Beyond Top-Class Agreement: Using Divergences to Forecast Performance under Distribution Shift

Mona Schirmer
UvA-Bosch Delta Lab
University of Amsterdam

Dan Zhang
Bosch Center for AI &
University of Tübingen

Eric Nalisnick
UvA-Bosch Delta Lab
University of Amsterdam

Abstract

Knowing if a model will generalize to data 'in the wild' is crucial for safe deployment. To this end, we study model disagreement notions that consider the full predictive distribution - specifically disagreement based on Hellinger distance, Jensen-Shannon and Kullback–Leibler divergence. We find that divergence-based scores provide better test error estimates and detection rates on out-of-distribution data compared to their top-1 counterparts. Experiments involve standard vision and foundation models.

1 Introduction

When deployed to the wild, machine learning systems often encounter conditions unlike those on which the system was trained. These deviations from the training distribution pose significant risks in safety-critical applications, such as autonomous driving, since performance can suddenly and precipitously degrade. It is therefore essential to develop techniques that can anticipate system failures caused by operating under distribution shift. However, estimating test-time performance is difficult in even in-distribution (ID) settings. Out-of-distribution (OOD) scenarios are especially challenging since a labeled OOD validation set is often impossible to obtain.

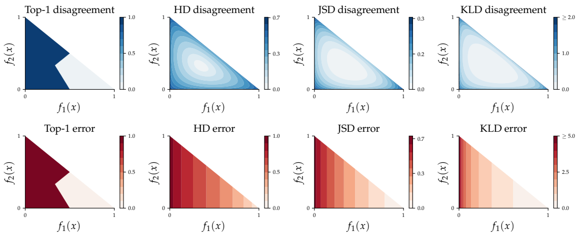

Figure 1: Disagreement and error notions for 3-class classification: x,y-axis represent predicted probability of model f for class $1, 2$. First row: disagreement between f, f' with fixed $f'(x) = (0.35, 0.325, 0.325)$. Second row: error for $y = 1$.

To help with this problem, recent studies [19, 10, 1, 11, 22] have identified strong empirical correlations and theoretical connections between test error and model disagreement. [19, 10] found that model disagreement rates serve as an effective approximation of the test error and [10, 11] show that the two are equivalent under calibration. Moreover, [1] found that ID vs. OOD disagreement correlates linearly and the slope and bias match those of ID vs. OOD test error. These discoveries motivate the use of model disagreement for unlabeled OOD performance estimation.

Previous approaches to estimate a model's OOD accuracy are often based on either the model's uncertainties [9, 5, 14] or top-1 disagreement with other models [10, 1, 22]. In this paper, we ask a natural next question; can we combine the best of both worlds by using model disagreement notions that consider the full predictive label distribution? Such distributional disagreement notions have already found applications in e.g. active learning [20, 17, 4] and uncertainty quantification [16]. However, its role for performance estimation has been unexplored. We argue that the use of the

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

predictive distribution provides a more nuanced view of model disagreement as visualized in Fig. 1. By investigating three classic divergences, we show their value in OOD detection and error estimation.

2 Preliminaries

Problem formulation We consider a classification task with a set of classifiers $\mathcal{F}$ where $f \in \mathcal{F}$ maps from the feature space $\mathcal{X} \subseteq \mathbb{R}^D$ to the probabilities over K classes, i.e. $f : \mathcal{X} \to [0,1]^K$. The classifiers have been trained on a training set $\mathcal{S}_{ID} = \{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = [K]$ sampled from the in-distribution $\mathcal{D}_{ID}$. In this paper, we ask the question what is a good measure of model disagreement $\texttt{dis}$ that can help to infer the model performance on an unseen test set $\mathcal{T}_{OOD} = \{(x_i, y_i)\}_{i=1}^{N'}$ sampled from a shifted distribution $\mathcal{D}_{OOD}$. To have a consistent translation from error to disagreement for each notion, we denote as error the respective disagreement between the model and the one-hot encoded class label, $\tilde{y}$. The disagreement of two classifiers f, f' as well as error for a data distribution $\mathcal{D}$ is then given by

$$\texttt{Dis}_{\mathcal{D}}(f, f') = \mathbb{E}_{x,y\sim\mathcal{D}}[\texttt{dis}(f(x), f'(x))], \quad \text{and} \quad \texttt{Err}_{\mathcal{D}}(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[\texttt{dis}(\tilde{y}, f(x))].$$

We can estimate $\texttt{Dis}_{\mathcal{D}}$ and $\texttt{Err}_{\mathcal{D}}$ as the empirical average over samples from $\mathcal{D}$. Note that while test error requires access to labels, disagreement can be computed using only unlabeled samples.

Previous approaches Past work on the link between agreement and generalization [10, 11, 1, 22] has focused on top-1 disagreement to predict test error rates. It states the disagreement rate of two classifiers on the predicted class label. Let $h : [0,1]^K \to [K]$ denote the function that converts the output probabilities to the predicted class label, i.e. $h(f(x)) = \operatorname{argmax}_{k\in[K]} f_k(x)$ where $f_k(x)$ denotes the predicted probability for class k. Then, top-1 disagreement and test error is defined by

$$\texttt{dis}^{Top1}(f(x), f'(x)) = \mathbb{1}\{h(f(x)) \neq h(f'(x))\} \quad \text{and} \quad \texttt{Err}_{\mathcal{D}}^{Top1}(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[\mathbb{1}\{h(f(x)) \neq y\}].$$

3 Divergence Disagreement

In this work, we investigate divergence-based disagreement notions arguing that the previous top-1 notion is too coarse to differentiate between in fact different disagreement scenarios. Consider this motivating example: Three multi-class classifiers A, B, and C predict class k for a given sample. Model A and B predict class k with probability 0.99 and model C with 0.2. Under top-1 disagreement, all models would be considered in agreement despite the clear alignment difference between models A and B versus model C. By considering a model's predictive distribution, we can define disagreement notions that differentiate between *discrepancies in uncertainties* across models. We examine disagreement notions based on three commonly used divergences: Hellinger distance, Jensen-Shannon and Kullback-Leibler divergence.

Hellinger distance (HD) disagreement The Hellinger distance is symmetric and satisfies the triangle inequality rendering it a true metric on the space of probability distributions. Notably, it lies in the interval $[0,1]$ making it straight forward for comparison. HD disagreement and test error with respect to the true class label y are given by

$$\texttt{dis}^{HD}(f(x), f'(x)) = \frac{1}{\sqrt{2}}\sqrt{\sum_k \left(\sqrt{f_k(x)} - \sqrt{f'_k(x)}\right)^2}$$

$$\texttt{Err}_{\mathcal{D}}^{HD}(f) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\frac{1}{\sqrt{2}}\sqrt{\left(\sqrt{f_y(x)} - 1\right)^2 + \sum_{k\neq y} f_k(x)}\right]$$

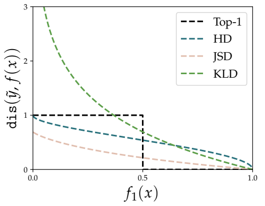

Figure 2: Error notions for binary classification for $y = 1$

Jensen–Shannon divergence (JSD) disagreement JSD is a symmetric divergence based on the Kullback-Leibler divergence (KLD). It averages the KLD of their arguments from their uniform mixture and is bounded by $\log 2$. The JSD disagreement notion is given by

$$\texttt{dis}^{JSD}(f(x), f'(x)) = \frac{1}{2}\Big(\sum_k f_k(x)\log\frac{f_k(x)}{\overline{f_k}(x)} + \sum_k f'_k(x)\log\frac{f'_k(x)}{\overline{f_k}(x)}\Big).$$

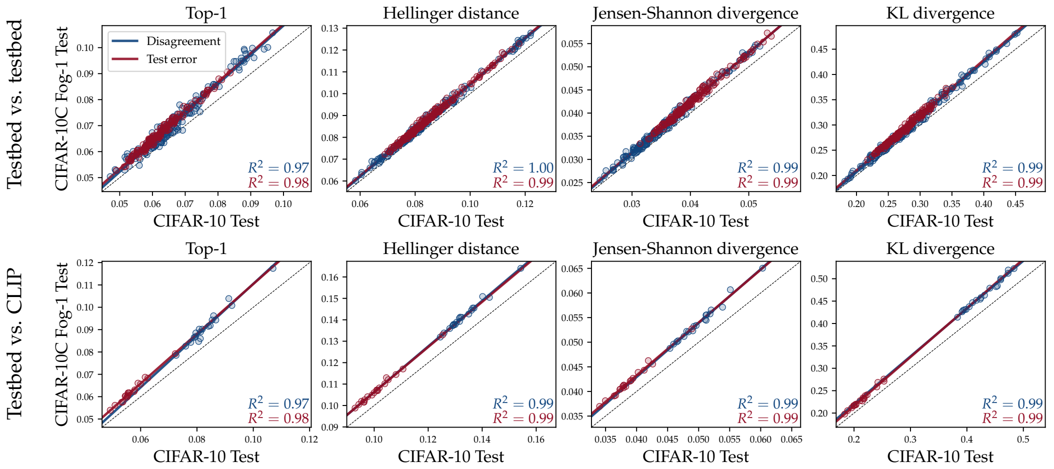

Figure 3: *Agreement-on-the-line* and *Accuracy-on-the-line* for CIFAR-10C fog: Distributional disagreement notions (columns 2-4) correlate stronger than top-1 disagreement (first column).

with $\overline{f_k}(x) = \frac{1}{2}(f_k(x) + f'_k(x))$ denoting the uniform mixture. The JSD error can be formulated as

$$\mathtt{Err}^{JSD}_{\mathcal{D}}(f) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\frac{1}{2}\Big(\log\frac{2}{1+f_y(x)} + f_y(x)\log\frac{2f_y(x)}{1+f_y(x)} + \sum_{k\neq y} f_k(x)\log 2\Big)\right].$$

KLD disagreement Unlike HD and JSD, the standard KLD is non-symmetric and unbounded. We consider a simple symmetrized version of the KLD previously employed in active learning [4] which averages over forward and reverse KLD,

$$\mathtt{dis}^{KLD}(f(x), f'(x)) = \frac{1}{2}\Big(\sum_k f_k(x)\log\frac{f_k(x)}{f'_k(x)} + \sum_k f'_k(x)\log\frac{f'_k(x)}{f_k(x)}\Big).$$

The KLD error notion takes the one-hot-encoded target label as the first argument simplifying to $\mathtt{Err}^{KLD}_{\mathcal{D}}(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[-\log f_y(x)]$. Interestingly, this error notion only considers the softmax output of the ground truth class. In this sense, $\mathtt{Err}^{KLD}_{\mathcal{D}}$ can be seen as a 'soft' version of $\mathtt{Err}^{Top1}_{\mathcal{D}}$ that incorporates uncertainty. Fig. 2 visualizes error notions. In comparison to HD and JSD, KLD has a steeper penalization for high uncertainties of the target class.

4 Experimental study

We conduct experiments on CIFAR-10 [12] and CIFAR-100 [13].[1] For the shift datasets, we use CIFAR-10C and CIFAR-100C [8] containing 18 types of synthetic corruptions in 5 severity levels. We employ 19 deep vision models pre-trained by [2] spanning a variety of vision models such as ResNet [6] and VGG [24]. In the testbed vs. testbed (TvT) scenario, we compute disagreement between all model pairs resulting in 171 disagreement instances.

The testbed vs. CLIP (TvC) setting measures disagreement of each vision model against the foundation model CLIP [21] finetuned on ID data. Here, CLIP serves as an anchor with superior generalisation abilities [21] possibly exposed to samples from $\mathcal{D}_{OOD}$ already during training. Model details are listed in Appendix A.

Does divergence disagreement pick up stronger on-the-line phenomena? We first inspect if the model's uncertainties incorporated in divergence disagreement and test error notions result in a stronger correlation between (i) ID vs. OOD disagreement (described as *agreement-on-the-line* by [1]) and (ii) ID vs. OOD test error (dubbed *accuracy-on-the-line* by [18]). We compute pairwise disagreement and test error for the ID test sets of CIFAR-10/100 and the corrupted OOD test sets of CIFAR-10C/100C and plot ID vs. OOD values for each model pair.

[1] Code made available at `https://github.com/monasch/divdis`

Fig. 3 shows results exemplary for the fog corruption of CIFAR-10C. We observe stronger correlations for divergence notions compared to the standard top-1 approach for both on-the-line phenomena in the setting with and without foundation model. As noticed by [1], the disagreement and test error line fit coincide. A stronger correlation is of interest because it could facilitate a more accurate line fit for estimating OOD test error, which will be our next point of discussion.

Can divergence disagreement provide better OOD performance estimation? To assess the potential of divergence disagreement for unlabeled OOD performance estimation, we employ a simple disagreement-based linear regression procedure, ALineD [1], and compare results across disagreement notions. AlineD exploits the observation that the slope and bias from the ID vs. OOD agreement matches the slope and bias from ID vs. OOD accuracy (see Fig. 3). Fitting regression coefficients to the disagreement relation and extrapolating with the ID test error thus gives an estimation for OOD test error without the need for labeled OOD data. Since the core requirement of the method is a correlation between ID and OOD agreement in the first place, we focus on CIFAR corruptions that exhibit such correlations ($R^2 > 0.95$) for all disagreement notions.

	OOD dataset	Top-1	HD	JSD	KLD
CIFAR-10C	brightness1	1.14	**0.39**	0.76	0.83
	brightness2	1.30	**0.47**	0.99	1.03
	brightness3	1.80	**0.81**	1.23	1.57
	brightness4	2.11	**1.26**	1.70	1.71
	contrast1	2.69	**1.49**	1.70	3.77
	defocus_blur1	1.63	**0.58**	0.91	1.31
	fog1	2.26	**0.70**	0.82	1.40
	gaussian_blur1	1.50	**0.58**	0.90	1.31
	saturate1	3.01	**2.90**	4.80	6.77
	saturate3	2.61	**1.22**	2.12	2.91
CIFAR-100C	brightness 1	0.63	**0.17**	0.23	0.39
	brightness 2	0.73	**0.41**	0.43	0.71
	brightness 3	**0.80**	1.06	0.93	1.37
	brightness 4	**0.94**	2.10	1.93	2.82
	contrast 1	1.28	**1.25**	1.61	2.51
	defocus blur 1	0.74	**0.46**	0.63	0.91
	defocus blur 2	**1.47**	2.36	2.59	4.42
	fog 1	0.86	**0.50**	0.53	0.88
	fog 2	**1.29**	2.01	1.53	3.65
	gaussian blur 1	0.80	**0.45**	0.63	0.91
	saturate 3	1.49	**1.28**	1.34	3.03
	saturate 4	**2.37**	4.70	4.85	10.36

Table 1: MAPE (↓) for OOD performance estimation: HD performs best overall, but is less performative under more labels.

Tab. 1 displays mean absolut percentage error (MAPE) of OOD performance estimation in the TvT setting for datasets that satisfy the R^2 threshold. We remark three main trends. Firstly, Hellinger disagreement provides the best OOD test error estimates on high correlation datasets reporting the smallest mean absolut percentage error (MAPE) in 17 out of 22 shifts. Secondly, HD appears to be slightly less robust on the more labels containing CIFAR-100C: Interestingly, for the same corruption (brightness 3+4) HD disagreement performs best on CIFAR-10C but top-1 performs better on CIFAR-100C. This is indeed surprising, as one might expect divergence disagreement to prove particularly advantageous on datasets with many labels. Thirdly, top-1 seems more robust on shifts with weak ID vs OOD disagreement: Considering all corruptions regardless of R^2, the median MAPE for top-1 (8.05) is lower than for HD (10.15), JSD (16.61) and KLD (19.62) on CIFAR-10C. We present results for the TvC scenario in Appendix B.

How does miscalibration affect ID vs. OOD correlations? Next, we compare the impact of miscalibration on OOD performance estimation across disagreement notions. For that, we ask if the disagreement and test error correlation on ID vs. OOD data is more prone to miscalibration for divergence disagreement. This hypothesis sounds plausible since such notions rely heavily on the predictive distribution and could potentially explain poorer median estimates than top-1 (third observation from above). To assess this, we plot the class-aggregated calibration error (CACE) [10] averaged over all models in the testbed against the R^2 of the on-the-line phenomena for all corrupted datasets. We fit a three-degree polynomial per notion to highlight trends.

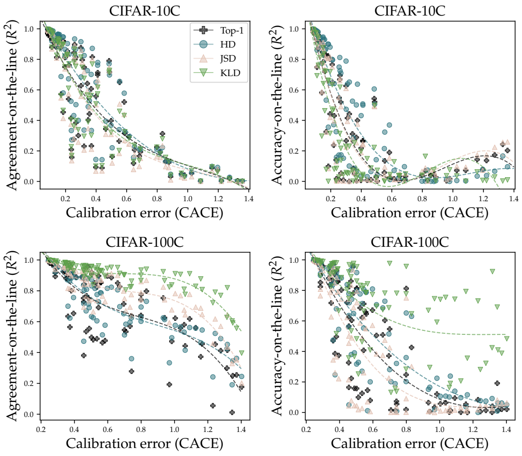

Figure 4: Calibration vs. agreement- and accuracy-on-the-line correlation (R^2): Miscalibration reduces correlation.

Fig. 4 shows results. We highlight two main observations. First, the on-the-line phenomena vanish under increasing calibration error for all disagreement notions. This is an important insight since the question of when the phenomena are valid has been largely unexplained so far. Second, perhaps surprisingly, top-1 correlation is not more robust than divergence-based correlations. On CIFAR-100, we observe however, that Hellinger disagreement correlation drops faster with increasing calibration error (potentially explaining the second observation from above).

Can divergence disagreement detect OOD samples better? By comparing disagreement notions through the lens of OOD detection, we like to provide a different point of comparison. Unlike in OOD performance estimation, disagreement in OOD detection does not require a solid link to test error. Instead, it assumes distribution shifts affect models in random ways which lead them to disagree. We hypothesis that these perturbations affect all output dimensions equally and can thus be better captured by divergence disagreement. We use CIFAR-10 and CIFAR-100 test sets as ID samples and each "Hendrycks-corruptions" set as the OOD samples in separate OOD detection tasks. Maximum softmax probability (MSP) [9] and maximum logit (MaxLogit) [7] serve as baseline OOD scores.

Tab. 2 seems to confirm our hypothesis. It reports ROC-AUC for separating OOD from ID samples. Divergence disagreement performs best capturing information about the shift better than top-1, MSP and MaxLogit. Interestingly, KLD performs best on this task, but worst on OOD error estimation indicating it picks up information about the distribution shift that could not be exploited for error estimation.

	Severity	−MaxLogit	−MSP	Top-1	HD	JSD	KLD
CIFAR-10C	1	60.10	60.28	58.15	**60.47**	60.33	60.46
	2	66.85	67.32	64.01	67.66	67.57	**67.67**
	3	71.48	72.08	68.25	**72.46**	72.34	72.44
	4	76.46	77.06	73.31	**77.59**	77.42	77.54
	5	82.60	82.89	79.47	**83.52**	83.26	83.34
CIFAR-100C	1	61.09	60.74	60.24	61.49	61.45	**61.80**
	2	67.59	66.97	66.46	68.22	68.31	**68.59**
	3	70.44	69.88	69.36	71.30	71.40	**71.66**
	4	73.87	73.21	72.75	74.92	75.06	**75.37**
	5	78.15	77.50	77.22	79.59	79.73	**80.10**

Table 2: ROC-AUC score for OOD detection in the TvT scenario: Scores per severity level are averaged over all corruption types. HD and KLD disagreement detect OOD samples best.

5 Conclusion

By moving beyond the traditional top-class focus, we investigated more fine-grained notions of model disagreement that consider the full predictive label distribution. We show that contrasting models in a more nuanced way unlocks great potential for detecting system failure under distribution shift even more accurately. Future work may investigate the link to calibration more closely and sketch how model over- and under confidence affects disagreement.

Acknowledgments

We thank Metod Jazbec, Alexander Timans and Putra Manggala for helpful feedback on the draft. We are also grateful for the constructive comments from the anonymous reviewers that helped to improve this work. MS and EN are generously supported by the Bosch Center for Artificial Intelligence.

References

[1] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022. 1, 2, 3, 4, 9

[2] Yaofo Chen. chenyaofo/pytorch-cifar-models, November 2021. 3, 8

[3] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 8

[4] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3723–3731, 2019. 1, 3

[5] Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 8

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 5

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 5

[10] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021. 1, 2, 4

[11] Andreas Kirsch and Yarin Gal. A note on" assessing generalization of sgd via disagreement". *arXiv preprint arXiv:2202.01851*, 2022. 1, 2

[12] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 3

[13] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 3

[14] Yuzhe Lu, Zhenlin Wang, Runtian Zhai, Soheil Kolouri, Joseph Campbell, and Katia Sycara. Predicting out-of-distribution error with confidence optimal transport. *arXiv preprint arXiv:2302.05018*, 2023. 1

[15] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 8

[16] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020. 1

[17] Prem Melville, Stewart M Yang, Maytal Saar-Tsechansky, and Raymond Mooney. Active learning for probability estimation using jensen-shannon divergence. In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*, pages 268–279. Springer, 2005. 1

[18] 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*, pages 7721–7735. PMLR, 2021. 3

[19] Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020. 1

[20] K Nigam. Employing em in pool-based active learning for text classification. machine learning. In *Proceeding (s) of the Fifteenth International Conference (ICML'98)*, pages 350–358, 1998. 1

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 8

[22] Elan Rosenfeld and Saurabh Garg. (almost) provable error bounds under distribution shift via disagreement discrepancy. *arXiv preprint arXiv:2306.00312*, 2023. 1, 2

[23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8

A Testbed

For the standard vision models, we use models pre-trained on CIFAR-10/CIFAR-100 by [2]. As foundation model, we finetune CLIP [21] on CIFAR ID data, which yield an accuracy increase on ID data from 88.8% to 94.8% on CIFAR-10 and from 61.7% to 79.3% on CIFAR-100 compared to the zeroshot version. We use a simple finetune procedure by adding a linear classification head to the image features outputted by CLIP. The classification head is trained for 10 epochs with a learning rate of 1×10^{-3}. Tabs. 3 and 4 summarize models, their test accuracy on ID data and parameter size.

Model	Top-1 accuracy (%)	Parameters (M)
ResNet20 [6]	92.60	0.27
ResNet32 [6]	93.53	0.47
ResNet44 [6]	94.01	0.66
ResNet56 [6]	94.37	0.86
VGG11 [24]	92.79	9.76
VGG13 [24]	94.00	9.94
VGG16 [24]	94.16	15.25
VGG19 [24]	93.91	20.57
MobileNetV2 (x0_5) [23]	92.88	0.70
MobileNetV2 (x0_75) [23]	93.72	1.37
MobileNetV2 (x1_0) [23]	93.79	2.24
MobileNetV2 (x1_4) [23]	94.22	4.33
ShuffleNetV2 (x0_5) [15]	90.13	0.35
ShuffleNetV2 (x1_0) [15]	92.98	1.26
ShuffleNetV2 (x1_5) [15]	93.55	2.49
ShuffleNetV2 (x2_0) [15]	93.81	5.37
RepVGG (a0) [3]	94.39	7.84
RepVGG (a1) [3]	94.89	12.82
RepVGG (a2) [3]	94.98	26.82
CLIP ViT-B/32 [21] (finetuned)	94.82	151.28

Table 3: Testbed for CIFAR-10

Model	Top-1 accuracy (%)	Parameters (M)
ResNet20 [6]	68.83	0.28
ResNet32 [6]	70.16	0.47
ResNet44 [6]	71.63	0.67
ResNet56 [6]	72.63	0.86
VGG11 [24]	70.78	9.80
VGG13 [24]	74.63	9.99
VGG16 [24]	74.00	15.30
VGG19 [24]	73.87	20.61
MobileNetV2 (x0_5) [23]	70.88	0.82
MobileNetV2 (x0_75) [23]	73.61	1.48
MobileNetV2 (x1_0) [23]	74.20	2.35
MobileNetV2 (x1_4) [23]	75.98	4.50
ShuffleNetV2 (x0_5) [15]	67.82	0.44
ShuffleNetV2 (x1_0) [15]	72.39	1.36
ShuffleNetV2 (x1_5) [15]	73.91	2.58
ShuffleNetV2 (x2_0) [15]	75.35	5.55
RepVGG (a0) [3]	75.22	7.96
RepVGG (a1) [3]	76.12	12.94
RepVGG (a2) [3]	77.18	26.94
CLIP ViT-B/32 [21] (finetuned)	79.28	151.28

Table 4: Testbed for CIFAR-100

B OOD performance estimation with CLIP

We evaluate performance estimation for both the TvT and TvC setting. In the TvC scenario, we are interested in estimating performance of the standard vision models based on disagreement of each standard vision model with the finetuned CLIP model. We found the simpler line fitting method, ALine-S [1], to work better than ALine-D [1] in this setting. ALine-D takes the disagreement of the model pair of interest directly into account (instead of indirectly through the fitted line coefficients, see [1] for details). However, this yields to poor estimates when disagreement and error are far apart as observed in Fig. 3.

Tab. 5 reports results for datasets where agreement-on-the-line holds ($R^2 > 0.95$). Interestingly, in this setting, JSD and top-1 disagreement perform equally well and best on 7 shift datasets.

	OOD dataset	Top-1	HD	JSD	KLD
CIFAR-10C	brightness 1	1.30	**0.47**	0.62	0.53
	brightness 3	**2.48**	3.86	3.92	3.87
	brightness 4	**4.62**	9.17	6.96	7.92
	contrast 1	5.07	5.09	**3.17**	4.03
	defocus blur 1	**1.95**	6.44	3.67	4.65
	fog1	1.97	1.96	**1.76**	2.98
	gaussian blur 1	**2.16**	6.68	3.76	4.83
	saturate 3	2.60	1.64	**1.62**	2.19
CIFAR-100C	brightness 1	0.63	**0.22**	0.30	0.32
	brightness 2	0.99	0.59	**0.56**	1.01
	brightness 3	**1.20**	1.54	0.96	3.05
	contrast 1	1.36	1.14	**0.97**	2.62
	defocus blur 1	1.10	0.59	**0.40**	0.70
	fog1	1.71	0.73	0.83	**0.67**
	gaussian blur 1	1.24	0.55	**0.40**	0.74
	saturate 3	**1.08**	2.17	1.31	4.59
	saturate 4	**1.29**	7.98	5.38	15.44

Table 5: MAPE (↓) for OOD performance estimation of vision models using disagreement against a finetuned CLIP model (TvC setting).