CALIBRATED ENSEMBLES - A SIMPLE WAY TO MITIGATE ID-OOD ACCURACY TRADEOFFS

Anonymous authors
Paper under double-blind review

ABSTRACT

We often see undesirable tradeoffs in robust machine learning where out-of-distribution (OOD) accuracy is at odds with in-distribution (ID) accuracy. A "robust" classifier obtained via specialized techniques like removing spurious features has better OOD but worse ID accuracy compared to a "standard" classifier trained via vanilla ERM. On six distribution shift datasets, we find that simply ensembling a standard and a robust model is a strong baseline—we match the ID accuracy of a standard model with only a small drop in OOD accuracy compared to the robust model. However, calibrating these models in-domain surprisingly improves the OOD accuracy of the ensemble and eliminates the tradeoff and we achieve the best of both ID and OOD accuracy over the original models.

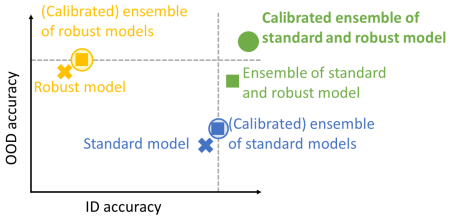
1 INTRODUCTION

Machine learning models typically suffer large drops in accuracy in the presence of distribution shift where the test distribution is different from the training distribution. As ML systems are widely deployed, it is important to train models that achieve good "out-of-distribution" (OOD) accuracy. For example, models trained on medical data from a few hospitals should work well when deployed broadly (Zech et al., 2018; AlBadawy et al., 2018). Similarly, when predicting poverty from satellite imagery, models trained on data from a few countries should work well on all countries, particularly those where labels are scarce due to resource constraints (Jean et al., 2016). There has been a lot of research interest in tackling this robustness problem under various settings such as robustness to spurious correlations (Heinze-Deml & Meinshausen, 2017; Sagawa et al., 2020a; Chen et al., 2020b; Liu et al., 2021), robustness to adversarial perturbations (Goodfellow et al., 2015; Madry et al., 2018), domain generalization (Arjovsky et al., 2019; Sun & Saenko, 2016), robustness to demographic shifts (Hashimoto et al., 2018; Duchi et al., 2019) among others. Almost universally across these different settings, an unfortunate tradeoff arises. Robustness interventions typically improve the OOD accuracy but simultaneously cause a drop in the "in-distribution" (ID) accuracy on new test points from the original distribution.

This tradeoff is a major hurdle in using the multitude of proposed robustness interventions. In practice, most inputs are likely to be ID, so it is unsatisfactory to use a "robust" model that has high OOD performance but performs less accurately on these majority ID points. On the other hand, "standard models" (trained without robustness interventions) fail catastrophically in the presence of even small shifts, and it can be highly dangerous to use a standard model even if OOD points are rare. In this work, we ask *is there a general strategy by which we can achieve high accuracy both in-distribution and out-of-distribution and mitigate tradeoffs arising in robustness?*

We consider four benchmark datasets (DomainNet, CIFAR $\rightarrow$ STL, ImageNet $\rightarrow$ ImageNet-R, and BREEDS-Entity-30) and two real world satellite remote sensing datasets (Landcover and Cropland), that have been used in prior work on robustness. Our work spans different types of robustness interventions (projecting out spurious correlations, zero-shot language prompting, freezing pretrained features), data modalities (image and time series data), and model architectures (vision transformers, deep convolutional networks, time series convolution). Averaged across these datasets, robustness interventions increase OOD accuracy from 63% to 74%, but decrease ID accuracy from 88% to 85%.

We first explore the natural strategy of ensembling the standard and robust models to combine their strengths. Concretely, we add the probabilities of each model to obtain a prediction with the hope



(a) Schematic of performance of different methods

	ID accuracy	OOD accuracy
Standard Model	87.9%	62.7%
Ensemble (Standard + Standard)	88.6%	63.7%
Robust Model	84.7%	74.1%
Ensemble (Robust + Robust)	84.9%	74.3%
Ensemble (Standard + Robust)	89.0%	73.2%
Calibrated ensemble (Standard + Robust)	**89.3%**	**74.6%**

(b) Performance of different methods averaged over 6 real distribution shifts

Figure 1: In many settings, we have a 'standard' model that performs better in-distribution, and a 'robust' model that performs better out-of-distribution. Simply ensembling these two models (e.g., by adding their probabilities), gets better ID accuracy than the standard and robust models, and closes most of the OOD gap. Calibrating the models in-distribution (no access to OOD data) before ensembling them leads to further improvements. Note that ensembling two standard or two robust models does not close the gap and only leads to small improvements.

that when the two models conflict, the more confident model (with larger probability) dictates the final prediction. We find that this surprisingly simple baseline already perfoms quite well—on average across all our datasets, this closes 80% of the gap between the OOD of standard models, while outperforming both models ID. In other words, this simple baseline improves the OOD accuracy over standard models without hurting ID accuracy unlike previous robustness interventions. However, vanilla ensembling still leaves a gap as it underperforms the robust model OOD.

We find that simply calibrating both models ID (adjusting their predicted confidence to match their accuracy, on *in-distribution* data) before ensembling them closes this gap. *Calibrated ensembles get an average accuracy of 89.3% ID and 74.6% OOD, and outperform both the standard and robust model, ID and OOD*. The other method in the literature to alleviate robustness induced tradeoffs is self-training that uses large amount of unlabeled data (Raghunathan et al., 2020; Xie et al., 2021; Khani & Liang, 2021). On the two remote-sensing datasets with additional unlabeled data, we find that calibrated ensembles match self-training on these datasets without requiring any unlabeled data. This shows that calibrated ensembles, though conceptually simple, can be highly effective in mitigating tradeoffs. As a sanity check, we find that the method also works when there is no tradeoff: even when the standard model (or robust model) dominates the robust model (or standard model) both ID and OOD, the calibrated ensemble accuracy matches up with the better model in both domains.

While our method is intuitive in a way, it is also intriguing that it works so well because ensembling seems to rely on good uncertainty estimates while it is common wisdom that uncertainty estimates of deep networks are unreliable out-of-distribution (Ovadia et al., 2019). Furthermore, it has been shown that calibration in-domain does not fix the issue of poor uncertainty estimates out-of-distribution. Indeed, on the six datasets we test on, the models fare poorly on standard uncertainty metrics OOD, even after calibrating ID. The expected calibration error (which roughly measures the difference between the model's confidence and accuracy) of the standard model across all datasets is 12%. A partial answer might be that even if the models have high calibration error on average, what determines the quality of ensembling is the relative confidence between the standard and robust model on a particular data point. For example, on the remote sensing dataset (Landcover), the standard model is on average 6% more confident in its OOD predictions than the robust model, even though the standard model is less accurate OOD—which seems bad. But at the granularity of individual points, we find the more confident model is more likely to be accurate, enabling calibrated ensembles to achieve higher OOD accuracy than both standard and robust models. While most prior work on confidence estimates has focused on a single model at a population level, our results suggest that examining individual data points and relative confidence of different models is an exciting line of future work.

Comparison with regular ensembling. Ensembling is an extremely common trick in ML to boost performance. However, when using ensembles, it is typical to have each ensemble member obtained by running the same training process with a different random seed which controls the stochastic aspects of the training algorithm like initialization, ordering of the training samples in a batch, which features are selected when using drop out and so on. The idea behind ensembling is that with different random seeds, the algorithm might converge to different solutions with decorrelated errors. In this work, we ensemble a standard and robust model which are obtain by *minimizing entirely different training losses*. This small but subtle change is very crucial. Experimentally, we find that regular ensembling of two standard models does poorly OOD (10% worse than a single robust model), and ensembling two robust models does poorly ID (3% worse than a single standard model). For these regular ensembles, calibration has no effect—the ID and OOD accuracies stay the same.

Why is our proposed ensembling so much better than regular ensembling? Consider the case of spurious correlations as studied in (Sagawa et al., 2020b) where standard classifiers use spurious features irrespective of their random initialization. In other words, all members of the regular ensemble would have the same failure mode. The intuition of decorrelated error does not pan out with regular ensembles when standard models make systematic errors, as is common in most distribution shift settings. On the other hand, in our ensemble, we have a standard model that uses spurious features to be more accurate but also a robust model trained specifically not to use spurious features. By combining these "diverse" models with qualitatively different failure modes, we are able to achieve the best of both ID and OOD performance.

In summary, we show that a simple method of calibrating and then ensembling can combine the best of standard and robust models, getting high ID and OOD accuracy—and calibration plays a key role in this method. This is a general practical method to alleviate tradeoffs which can likely be combined with other innovations to completely eliminate robustness induced tradeoffs in practice. This work also raises several interesting conceptual questions regarding ensembling and uncertainty estimation.

2 Setup

Consider a K-class classification task, where the goal is to predict targets $y \in [K]$ from inputs $x \in R^d$.

Models: A model $f : \mathbb{R}^d \to \mathbb{R}^k$ takes an input $x \in \mathbb{R}^d$ and outputs $f(x) \in R^k$ where $f(x)_i$ denotes the model's confidence that the output is $y = i$. The model predicts the label $\widehat{y} = \arg\max_k f(x)_k$. The confidences can be converted into probabilities using a softmax function:

$$\text{softmax}(f(x))_i = \frac{\exp(f(x)_i)}{\sum_{j=1}^{k} \exp(f(x)_j)} \qquad \text{softmax} : \mathbb{R}^k \to \mathbb{R}^k \tag{2.1}$$

Training data: Let P_{id} and P_{ood} denote the underlying distribution of (x, y) pairs in-distribution and out-of-distribution, respectively. We have a validation set $\{(x_i^{\mathsf{val}}, y_i^{\mathsf{val}})\}_{i=1}^{n_{\mathsf{val}}} \sim P_{\mathsf{id}}$ used for early stopping and calibration, a held-out in-distribution test set $\{(x_i^{\mathsf{test}}, y_i^{\mathsf{test}})\}_{i=1}^{n_{\mathsf{test}}} \sim P_{\mathsf{id}}$, and a held-out out-of-distribution test set $\{(x_i^{\mathsf{ood}}, y_i^{\mathsf{ood}})\}_{i=1}^{n_{\mathsf{ood}}} \sim P_{\mathsf{ood}}$.

Evaluation: All methods can use the ID validation set for tuning hyperparameters, early stopping, and calibration. The ID and OOD test set are only used for evaluation. Given a model f, the ID and OOD accuracies are the average accuracies on these test sets:

$$\text{ID-Acc}(f) = \frac{1}{n_{\mathsf{test}}} \sum_{i=1}^{n_{\mathsf{test}}} \mathbb{I}(y_i^{\mathsf{test}} = \arg\max_k f(x_i^{\mathsf{test}})_k) \tag{2.2}$$

$$\text{OOD-Acc}(f) = \frac{1}{n_{\mathsf{ood}}} \sum_{i=1}^{n_{\mathsf{ood}}} \mathbb{I}(y_i^{\mathsf{ood}} = \arg\max_k f(x_i^{\mathsf{ood}})_k) \tag{2.3}$$

3 METHODS

Calibrated ensembles. Given a standard models f_{std} and robust model f_{rob}, calibrated ensembles first calibrate each model using temperature scaling (Guo et al., 2017) with the cross-entropy loss l on the *in-distribution* validation data:

$$T_{\mathsf{std}} = \arg\min_{T} \frac{1}{n_{\mathsf{val}}} \sum_{i=1}^{n_{\mathsf{val}}} l\Big(\frac{f_{\mathsf{std}}(x_i^{\mathsf{val}})}{T}, y_i^{\mathsf{val}}\Big) \tag{3.1}$$

$$T_{\mathsf{rob}} = \arg\min_{T} \frac{1}{n_{\mathsf{val}}} \sum_{i=1}^{n_{\mathsf{val}}} l\Big(\frac{f_{\mathsf{rob}}(x_i^{\mathsf{val}})}{T}, y_i^{\mathsf{val}}\Big) \tag{3.2}$$

The goal of temperature scaling is to adjust each model's confidence on the in-distribution validation data. We then ensemble the two models by adding up the probabilities that they predict (Lakshminarayanan et al., 2017). So given an example x, we simply output the average $\widehat{p}$ of the probabilities predicted by the standard and robust models after scaling by T_{std} and T_{rob} respectively.

$$\widehat{p} = \frac{1}{2}\Big(\text{softmax}\Big(\frac{f_{\mathsf{std}}(x)}{T_{\mathsf{std}}}\Big) + \text{softmax}\Big(\frac{f_{\mathsf{rob}}(x)}{T_{\mathsf{rob}}}\Big)\Big) \tag{3.3}$$

Other ensembles. As baselines, we consider a number of additional ensembles.

Vanilla Ensembling: Here we simply return the average of the standard and robust model's probabilities, without temperature scaling.

$$\widehat{p} = \frac{1}{2}(\text{softmax}(f_{\mathsf{std}}(x)) + \text{softmax}(f_{\mathsf{rob}}(x))) \tag{3.4}$$

Tuned Ensembling: As a potentially stronger baseline, we consider outputting a weighted average of the standard and robust model's probabilities, where the weight $\alpha \in [0, 1]$ is tuned to maximize accuracy on the in-distribution set.

$$\widehat{p} = \alpha\text{softmax}(f_{\mathsf{std}}(x)) + (1-\alpha)\text{softmax}(f_{\mathsf{rob}}(x))) \tag{3.5}$$

Alternative combination methods: We could also combine the logits of the models before taking the softmax, or we could simply output the prediction of the more confident model. We chose to add up the probabilities since that is commonly done by prior work on ensembling (Lakshminarayanan et al., 2017) and we found it to work slightly better than these alternatives.

Other models: Each of these ensembling methods takes two models and combines them. As such, we can apply each of these ensembling methods to two standard models or two robust models as well.

4 DATASETS

We run experiments spanning three different types of robustness interventions: projecting out spurious metadata, language prompting, and freezing pretrained features. These experiments span multiple data modalities and model architectures.

4.1 SPURIOUS METADATA

We run experiments on two remote sensing datasets used in prior work studying ID-OOD tradeoffs (Xie et al., 2021). These datasets consist of a core input x (image data or time series data) and metadata z (e.g., location, meteorological climate data). The metadata is spuriously correlated with the target—using the metadata to predict labels improves accuracy in-distribution (ID), but hurts accuracy out-of-distribution. Xie et al. (2021) consider a standard model that takes in both the core inputs and metadata to predict the target, and a robust model that only takes in the core inputs and does some additional pretraining. They call these the 'aux-in' and 'aux-out' models respectively.

Cropland. The goal is to predict whether a satellite image is of a cropland or not. The core input x is an RGB satellite image, and the metadata z consists of location coordinates and vegetation bands. The original dataset is from Wang et al. (2020), and we use U-net model checkpoints from Xie et al. (2021).

Landcover. The goal is to predict the land type from satellite data at a given location. Here, the core input x is a time series measured by NASA's MODIS satellite (Vermote, 2015), and z is climate data (e.g., temperature) at that location. The dataset is from Gislason et al. (2006); Rußwurm et al. (2020). We use model checkpoints from Xie et al. (2021) where they use 1D convolutions for time series data.

4.2 Zero-shot language prompting

Radford et al. (2021) (CLIP) pretrain a model on a large multi-modal language and vision dataset. The model can then predict the label of an image by comparing the image embedding, with the language embedding for prompts such as 'photo of an apple' or 'photo of a banana'. They show that this zero-shot language prompting approach can be much more accurate out-of-distribution than the traditional method of fine-tuning the entire model.

ImageNet → ImageNet-R. We use a CLIP vision transformer, specifically a ViT-B/16, which is the best publicly available model. The robust model uses language prompts to make zero-shot predictions on ImageNet-Renditions (Hendrycks et al., 2020), a dataset containing cartoon, graffiti, video game, etc, renditions of ImageNet classes. The standard model initializes with weights from the CLIP model, and fine-tunes on ImageNet (Russakovsky et al., 2015) training data for 10 epochs, before making predictions on ImageNet-R. We note that the robust model gets 10% lower accuracy ID (on ImageNet validation examples), but gets 20% higher accuracy OOD (on ImageNet-R test examples)

4.3 Freezing pretrained features

When adapting a pretrained model to an ID dataset, typically all the model parameters are fine-tuned. Recent work looks at 'lightweight' fine-tuning, where only parts of the model are adapted—this can often do better OOD even though the ID performance is worse (Li & Liang, 2021; Houlsby et al., 2019). We consider three distribution shift datasets where the standard model starts from a pretrained initialization and fine-tunes all parameters on an ID dataset, and the robust model only learns the top linear 'head' layer.

DomainNet. A standard domain adaptation dataset (Peng et al., 2019). Here, our ID dataset contains 'sketch' images (e.g., drawings of apples, elephants, etc), and the OOD dataset contains 'real' photos of the same categories. We use the version of the dataset from Tan et al. (2020). We start from a CLIP pretrained ResNet50 and either fine-tune (to get a standard model) or train the head layer (to get a robust model).

CIFAR-10 → STL. Another standard domain adaptation dataset (French et al., 2018), where the ID is CIFAR-10 (Krizhevsky, 2009), and the OOD is STL (Coates et al., 2011). We start from a ResNet50 pretrained on unlabeled ImageNet examples using MoCo-v2 (Chen et al., 2020a) and either fine-tune (to get a standard model) or train the head layer (to get a robust model).

Living-17. Part of the BREEDS benchmark (Santurkar et al., 2020), here the goal is to classify an image as one of 17 animal categories such as 'bear'—the ID dataset contains images of black bears and sloth bears and the OOD dataset has images of brown bears and polar bears. We start from a ResNet50 pretrained on unlabeled ImageNet examples using MoCo-v2 (Chen et al., 2020a) and either fine-tune (to get a standard model) or train the head layer (to get a robust model).

Table 1: *In-distribution (ID)* accuracies for the standard model, robust model, and calibrated ensembling, across six datasets. Calibrated ensembling matches or outperforms the better model in all cases, and on average outperforms both the standard and robust models.

	Ent30	DomNet	CIFAR10	Land	Crop	ImNet
Standard	**93.6 (0.2)**	83.9 (1.0)	**97.4 (0.1)**	**76.9 (0.3)**	95.3 (0.0)	80.5 (-)
Robust	90.7 (0.2)	89.2 (0.1)	92.0 (0.0)	72.7 (0.2)	95.1 (0.1)	68.4 (-)
Cal ensemble	**93.7 (0.1)**	**91.2 (0.6)**	**97.2 (0.1)**	**77.1 (0.2)**	**95.6 (0.1)**	**81.1 (-)**

Table 2: *Out-of-distribution (OOD)* accuracies for the standard model, robust model, and calibrated ensembles, across six datasets. Calibrated ensembling matches or outperforms the better model in 4/6 cases, and on average outperforms both the standard and robust models. For the remaining two datasets, DomainNet and ImageNet-R, calibrated ensembles close 96% and 93% of the gap between the standard and robust model.

	Ent30	DomNet	STL	Land	Crop	ImNet-R
Standard	60.7 (0.1)	55.3 (0.4)	82.4 (0.3)	55.7 (1.1)	**85.6 (5.8)**	36.2 (-)
Robust	63.2 (1.1)	**87.2 (0.1)**	85.1 (0.2)	**60.4 (1.1)**	89.8 (0.4)	**59.1 (-)**
Cal ensemble	**64.8 (0.5)**	85.9 (0.2)	**87.3 (0.2)**	**60.8 (0.8)**	**91.3 (0.8)**	57.4 (-)

5 EXPERIMENTS

5.1 MAIN RESULTS

Strong ID and OOD accuracy: Calibrating and then ensembling a standard and a robust model, gets the best of both worlds, typically outperforming the standard and robust model both ID (Table 1) and OOD (Table 2). Averaged across the datasets, calibrated ensembles get 89.3% ID (vs 87.9% for the standard model and 84.7% for the robust model) and 74.6% OOD (vs 74.1% for the robust model and 62.7% for the standard model). The method works across the board—calibrated ensembles achieve the best performance on 6/6 ID datasets, and on 4/6 OOD datasets. For the remaining two datasets, DomainNet and ImageNet-R, calibrated ensembles close most of the gap between the standard and robust model (96% for DomainNet, 93% for ImageNet).

Competitive with self-training: The remote sensing datasets have lots of unlabeled data so prior work uses self-training on these datasets to mitigate the ID-OOD accuracy tradeoff—we take checkpoints from the official implementation in (Xie et al., 2021) and compare calibrated ensembles with self-training. Table 3 shows that calibrated-ensembles match or outperform self-training on both datasets, both ID and OOD. We believe this is an interesting result because calibrated ensembling is a simple method and does not need additional unlabeled data.

5.2 ENSEMBLING ABLATIONS

We run ablations to show the importance of calibration, and of including both a robust and a standard model in the ensemble. We find that a strong baseline of tuning the ensemble weights on ID data has lower accuracy than calibrated ensembles OOD.

Standard ensembles do not mitigate tradeoffs: Ensembling two standard models or two robust models does not mitigate the ID-OOD accuracy tradeoff. As seen in Figure 1, ensembling two standard models gets an average OOD accuracy of 63.7%, which is substantially worse than our method (74.6%) or even a single robust model (74.1%). Calibrating the standard models does not change its accuracy. Similarly calibrating and ensembling two robust models gets an average ID accuracy of 85.9% which is worse than our method (89.3%) or even a single standard model (87.9%). In summary, calibrated ensembles outperform standard and robust models, ID and OOD (Figure 1).

Calibration is important: We compared calibrated ensembles with vanilla ensembles (no calibration) and tuned ensembles where how to weight the predictions of the standard versus robust model

Table 3: Calibrated ensembles are competitive with self-training (Xie et al., 2021) ID and OOD, which requires unlabeled data.

	Cropland		Landcover	
	ID Acc	OOD Acc	ID Acc	OOD Acc
Standard model	95.3 (0.0)	85.6	76.9	55.7
Robust model	95.1 (0.1)	89.8	72.7	60.4
Self-training	95.3 (0.2)	**90.6 (0.6)**	**77.0 (0.4)**	**61.0 (0.7)**
Cal ensembling	**95.6 (0.1)**	**91.0 (0.8)**	**77.1 (0.2)**	**60.8 (0.8)**

Table 4: OOD accuracies: calibrated ensembles outperform vanilla ensembles and even tuned ensembles where the combination weights are tuned to maximize in-distribution accuracy. Averaged across the datasets, calibrated ensembles get an OOD accuracy of 74.6%, while tuned ensembles get an accuracy of 71.3%. The in-distribution accuracies of the methods are very close (within 0.2% of each other).

	Ent30	DomNet	STL	Land	Crop	ImNet-R
Vanilla	**64.6 (0.4)**	78.7 (1.3)	**87.2 (0.2)**	59.5 (1.0)	**90.9 (0.2)**	**58.0 (-)**
Tuned ID	**64.6 (0.6)**	**86.3 (0.6)**	85.7 (0.9)	58.7 (1.2)	**87.3 (5.7)**	45.4 (-)
Calibrated ID	**64.8 (0.5)**	**85.9 (0.2)**	**87.3 (0.2)**	**60.8 (0.8)**	**91.3 (0.8)**	**57.4** (-)

is learned on ID data. Interestingly, calibrated ensembles substantially outperform tuned ensembles OOD (calibrated ensembles: 74.6%, tuned ensembles: 71.3%) with only a 0.2% drop in ID accuracy relative to tuned ensembles. Naturally, we expect the tuned ensemble to do the best ID since its weights are tailored for ID—what is surprising is that the calibrated ensembles do so much better OOD without using any OOD data either. Calibrated ensembles outperform vanilla ensembles both ID and OOD as well.

6 Related Works and Discussion

Calibration. Calibration has been widely studied in machine learning (Naeini et al., 2014; Guo et al., 2017; Kumar et al., 2019), and applications such as meteorology (Murphy, 1973; DeGroot & Fienberg, 1983; Gneiting & Raftery, 2005), fairness (Hebert-Johnson et al., 2018), and healthcare (Jiang et al., 2012). Many of these works focus on the in-distribution (ID) setting, where models are calibrated on the same distribution that they are evaluated on. Ovadia et al. (2019) show that if we calibrate (e.g., via temperature scaling) a model ID, it still has poor uncertainties OOD. Jones et al. (2021) also show that model uncertainties can be quite unreliable out-of-distribution. However, we show that despite having poor uncertainties on traditional metrics, calibrated models can be combined effectively to mitigate ID-OOD tradeoffs.

Ensembling. Ensembling models is a common way to get an accuracy boost—typically the ensemble members are trained on the same data, but with a different random seed (Lakshminarayanan et al., 2017) or augmentation (Stickland & Murray, 2020). In the setting where the ensemble members mostly differ by random seeds or augmentations, prior work has shown that calibrating the members of an ensemble does not help (Wu & Gales, 2021; Ovadia et al., 2019). Indeed, we find that calibration has minimal effect when we ensemble two standard, or two robust models, that are trained from different seeds. However when we combine two very different models (standard and robust), calibration leads to clear improvements.

Mitigating ID-OOD tradeoffs. Tradeoffs between ID and OOD accuracy are widely studied and prior work self-trains on large amounts of unlabeled data to mitigate such tradeoffs (Raghunathan et al., 2020; Xie et al., 2021; Khani & Liang, 2021). In contrast, our approach uses no extra unlabeled data and is a simple method where we just add up the model probabilities after a quick calibration step. In concurrent and independent work, (Wortsman et al., 2021) show on ImageNet and variants

(e.g. ImageNet-v2, ImageNet-sketch, ImageNet-R) that there *exists* a way to ensemble a CLIP zero-shot and fine-tuned model to get good ID and OOD accuracy—however learning exactly how to ensemble the models might require OOD data, which is not available. We show that the natural way to learn how to weight ensemble members—selecting the weights to optimize in-distribution accuracy—does not mitigate the ID-OOD gap, but calibrated ensembles do.

Discussion. In this paper, we show that a simple method of calibrating and then ensembling models can eliminate the tradeoff between in-distribution (ID) and out-of-distribution (OOD) accuracy. We hope that this leads to more widespread use and deployment of robustness interventions. One mystery is why calibrated ensembles work so well—especially when tuning ensemble weights ID does not, and the worse standard model can still be more confident OOD.

REFERENCES

EA AlBadawy, A Saha, and MA Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys.*, 45, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv*, 2020a.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Adam Coates, Andrew Ng, and Honlak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 215–223, 2011.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:12–22, 1983.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. `https://cs.stanford.edu/~thashim/assets/publications/condrisk.pdf`, 2019.

Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.

Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

Tilmann Gneiting and Adrian E. Raftery. Weather forecasting with ensemble methods. *Science*, 310, 2005.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

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

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

Ursula Hebert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, 2018.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *arXiv*, 2019.

Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 2016.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations (ICLR)*, 2021.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Association for Computational Linguistics (ACL)*, 2021.

Evan Zheran 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 (ICML)*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Binary classifier calibration: Non-parametric approach. *arXiv*, 2014.

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

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, 2021.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 200–201, 2020.

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 (ICLR)*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020b.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv*, 2020.

Asa Cooper Stickland and Iain Murray. Diverse ensembles improve calibration. *arXiv*, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV)*, 2016.

Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: An empirical odyssey. *arXiv preprint arXiv:1910.10320*, 2020.

E. Vermote. MOD09A1 MODIS/terra surface reflectance 8-day L3 global 500m SIN grid V006. `https://doi.org/10.5067/MODIS/MOD09A1.006`, 2015.

Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12, 2020.

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

Xixin Wu and M. Gales. Should ensemble members be calibrated? *arXiv*, 2021.

Sang Michael Xie, Ananya Kumar, Robert Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-N-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations (ICLR)*, 2021.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. In *PLOS Medicine*, 2018.