test samples.

Proof of concept. The above works imply the prediction probability is likely to be effective in measuring OOD performance of a pool of models. Given an OOD test set (ImageNet-R) and various ImageNet models, we conduct correlation study between MaxPred and classification accuracy. In Fig. 1, we show that there is a moderate correlation between MaxPred and model accuracy ($\rho = 0.613$ and $\tau_w = 0.776$) on ImageNet-R. This indicates that MaxPred is feasible in assessing OOD performance. Based on this observation, we further explore more empirical prediction probability-based measures and develop a more effective measure which exploits more semantics reflected in the prediction probability.

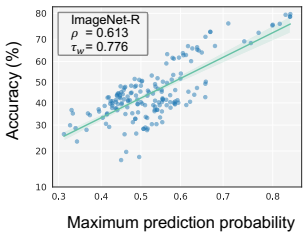

Figure 1: Correlation study between MaxPred and accuracy (%) on ImageNet-R.

4.2 EXPLORING MORE EMPIRICAL MEASURES

In addition to MaxPred, we investigate other empirical measures below.

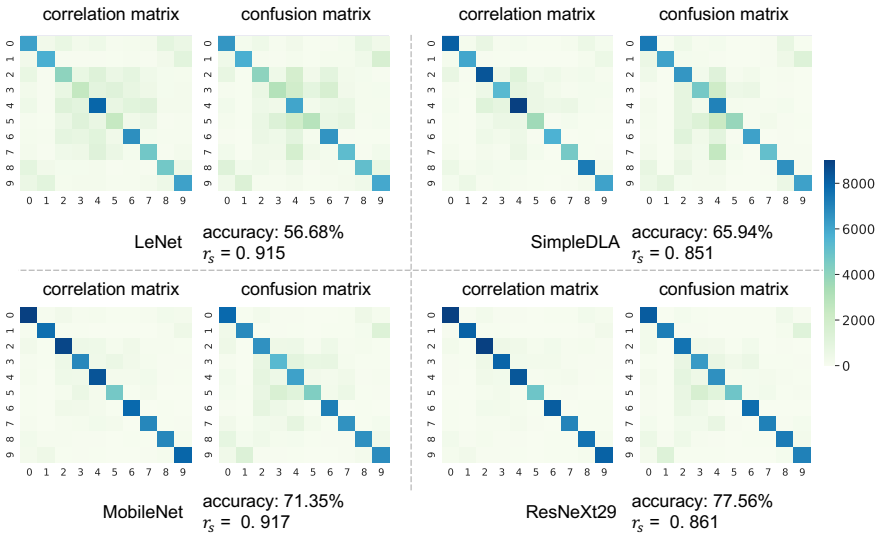

Figure 2: Comparison between the correlation matrix computed from Softmax vectors and confusion matrix. We show comparisons obtained by four classifiers: LeNet, SimpleDLA, MobileNet and ResNeXt29. The test set is CINIC. For each pair of matrices, we use representational similarity analysis (r_s) to measure their similarity. The high $r_s > 0.85$ indicate that the class correlation matrix reveals knowledge of confusion matrix and therefore model accuracy.

Softmax Gap (SoftGap). MaxPred only utilizes the maximal value of Softmax vectors while disregarding values on other entries. We introduce SoftGap based on MaxPred which further considers the second largest entry in a prediction vector. Specifically, it calculates the average of difference between the largest and second largest Softmax prediction probabilities over all samples. A larger Softgap indicates prediction is more certain, suggesting better generalization.

Negative Predictive Entropy (Entropy) (Pereyra et al., 2017; Vedantam et al., 2021). This computes negative entropy of $\boldsymbol{P}$ by $-\frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{P}_{i,:})$, where $H(\cdot)$ is Shannon entropy (Shannon, 1948). A larger Entropy means a model makes more confident predictions.

Information Maximization (InfoMax). It is a loss function used in unsupervised domain adaptation algorithms (Shi & Sha, 2012; Liang et al., 2020). Its formula is $H(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{P}_{i,:}) - \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{P}_{i,:})$. The idea is that a well-generalized model has diverse predictions over all samples (first term) while being certain on individuals (second term). Specifically, an accurate model tends to have a large entropy of the average prediction probability of all samples and a small entropy on each data point.

4.3 Our Method: Softmax Correlation (SoftmaxCorr)

Given the prediction matrix $\boldsymbol{P} \in \mathbb{R}^{N\times K}$ predicted by ϕ_m, a class correlation matrix $\boldsymbol{C} \in \mathbb{R}^{K\times K}$ can be computed by $\boldsymbol{C} = \boldsymbol{P}^\intercal \boldsymbol{P}$. An entry $C_{i,j}$ indicates the correlation between prediction probabilities of class i and class j over all samples, and is computed by $C_{i,j} = \sum_{n=1}^{N} P_{n,i}P_{n,j}$. We define the sum of diagonal entries of the correlation matrix as intra-class correlation (IntraCorr) and off-diagonal elements as inter-class correlation (InterCorr).

Our motivation is two-fold. **First**, a well-generalized model is certain on each sample. This means that samples are classified with high probabilities. Hence, the corresponding correlation matrix has large IntraCorr and small InterCorr. **Second**, when a model always confidently classifies images to a certain class, IntraCorr is large while the ground truth generalization is poor. To mitigate this limitation, we additionally measure whether models make predictions different from each other, while similar to one-hot encoding. Specifically, we expect a high score assigned to the correlation matrix which has large and evenly distributed values on diagonal entries and small values on off-diagonal elements. To this end, we define Softmax Correlation (SoftmaxCorr) as the *cosine similarity* between correlation matrix $\boldsymbol{C}$ and an identity matrix $\boldsymbol{I}_K$. SoftmaxCorr assigns high scores to models which are confident on each individual and make diverse predictions over all samples.

In Fig. 2, we compare the computed class correlation matrices and ground-truth confusion matrices of LeNet (LeCun et al., 1998), SimpleDLA (Yu et al., 2018), MobileNet (Howard et al., 2017) and ResNeXt29 (Xie et al., 2017) on CINIC. We use representational similarity analysis (r_s) (Kriegeskorte et al., 2008; Dwivedi & Roig, 2019) to calculate the similarity between each pair of class correlation and confusion matrices. We observe that r_s are high ($r_s > 0.85$). This suggests that the correlation matrix encodes critical and informative knowledge of the level of confusion between classes and thus directly relates to model accuracy.

5 EXPERIMENT

In this section, we firstly describe the ImageNet, CIFAR-10 and WILDS setups. Then, we analyze the experiment results of prediction probability-based measures on three setups. After that, we compare SoftmaxCorr with other characterizations of correlation matrix. Moreover, we test whether SoftmaxCorr can model checkpoints along the training trajectory and assess models trained by different domain adaption algorithms. Lastly, we study the correlation between SoftmaxCorr and accuracy of a single model on various OOD test sets. We note that for the completeness of experiment, we also include the performance of Softmax measures on ID test sets.

5.1 EXPERIMENTAL SETUP

ImageNet setup. We collect 173 models publicly accessible from TIMM (Wightman, 2019). They are trained or fine-tuned on ImageNet (Deng et al., 2009) and vary in architectures and training strategies. We use both ID and OOD datasets for correlation study. Specifically, ImageNet-Val is ID test set. OOD datasets are: (1) ImageNet-V2 (Recht et al., 2019); (2) ObjectNet (Barbu et al., 2019); (3) ImageNet-S(ketch) (Wang et al., 2019); (4) ImageNet-Blur severity 5 (Hendrycks & Dietterich, 2019); (5) ImageNet-R(endition) (Hendrycks et al., 2021); (6) Stylized-ImageNet (Geirhos et al., 2019). Note that, ImageNet-V2 has three versions (ImageNet-V2-A/B/C) resulting from different sampling strategies. ImageNet-R and ObjectNet contain 200 and 113 ImageNet classes respectively.

CIFAR-10 setup. We collect 65 networks trained with the strategy provided by Wightman (2017) on CIFAR-10 training set (Krizhevsky et al., 2009). CIFAR-10-Val(idation) is the ID test set. For OOD datasets, we use (1) CIFAR-10.2 (Recht et al., 2018) (2) CINIC (Darlow et al., 2018) (3) CIFAR-10-Noise with severity 5 (Hendrycks & Dietterich, 2019).

WILDS setup. We follow the same training schema provided by Koh et al. (2021) to train or fine-tune models for Camelyon17 (Bandi et al., 2018) and iWildCam (Beery et al., 2021). **Camelyon17** is a binary classification dataset where the objective is to classify whether a slide contains tumor issue. We use 45 models varying in architectures and random seeds. **iWildCam** is a 182-way animal classification dataset. We collect 66 models the variations of which results from different network architectures and learning rates. Model performance is measured by macro-$F1$ score.

5.2 MAIN OBSERVATIONS

SoftmaxCorr exhibits a strong correlation with model generalization. In Fig. 3 and Table 1, we observe that SoftmaxCorr is indicative of model performance under the three setups. Particularly on the ImageNet setup, SoftmaxCorr has consistently stronger correlations than the other empirical measures. For example, the average Spearman's Rank Correlation coefficient ρ is 0.910, 0.671 and 0.764, and 0.361, and 0.433 for our method, MaxPred, SoftGap, Entropy and InfoMax, respectively. On CIFAR-10 and WILDS, while on some test sets it does not present the strongest correlation, we still think that SoftmaxCorr is a preferred measure because it has higher or very competitive average correlation scores ($\rho = 0.851, 0.861$ and $\tau_w = 0.645, 0.717$).

SoftmaxCorr gives more stable correlation, while the other four measures have mixed performance on different test sets. On CIFAR-10-Noise, we find that SoftmaxCorr correlates well with model performance ($\rho = 0.813$ and $\tau_w = 0.771$). In contrast, Entropy, MaxPred and SoftGap show no correlation ($\rho < 0.05$ and $\tau_w < 0$). Although InfoMax has higher rank correlation than SoftmaxCorr ($\rho = 0.896$ *vs.* 0.813), it does not correlate with accuracy on ImageNet-Val, ImageNet-V2-A/B/C ($\rho < 0.25$ and $\tau_w < 0.35$). On ID and OOD sets of iWildCam, SoftGap and MaxPred have higher correlations than SoftmaxCorr. However, they fail to measure the macro-F1

Table 1: Method comparison on ImageNet, CIFAR-10, and WILDS. We compare our SoftmaxCorr with four empirical Softmax prediction probability-based measures: MaxPred, SoftGap, Entropy and InfoMax. We use Spearman's rank correlation (ρ) and weighted Kendall's correlation (τ_w) for comparison. The highest correlation in each row is highlighted in **bold**. We show that our method is stable and yields the highest average correlation across three setups.

Setup	Dataset	MaxPred		SoftGap		Entropy		InfoMax		SoftmaxCorr	
		ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w
ImageNet	ImageNet-Val	0.542	0.515	0.665	0.568	0.142	0.286	0.145	0.288	**0.919**	**0.752**
	ImageNet-V2-A	0.645	0.574	0.747	0.622	0.217	0.327	0.222	0.330	**0.922**	**0.792**
	ImageNet-V2-B	0.506	0.483	0.636	0.554	0.109	0.256	0.113	0.266	**0.912**	**0.785**
	ImageNet-V2-C	0.577	0.528	0.699	0.594	0.169	0.303	0.173	0.306	**0.917**	**0.795**
	ImageNet-R	0.613	0.776	0.797	0.839	0.295	0.623	0.423	0.688	**0.925**	**0.880**
	ImageNet-S	0.784	0.798	0.839	0.825	0.471	0.652	0.614	0.730	**0.875**	**0.883**
	Stylized-ImageNet	0.580	0.609	0.661	0.656	0.233	0.431	0.512	0.613	**0.828**	**0.834**
	ObjectNet	0.912	0.814	**0.934**	0.848	0.820	0.735	0.836	0.761	0.927	**0.897**
	ImageNet-Blur	0.877	0.828	0.898	0.843	0.780	0.702	0.850	0.807	**0.967**	**0.929**
	Average	0.671	0.665	0.764	0.705	0.361	0.480	0.433	0.520	**0.910**	**0.839**
CIFAR-10	CIFAR-10-Val	0.860	0.512	**0.871**	**0.551**	0.844	0.494	0.844	0.494	0.865	0.536
	CIFAR-10.2	0.857	0.696	0.868	**0.705**	0.826	0.663	0.832	0.669	**0.884**	0.694
	CINIC	0.753	0.375	0.777	0.402	0.722	0.331	0.735	0.348	**0.844**	**0.578**
	CIFAR-10-Noise	0.032	-0.082	0.041	-0.077	0.035	-0.090	**0.896**	0.750	0.813	**0.771**
	Average	0.626	0.375	0.639	0.395	0.607	0.350	0.827	0.565	**0.851**	**0.645**
WILDS	iWildCam-ID	**0.972**	**0.889**	0.969	0.887	0.967	0.885	0.960	0.810	0.945	0.792
	iWildCam-OOD	0.917	0.840	**0.936**	**0.846**	0.853	0.798	0.883	0.792	0.916	0.839
	Camelyon17-ID	0.847	0.718	0.847	0.718	0.728	0.611	0.833	0.653	**0.893**	**0.746**
	Camelyon17-OOD	0.192	0.320	0.192	0.320	0.167	0.299	**0.695**	**0.635**	0.689	0.492
	Average	0.732	0.691	0.736	0.692	0.679	0.647	0.842	**0.725**	**0.861**	0.717

on Camelyon17-OOD ($\rho = 0.192$). While in some cases SoftmaxCorr does not present the highest correlation, we emphasize that it overall gives more stable and stronger correlations. Thus, we think that SoftmaxCorr is generally more indicative of model generalization than other measures.

Compared to SoftGap and InfoMax, SoftmaxCorr better utilizes Softmax prediction probabilities. We use SoftGap on top of MaxPred as a simple approach to explicitly consider more entries (the second largest probability) in Softmax predictions. Based on the entropy, InfoMax further considers the prediction diversity (its first term) and thereby it also leverages more information of Softmax predictions. As shown in Table 1, both methods achieve higher correlation results than MaxPred. This indicates that using more information of Softmax predictions is useful. Compared with SoftGap and InfoMax, the class-wise correlation considered in our SoftmaxCorr better reveals the knowledge encoded by Softmax predictions. This is supported by the higher average correlation and more stable performance of SoftmaxCorr under three setups.

5.3 Discussion on the Characterization of Correlation Matrix

To investigate the importance of prediction diversity and inter-class correlation for predicting OOD generalization, we compare SoftmaxCorr with two approaches: (1) the sum of intra-class correlation (Variant-A); (2) cosine similarity between an identify matrix and correlation matrix with only diagonal elements (Variant-B). In Table 2, we present rank correlation of three methods on six datasets from three setups. We see that SoftmaxCorr has average higher correlation than Variant-A and Variant-B ($\rho = 0.863$ *vs.*

Dataset	Variant-A	Variant-B	SoftmaxCorr
ImageNet-V2-A	0.503	**0.924**	0.922
ImageNet-R	0.669	0.902	**0.925**
CIFAR-10.2	0.854	0.860	**0.884**
CINIC	0.752	0.817	**0.844**
iWildCam-OOD	**0.916**	-0.731	**0.916**
Camelyon17-OOD	0.174	0.209	**0.689**
average	0.645	0.497	**0.863**

Table 2: Comparison between SoftmaxCorr, Variant-A and Variant-B. Rank correlation (ρ) is the metric.

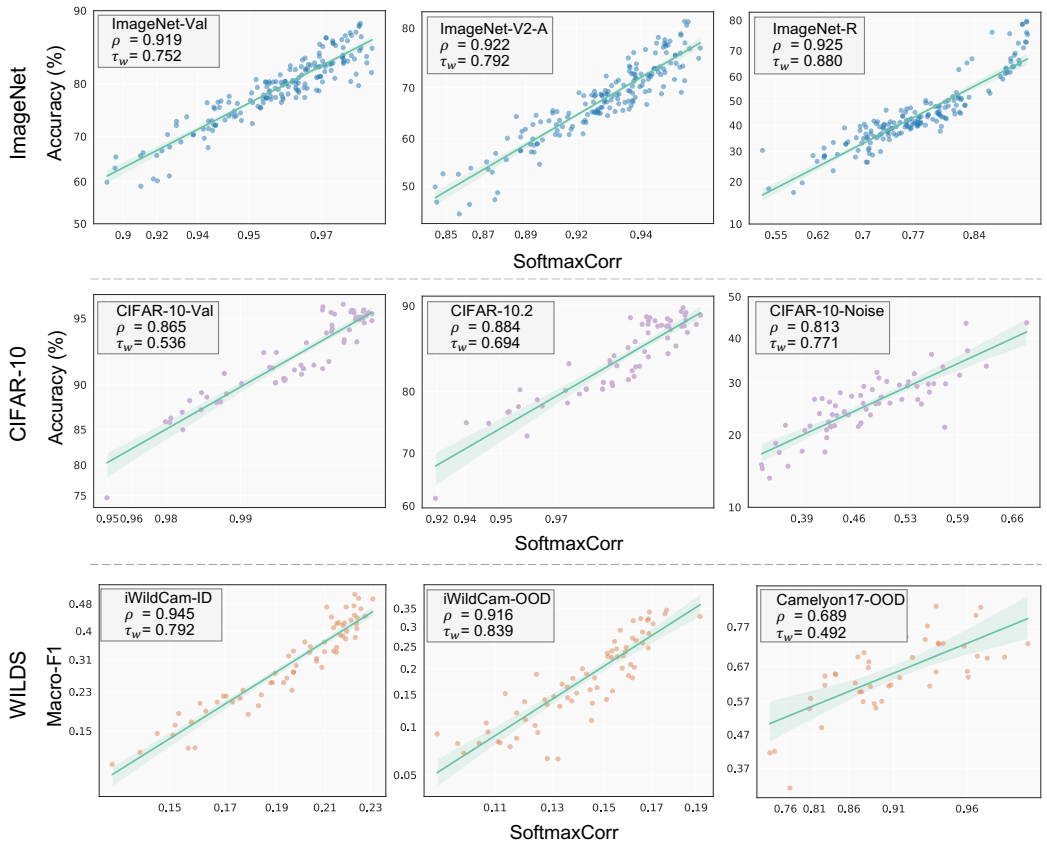

Figure 3: SoftmaxCorr *vs.* model generalization under ImageNet, CIFAR-10 and WILDS setups. For the ImageNet setup, the ID test set is ImageNet-Val; OOD test sets are ImageNet-V2-A and ImageNet-R. For CIFAR-10 setup, the ID test set is CIFAR-10-Val, while OOD test sets are CIFAR-10.2 and CIFAR-10-Noise. For WILDS, the ID test set is iWildCam-ID, and OOD test sets are iWildCam-OOD and Camelyon17-OOD. The y-axis is top-1 accuracy, top-1 accuracy and macro-F1 for the three setups, respectively. The straight lines are fit with robust linear regression (Huber, 2011). Axes are probit scaled as described in Section 3. We observe that SoftmaxCorr is a reliable and effective metric. Particularly on ImageNet, SoftmaxCorr is predictive of model generalization and selective of models with strong performance ($\rho > 0.92$ and $\tau_w > 0.75$).

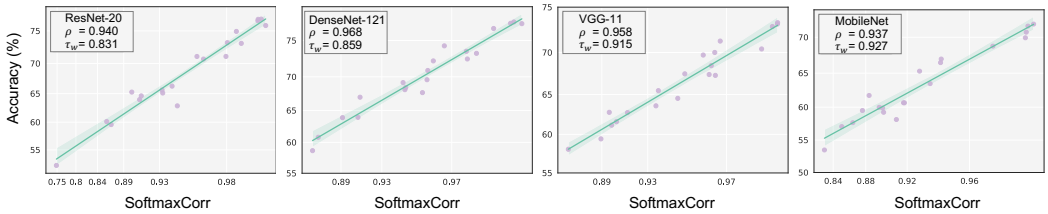

Figure 4: Correlation analysis: SoftmaxCorr and accuracy on CINIC. Each point represents a checkpoint. We consider CIFAR-10 models: ResNet-20, DenseNet-121, VGG-11 and MobileNet. Axes are probit scaled as in Section 3. For four models, we see strong correlations ($\rho > 0.93$). This suggests that SoftmaxCorr is helpful in assessing checkpoints along the training process.

0.645 *vs.* 0.497). On ImageNet-V2-A and ImageNet-R, SoftmaxCorr is a good indicator of accuracy, while IntraCorr shows a weaker correlation. This suggests that measuring prediction diversity benefits generalization assessment. On iWildCam-OOD and Camelyon17-OOD, Variant-B presents a negative or very weak correlation, but SoftmaxCorr remains informative. This means that considering inter-class correlation is important for predicting OOD generalization.

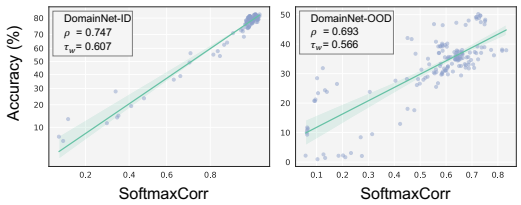

Figure 5: SoftmaxCorr *v.s.* accuracy (%) under domain adaptation. We observe meaningful correlation especially under ID test set.

Test set	DomainNet-ID	DomainNet-OOD
MaxPred	0.725/0.570	0.403/0.274
SoftGap	0.744/0.597	0.407/0.258
Entropy	0.677/0.523	0.386/0.253
InfoMax	0.698/0.539	**0.763/0.721**
SoftmaxCorr	**0.747/0.607**	0.693/0.566

Table 3: Comparison of five measures under domain adaptation (ρ/τ_w). Our method is competitive on both ID and OOD test sets.

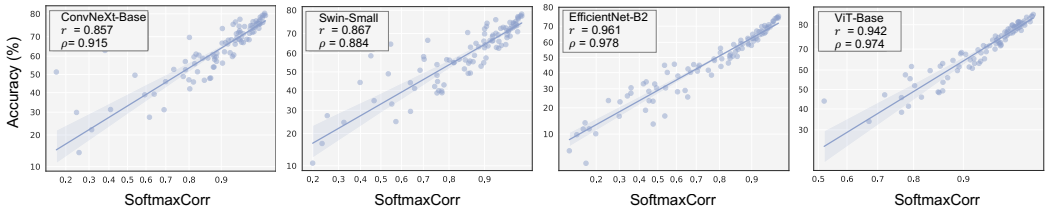

Figure 6: SoftmaxCorr *v.s.* accuracy on ImageNet-C benchmark. We observe strong linear and rank correlation for ImageNet models: ConvNext-Base, Swin-Small, EfficientNet-B2 and ViT-Base.

5.4 Evaluating Generalization along A Training Trajectory

In previous sections, SoftmaxCorr is utilized to measure the performance of models varying in different architectures and training strategies. In practice, we are sometimes interested in evaluating the models at different training check points. Hence, we also analyze whether SoftmaxCorr is helpful in this case. We collect prediction probabilities of CINIC every 10 epochs along the training process of ResNet-20, DenseNet-121 (Huang et al., 2017), VGG11 (Simonyan & Zisserman, 2014) and MobileNet (Howard et al., 2017) trained on the CIFAR-10. In Fig. 4, we observe that SoftmaxCorr has high rank correlation ($\rho > 0.93$) with model performance for four networks. This means that we can potentially apply SoftmaxCorr to assay model generalization along the training process.

5.5 Assessing Predictive Ability of SoftmaxCorr under Domain Adaptation.

So far in our experiment, models are trained by empirical risk minimization (Vapnik, 1991) without using unlabeled test samples. In domain adaptation (DA) (Kouw & Loog, 2019; Zhou et al., 2022), there are various algorithms proposed to improve generalization by training models on both labeled training set and unlabeled OOD test set. Here, we discuss whether SoftmaxCorr remains effective under such scenario. To this end, we conduct experiment on the commonly used domain adaptation benchmark DomainNet (Peng et al., 2019; Sagawa et al., 2021), which is a 345-way object classification task. Following Sagawa et al. (2021), we use the "Real" domain to train models, use its validation set as the ID test set, and take the validation set of "Sketch" domain as the OOD test set. There are 8 different optimization algorithms considered. In total, we collect 152 models for the correlation study. In Fig. 5, we notice SoftmaxCorr remains informative on ID and OOD test sets of DomainNet ($\rho = 0.747, 0.693$). According to Table 3, InfoMax provides a higher correlation on OOD test set but not on ID test set. InfoMax is commonly used in DA (Shi & Sha, 2012; Liang et al., 2020), and it might better reflect models' target-adaptation ability than source-adaptation ability. While SoftmaxCorr has lower correlation on OOD test set than InfoMax, it still presents a meaningful result on OOD test set and higher correlation on ID test set.

5.6 SoftmaxCorr v.s. Generalization on Various OOD Test Sets

We investigate how a given trained model generalize to various OOD datasets. In detail, we evaluate a single model on all test sets of ImageNet-C benchmark (Hendrycks & Dietterich, 2019) and conduct correlation study between accuracy and SoftmaxCorr. We additionally use Pearson's corre-

lation (r) to measure overall linear trend. This coefficient varies in $[-1, 1]$. A value closer to $-1/1$ indicates better negative/positive linearity and 0 means no correlation. We use ImageNet networks: ConvNeXt-Base (Liu et al., 2022), Swin-Small (Liu et al., 2021b), EfficientNet-B2 (Tan & Le, 2019) and ViT-Base (Dosovitskiy et al., 2020). Figure 6 shows a strong linear relationship and high rank correlation ($r > 0.85$ and $\rho > 0.91$). It indicates that with a linear regressor SoftmaxCorr can also help estimate the accuracy of a given model on various test sets.

6 Discussion and Potential Directions

What makes OOD measures interesting? Miller et al. (2021) report an accuracy-on-the-line phenomenon where there exists a very strong linear correlation between probit-scaled ID and OOD generalization. It means ID generalization is a good predictor of OOD generalization. There are three reasons that make it meaningful to investigate OOD measures. **First**, *Softmax prediction probability can be attained with just unlabeled data.* This makes prediction probability-based measures of important practical value as indicators of OOD generalization. **Second**, Miller et al. (2021) demonstrate that accuracy-on-the-line is not universal. That is, on some datasets, ID and OOD generalization do not show a clear positive correlation. This finding is further discussed in recent work (Wenzel et al., 2022). Specifically, Wenzel et al. (2022) suggest two patterns preventing this phenomenon: (1) underspecification (*e.g.*, Camelyon17) where the same ID performance leads to different OOD behavior; (2) models do not transfer information from ID to OOD domains (*e.g.*, DomainNet). That is, despite of various ID performance, all models perform poorly on OOD datasets. Paradoxically, identifying the patterns itself requires labeled OOD data. Therefore, while ID performance is indicative in some scenarios, we emphasize that SoftmaxCorr is overall a good alternative when labeled ID data is inaccessible or accuracy-on-the-line does not hold. **Third**, it is demanding and sophisticated to design ID test sets (Engstrom et al., 2020). They are expected to be unbiased and representative of distribution to effectively measure model ID performance. Further, it is a trade-off to split a full dataset into training, validation and test sets in terms of training and evaluation quality.

Discussion on imbalanced test sets. SoftmaxCorr is defined as the cosine similarity between correlation matrix $\boldsymbol{C}$ and an identity matrix $\boldsymbol{I}_K$. This implicitly assumes the test set is balanced: the number of samples in each class is roughly equal. This definition may not optimal when the test set is extreme imbalanced (*e.g.*, some classes may only have one image being predicted into). For this point, we note that some test sets in our study are moderately imbalanced, such as ImageNet-R, ObjectNet, iWildCam and DomainNet. We empirically observe that SoftmaxCorr still maintains a high correlation on these datasets. Under extremely imbalanced case, the definition needs to consider the importance weighting of each class. It would be helpful to use techniques of label shift estimation (Lipton et al., 2018; Garg et al., 2020; Sun et al., 2022), and we leave it as our future work.

Potential OOD measures. This work proposes SoftmaxCorr to use class-wise relationship encoded by Softmax prediction probabilities. There are other potential ways. For example, SoftGap computes the difference of largest and second largest prediction probabilities. We show that SoftGap exhibits a stronger correlation with performance compared to MaxPred. It would be interesting to improve SoftGap by utilizing more probabilities (*e.g.*, top five probabilities). Furthermore, this work focuses on prediction probability. We tested our method based on logits but no obvious correlation is exhibited. This may be because the logits of different models vary in significantly different ranges.

7 Conclusion

This paper studies an under-explored problem of assaying model generalization under distribution shift. To this end, we explore the use of Softmax prediction probability for developing OOD measures. We start by finding that maximum Softmax prediction probability is to some extent useful to measure the OOD performance. We then propose Softmax Correlation (SoftmaxCorr) which leverages class confusion encoded by class-class correlation matrix and thus better reflects the overall quality of the classifier predictions. To validate the usefulness of SoftmaxCorr, we compare with four other empirical measures across 19 datasets under ImageNet, CIFAR-10 and WILDS setups. We observe SoftmaxCorr generally presents more stable and higher correlation with model generalization on both ID and OOD datasets. This paper establishes some baseline usage of Softmax prediction probability and a specific improvement, and more investigation will be made in future.

REFERENCES

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pp. 66–75, 2021.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263, 2018.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, 2022.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, 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.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 2019.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, and Stefanie Jegelka. Measuring generalization with optimal transport. In *Advances in Neural Information Processing Systems*, pp. 8294–8306, 2021.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Ailin Deng, Shen Li, Miao Xiong, Zhirui Chen, and Bryan Hooi. Trust, but verify: Using self-supervised probing to improve trustworthiness. In *European Conference on Computer Vision*, 2022a.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Weijian Deng, Stephen Gould, and Liang Zheng. On the strong correlation between model invariance and generalization. In *Advances in Neural Information Processing Systems*, 2022b.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028, 2017.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16458–16468, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12387–12396, 2019.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*, pp. 2922–2932. PMLR, 2020.

Ustinova Ganin, Germain Ajakan, Laviolette Larochelle, Marchand, and Lempitsky. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*, 2016.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, pp. 3290–3300, 2020.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134–1144, 2021.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer, 2011.

Yiding Jiang, Pierre Foret, Scott Yak, Daniel M Roy, Hossein Mobahi, Gintare Karolina Dziugaite, Samy Bengio, Suriya Gunasekar, Isabelle Guyon, and Behnam Neyshabur. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020a.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020b.

Maurice George Kendall. *Rank correlation methods.* Griffin, 1948.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021.

Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, pp. 4, 2008.

Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images*. Citeseer, 2009.

Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning*, number 2, pp. 896, 2013.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039, 2020.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 888–896, 2019.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130, 2018.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792, 2021a.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 21464–21475, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante. Domain generalization via gradient surgery. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735, 2021.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, pp. 15682–15694, 2021.

Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing systems*, 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations*, 2017.

Fengchun Qiao and Xi Peng. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6790–6800, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.

Yair Schiff, Brian Quanz, Payel Das, and Pin-Yu Chen. Predicting deep neural network generalization with perturbation response curves. In *Advances in Neural Information Processing Systems*, pp. 21176–21188, 2021.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2012.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pp. 443–450, 2016.

Tao Sun, Cheng Lu, and Haibin Ling. Prior knowledge guided unsupervised domain adaptation. In *European Conference on Computer Vision*, 2022.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Advances in Neural Information Processing Systems*, pp. 19276–19289, 2020.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114, 2019.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, 1991.

Ramakrishna Vedantam, David Lopez-Paz, and David J Schwab. An empirical investigation of domain generalization with empirical risk minimizers. In *Advances in Neural Information Processing Systems*, pp. 28131–28143, 2021.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *arXiv preprint arXiv:2207.09239*, 2022.

Ross Wightman. Train cifar10 with pytorch. `https://github.com/kuangliu/pytorch-cifar`, 2017.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision*, 2019.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pp. 12133–12143, 2021.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2403–2412, 2018.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413, 2019.

Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *Advances in Neural Information Processing Systems*, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

A APPENDIX

A.1 EXPERIMENTAL SETUP

A.1.1 IMAGENET SETUP

(1) ImageNet Models

There are 173 models used in experiment. All of models are trained or fine-tuned on ImageNet-1k (Deng et al., 2009). The weights of the models are publicly avaiable via TIMM-1.5 (Wightman, 2019). For the choice of the models, we refer to the supplementary materials of Deng et al. (2022b); Miller et al. (2021); Baek et al. (2022). Moreover, we additionally include following models: { *'resnet10t', 'resnet14t', 'darknet53', 'cs3darknet_m', 'cs3darknet_focus_m', 'cs3darknet_l', 'cs3darknet_focus_l', 'regnety_040', 'regnety_064', 'regnety_080', 'regnetv_040', 'regnetv_064', 'regnetz_040', 'regnetz_040h'* }

(2) ImageNet Datasets

The datasets we use are publicly available standard benchmarks. We have double checked their license. We list their open-source as follows:

ImageNet-Validation (Deng et al., 2009) (https://www.image-net.org);
ImageNet-V2-A/B/C (Recht et al., 2019) (https://github.com/modestyachts/ImageNetV2);
ImageNet-Corruption (Hendrycks & Dietterich, 2019) (https://github.com/hendrycks/robustness);
ImageNet-Sketch (Wang et al., 2019) (https://github.com/HaohanWang/ImageNet-Sketch);
ImageNet-Rendition (Hendrycks et al., 2021) (https://github.com/hendrycks/imagenet-r);
ObjectNet (Barbu et al., 2019) (https://objectnet.dev);
Stylized-ImageNet (Geirhos et al., 2019) (https://github.com/rgeirhos/Stylized-ImageNet).

A.1.2 CIFAR-10 SETUP

(1) CIFAR-10 Models

Follow the practice in (Miller et al., 2021), we train 65 CIFAR models using the implementations from https: //github.com/kuangliu/pytorch-cifar. The models span a range of manually designed architectures and the results of automated architecture searches. Specifically, we use: {*'DenseNet-121/169/201/161/201', 'Densenet-cifar', 'DLA', 'DPN26/92', 'EfficientNetB0', 'GoogLeNet', 'LeNet', 'MobileNet', 'MobileNetV2', 'PNASNetA/B', 'PreActResNet18/34/50/101/152', 'RegNetX-200MF/400MF', 'RegNetY-400MF', 'ResNet-18/34/50/101/152', 'ResNeXt29-8x64d/32x4d/4x64d/2x64d', 'SENet18', 'ShuffleNetV2', 'ShuffleNetG2/G3', 'SimpleDLA', 'VGG-11/13/16/19'*}.

Furthermore, we use the trained models publicly provided by https://github.com/chenyaofo/pytorch-cifar-models. They are {*'cifar10-mobilenetv2-x0-5', 'cifar10-mobilenetv2-x0-75', 'cifar10-mobilenetv2-x1-0', 'cifar10-mobilenetv2-x1-4', 'cifar10-repvgg-a0', 'cifar10-repvgg-a1', 'cifar10-repvgg-a2', 'cifar10-resnet20', 'cifar10-resnet32', 'cifar10-resnet44', 'cifar10-resnet56', 'cifar10-shufflenetv2-x0-5', 'cifar10-shufflenetv2-x1-0', 'cifar10-shufflenetv2-x1-5', 'cifar10-shufflenetv2-x2-0', 'cifar10-vgg11-bn', 'cifar10-vgg13-bn', 'cifar10-vgg16-bn', 'cifar10-vgg19-bn'*} .

(2) CIFAR-10 Datasets

The datasets are publicly available benchmarks. Their source are as follows:

CIFAR-10 (Krizhevsky et al., 2009) (https://www.cs.toronto.edu/ kriz/cifar.html);
CIFAR-10-C (Hendrycks & Dietterich, 2019) (https://github.com/hendrycks/robustness);
CIFAR-10.1 (Recht et al., 2018) (https://github.com/modestyachts/CIFAR-10.1);
CINIC (Darlow et al., 2018) (https://github.com/BayesWatch/cinic-10).

A.1.3 WILDS SETUP

(1) iWildCam Models

We use the publicly available code (https://github.com/p-lambda/wilds) to finetune 66 models that are pretrained on ImageNet. Following the practice in (Miller et al., 2021), model variation results

from model architecture and learning rate. We train for 12 epochs with batch size 16 using Adam and sweep over learning rate in the grid $\{1e^{-3}, 1e^{-2}\}$. The other Adam parameters were set to the Pytorch defaults. The models are: { *'alexnet', 'mobilenet_v2', 'resnet18', 'resnet34', 'resnet50', 'vgg11', 'densenet121', efficientne_b0, efficientnet_b1, 'regnet_y_400mf', 'regnet_y_800mf', 'shufflenet_v2_x0_5' 'shufflenet_v2_x1_0'* }.

(2) Camelyon17 Models

We use the publicly available code (https://github.com/p-lambda/wilds) to finetune 45 models that are pretrained on ImageNet. Models vary in model architecture and random seed. The model architectures are: { *'alexnet', 'vgg11', 'densenet121', 'resnet18', 'resnet34', 'resnet50', 'resnet101', 'resnext50_32x4d' 'wide_resnet50_2' 'mobilenet_v3_large', 'regnet_y_1_6gf', 'shufflenet_v2_x0_5', 'shufflenet_v2_x1_5', 'shufflenet_v2_x2_0', 'regnet_y_400mf', 'regnet_y_800mf',* }.

(3) WILDS Datasets

iWildCam (Beery et al., 2021) https://www.kaggle.com/c/iwildcam-2020-fgvc7;
Camelyon17 (Bandi et al., 2018) https://camelyon17.grand-challenge.org/

A.2 DomainNet

Domain adaptation algorithms. We use publicly accessible weights provided in: https://worksheets.codalab.org/worksheets/0x52cea64d1d3f4fa89de326b4e31aa50a. These models are trained by following optimization algorithms: ERM (Vapnik, 1991), DeepCORAL (Sun & Saenko, 2016), DANN (Ganin et al., 2016), AFN (Xu et al., 2019), PseudoLabel (Lee et al., 2013), NoisyStudent (Xie et al., 2020), FixMatch (Sohn et al., 2020), SwAV (Caron et al., 2020).

DomainNet Dataset (Peng et al., 2019) http://ai.bu.edu/M3SDA/.

A.3 Computation Resources

All experiment is run on one 2080Ti and the CPU AMD Ryzen Threadripper 2950X 16-Core Processor. We use PyTorch (1.10.2+cu102) for all the experiments. .

A.4 RESULTS OF SOFTMAX CORRELATION

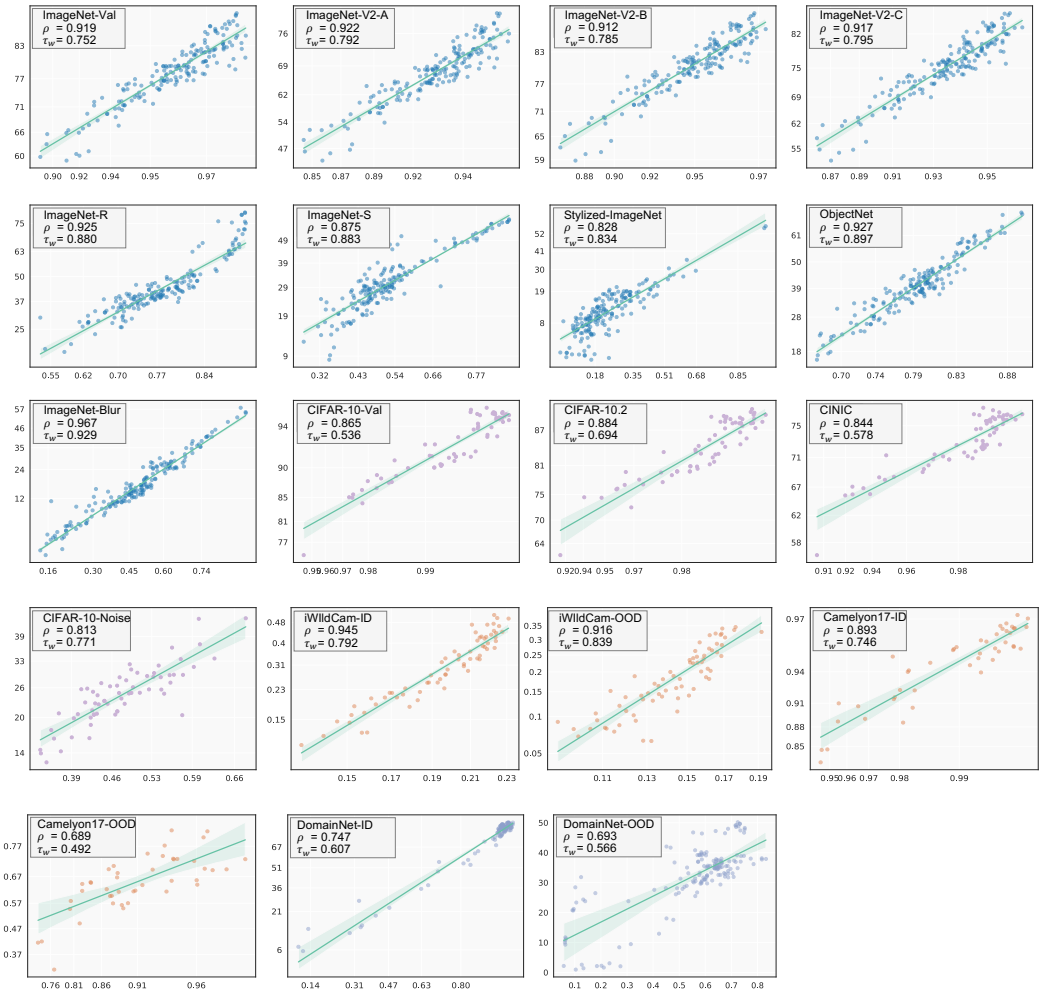

Figure 7: SoftmaxCorr *vs.* model generalization under ImageNet, CIFAR-10 and WILDS setups. The y-axis is accuracy, accuracy and macro-F1 for three setups, respectively. The straight lines are fit with robust linear regression (Huber, 2011). Axes are probit scaled as described in Section 3.

A.5 RESULTS OF MAXIMUM PREDICTION PROBABILITY

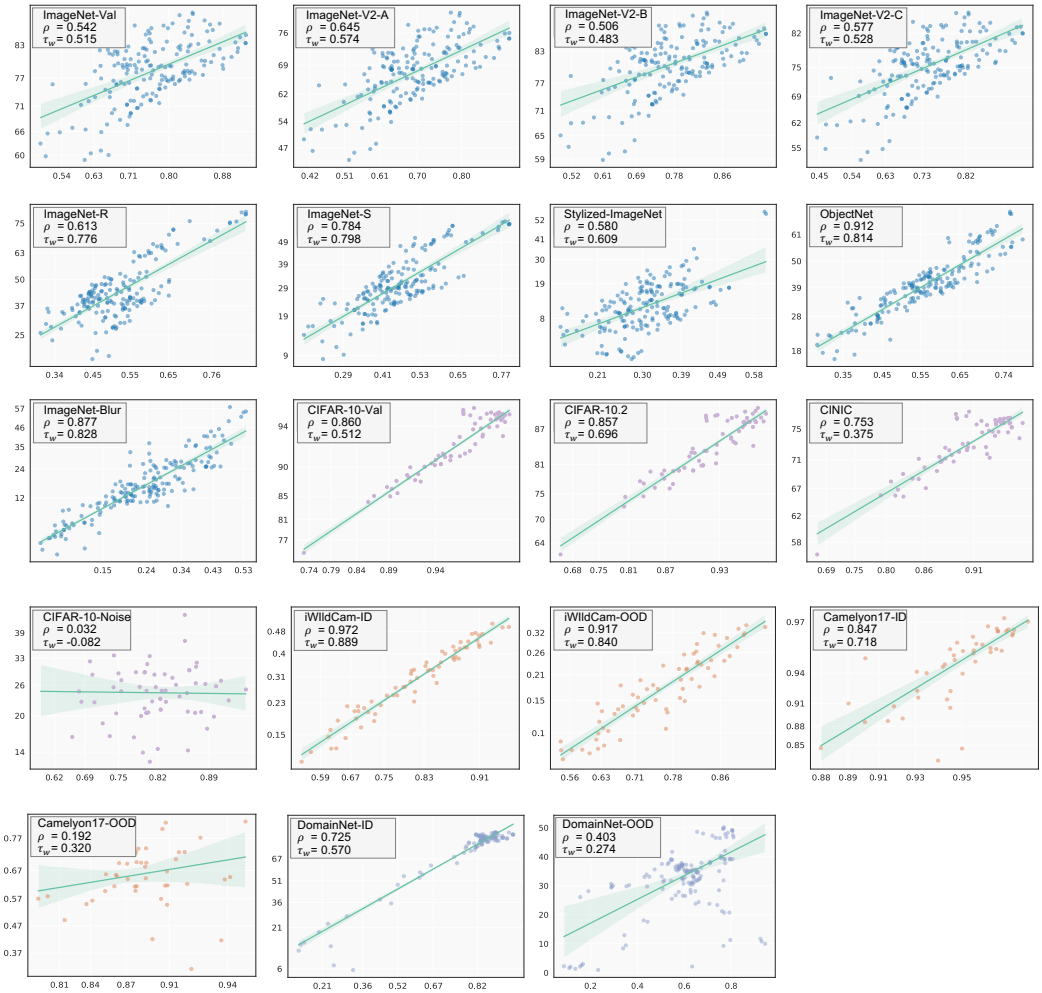

Figure 8: MaxPred *vs.* model generalization under ImageNet, CIFAR-10 and WILDS setups. The y-axis is accuracy, accuracy and macro-F1 for three setups, respectively. The straight lines are fit with robust linear regression (Huber, 2011). Axes are probit scaled as described in Section 3.

A.6 RESULTS OF SOFTMAX GAP

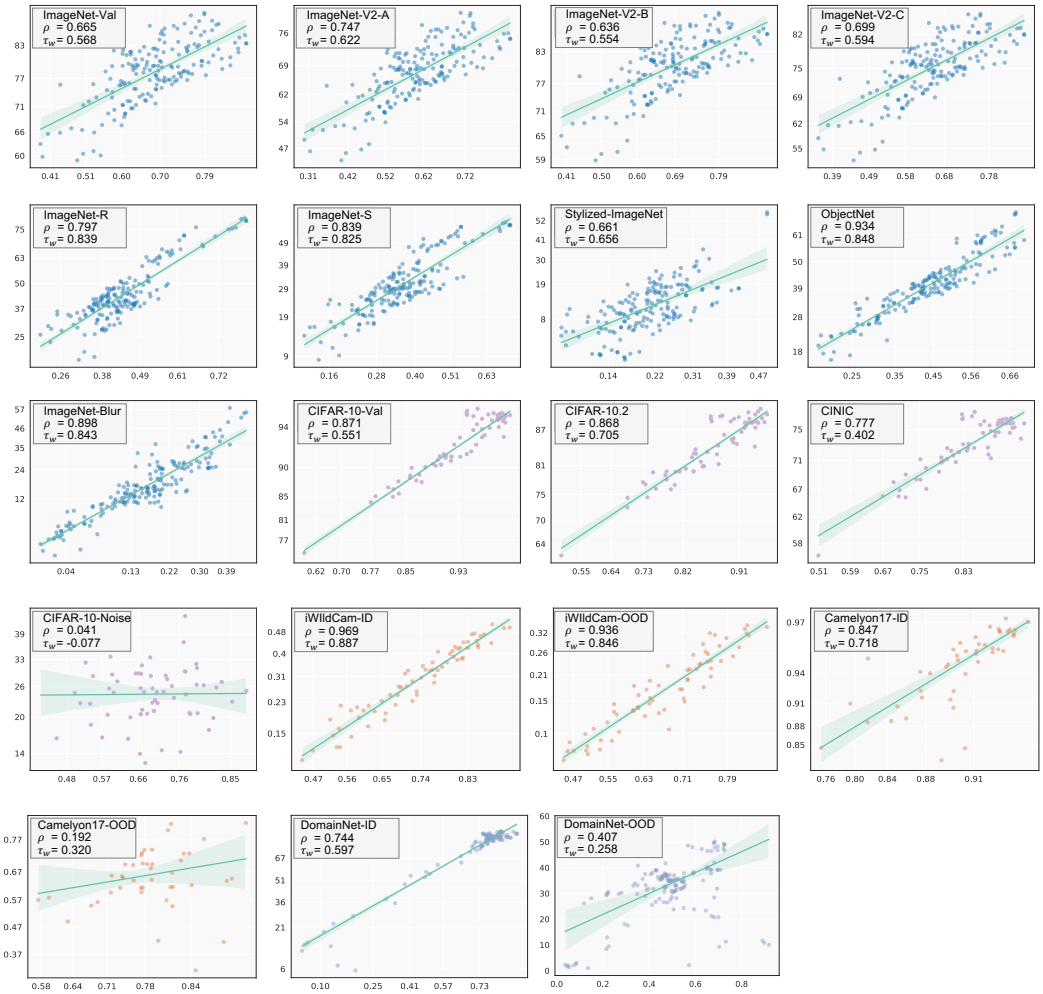

Figure 9: SoftGap *vs.* model generalization under ImageNet, CIFAR-10 and WILDS setups. The y-axis is accuracy, accuracy and macro-F1 for three setups, respectively. The straight lines are fit with robust linear regression (Huber, 2011). Axes are probit scaled as described in Section 3.

A.7 RESULTS OF NEGATIVE PREDICTION ENTROPY

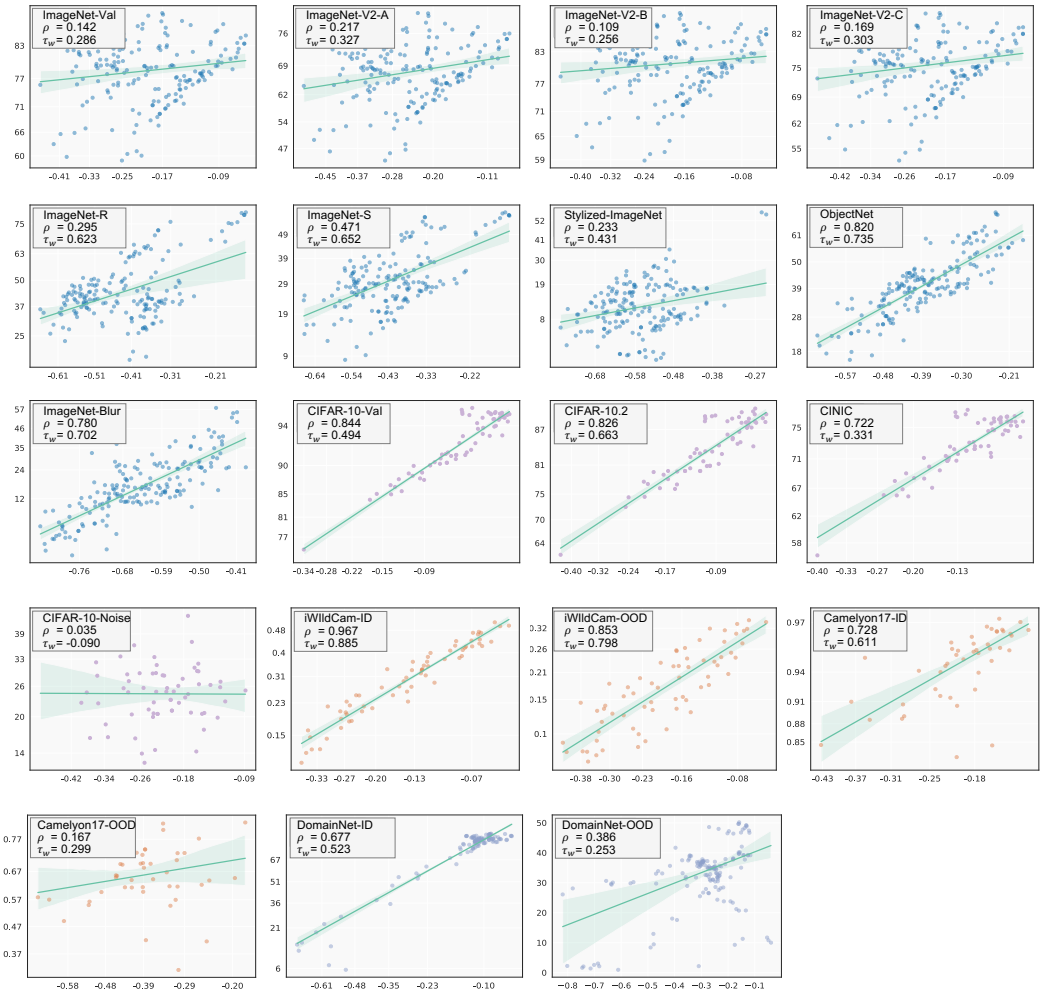

Figure 10: Entropy *vs.* model generalization under ImageNet, CIFAR-10 and WILDS setups. The y-axis is accuracy, accuracy and macro-F1 for three setups, respectively. The straight lines are fit with robust linear regression (Huber, 2011). Axes are probit scaled as described in Section 3.

A.8 RESULTS OF INFORMATION MAXIMIZATION

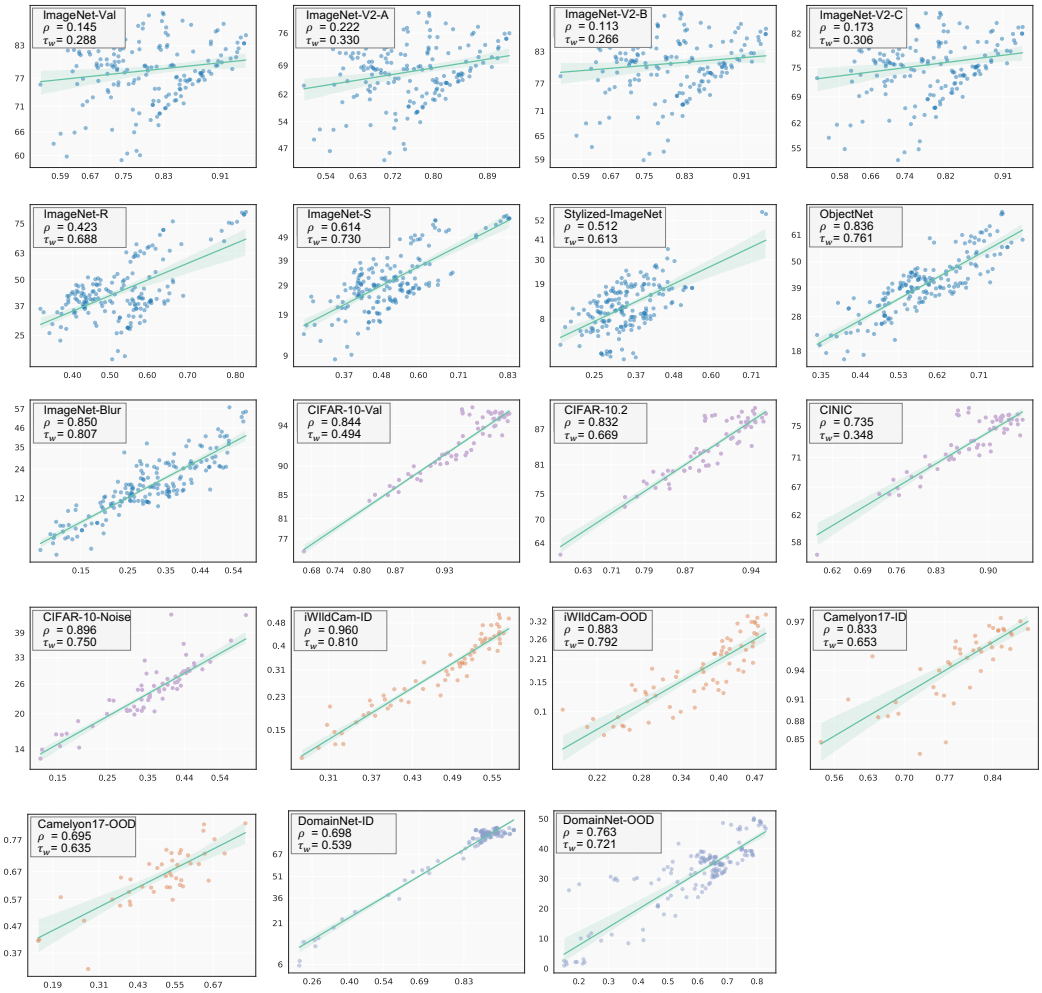

Figure 11: InfoMax *vs.* model generalization under ImageNet, CIFAR-10 and WILDS setups. The y-axis is accuracy, accuracy and macro-F1 for three setups, respectively. The straight lines are fit with robust linear regression (Huber, 2011). Axes are probit scaled as described in Section 3.