Reconsidering Deep Ensembles

Anonymous Author(s)
Affiliation
Address
email

Abstract

Ensembling neural networks is an effective way to increase accuracy, and can often match the performance of individual larger models. This observation poses a natural question: given the choice between a deep ensemble and a single neural network with similar accuracy, is one preferable over the other? Recent work suggests that deep ensembles may offer distinct benefits beyond predictive power: namely, uncertainty quantification and robustness to dataset shift. In this work, we demonstrate limitations to these purported benefits, and show that a single (but larger) neural network can replicate these qualities. First, we show that ensemble diversity, by any metric, does not meaningfully contribute to an ensemble's ability to detect out-of-distribution (OOD) data, and that one can estimate ensemble diversity by measuring the relative improvement of a single larger model. Second, we show that the OOD performance afforded by ensembles is strongly determined by their in-distribution (InD) performance, and—in this sense—is not indicative of any "effective robustness." While deep ensembles are a practical way to achieve improvements to predictive power, uncertainty quantification, and robustness, our results show that these improvements can be replicated by a (larger) single model.

1 Introduction

In many real-world settings, practitioners deploy ensembles of neural networks that combine the outputs of several individual models [e.g. 58, 36, 63]. Though training and evaluating multiple models is computationally expensive, a wide body of research demonstrates that ensembles achieve better performance (as measured by accuracy, negative log likelihood, or a variety of other metrics) than their constituent single models, provided that these models make diverse errors [13]. This benefit is well-established in the literature: theoretically proven for ensembles formed via boosting or bagging [54, 4], and demonstrated for *deep ensembles* that solely rely on the randomness of SGD coupled with non-convex loss surfaces [38, 15].

Of course, ensembling is not the only way to increase performance; one could also increase the depth or width of a single neural network. In many settings, a single large model performs similarly to an ensemble of (smaller) models with a similar number of parameters [39, 32, 60]. This observation poses a natural question: are there reasons to choose a deep ensemble over a single (larger) neural network with comparable performance?

Recent research suggests that deep ensembles may be preferable to single models in safety-critical applications and settings where data shifts significantly away from the training distribution. First, Lakshminarayanan et al. [37] demonstrate that deep ensembles provide *well-calibrated estimates of uncertainty* on classification and regression tasks. Compared with other uncertainty quantification (UQ) methods, ensembles offer larger (i.e. less overconfident) uncertainty estimates on out-of-distribution (OOD) or shifted data [49]. Second, recent work indicates that—beyond calibration—ensemble performance (as measured by accuracy, NLL, or other metrics) also tends to be *robust against dataset shift*, again often outperforming other methods in these regimes [24].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Intuitions in recent papers [e.g. 38, 15] attribute these UQ/robustness benefits to the fact that ensembles produce multiple diverse predictions, rather than a single point prediction. If diversity does in fact explain UQ/robustness improvements, this would suggest that deep ensembles can indeed offer benefits that could not be obtained by (standard) single neural networks. In this paper, we rigorously test hypotheses that formalize this intuition. Surprisingly, after controlling for factors related to single model performance, we find no evidence that having a diverse set of predictions is responsible for these purported benefits. Put differently, we find that these UQ/robustness benefits are not unique to deep ensembles, *as they can be replicated through the use of (larger) single models*.

Hypothesis: ensemble diversity is responsible for improved UQ. Two components contribute to ensemble uncertainty estimates: the uncertainties expressed by individual ensemble members, and diversity among ensemble member predictions. Recent work suggests that the diversity component is primarily responsible for larger (i.e. better calibrated) OOD uncertainty estimates, as ensemble members should agree less (i.e. offer more diverse predictions) as data shift away from the training distribution [37, 15, 24]. In contrast, we find that—after conditioning on the uncertainty of individual ensemble members—*the level of ensemble disagreement does not statistically differ between InD and OOD data* (Fig. 1), and thus ensemble diversity is not directly responsible for larger OOD uncertainty estimates. Furthermore, ensemble diversity—on a per-datapoint basis—is correlated with the expected improvement we would obtain by increasing model capacity (Fig. 2), implying that ensemble diversity does not capture a quantity inaccessible to a single (larger) model.

Hypothesis: ensemble diversity is responsible for improved robustness. Independent work demonstrates a deterministic relationship between a (single) neural network's 0-1 accuracy on InD and OOD datasets [59, 43], whereby the OOD performance of a model can be predicted from its InD performance. It is therefore natural to ask whether having multiple diverse predictions provides direct, outsize benefits to OOD robustness (as suggested by [15, 24]), or whether this diversity is simply improving InD performance and reaping the benefits to OOD performance expected given our single network baseline. Our results demonstrate that deep ensembles are not "effectively robust" relative to single models—i.e. their OOD performance (as measured by accuracy, NLL, Brier score, and calibration error) follows the same deterministic relationship to InD performance (Fig. 4). Therefore, ensemble diversity does not yield additional robustness over what standard single networks achieve.

Implications. This paper does not disagree per se with prior claims about the benefits of ensembles. Indeed, we confirm that ensembles improve many metrics of interest. At the same time, our results also indicate that—after controlling for individual model uncertainty and InD performance— ensemble diversity does not further contribute to improved UQ/robustness. We confirm these results for a wide variety of model architectures, as well as for *heterogeneous deep ensembles* that combine multiple different neural network architectures and *"implicit" deep ensembles* like MC Dropout [18], BatchEnsemble [61], and MIMO [26] (Appx. H.4). Our results suggest that—while deep ensembles are a convenient mechanism to improve predictive performance, UQ, and robustness—they do not provide benefits distinct from what could be achieved by a standard neural network.

2 Related work

Ensembling is an established technique to reduce generalization error [e.g. 54, 50, 14, 48], where the predictions of diverse models are aggregated to reach a consensus. It is well established that diversity amongst ensemble members is necessary to achieve performance improvements [13]. This diversity can be achieved through many means. Randomization approaches introduce diversity by training each model on a random subset of data [4] or a random subset of features [5]. Alternatively, boosting approaches [16, 17] achieve diversity by manipulating the weighting of training data. Other methods include using a diverse set of model classes [e.g. 8] or joint training objectives [e.g. 45].

Ensembles of neural networks. Historically, neural network ensembles have relied on a variety of mechanisms to introduce diversity [e.g. 25, 50, 44, 65]. Recently, diversity is often obtained by training multiple copies of the same neural network architecture with different intializations and minibatch orderings, as the inherent randomness of SGD has been shown to introduce a sufficient amount of diversity in these (non-convex) models. [38, 21, 15]. Importantly, this approach can exploit parallel computation [37], because none of the ensemble members depend on one another.

Deep ensembles for predictive uncertainty. It has been suggested that ensembles of neural networks not only improve accuracy but also estimates of predictive uncertainty [37]. Some research aims

to connect ensembles and Bayesian neural networks, suggesting that these improved uncertainty estimates are the result of performing approximate Bayesian model averaging [18, 62]. While the degree of this connection remains an ongoing discussion, deep ensembles are widely utilized for UQ as they have been shown to improved calibration and OOD detection over single networks [49, 24].

Deep ensembles and robustness. Robustness is the ability to maintain good accuracy and calibration under conditions of distributional shift. Deep ensembles outperform other approaches in maintaining both accuracy and calibration on OOD data [49, 24], although their limitations have also been demonstrated [35, 51]. This robustness is attributed to the diversity between ensemble members [15].

Other related work. Recent work investigates whether it is possible to achieve the benefits of an ensemble with reduced computation during training and/or test time [29, 40, 61, 26]. Other work aims to leverage the optimization landscape of single models to produce ensemble-like performance with reduced compute during training [29, 40]. Both historical and more recent work has analyzed diversity metrics for ensemble related quantities similar to those we examine here [e.g. 34, 42, 2], with applications such as improving ensemble training and the design of active learning strategies.

3 Setup

Consider multiclass classification: inputs $\boldsymbol{x} \in \mathbb{R}^D$ with targets $y \in [1, \ldots, C]$, where D is the number of features and C is the number of classes. We assume that we have access to M distinct neural networks $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$, where each model $\boldsymbol{f}_i : \mathbb{R}^D \to \Delta^C$ maps an input to the C-class probability simplex. We will primarily focus on the common case of **homogeneous ensembles**, where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$ represent the same neural network architecture and training procedure, relying on the inherent randomness of initialization and SGD to produce diverse models (See Sec 2 for a broad discussion). However, in Sec. 5.3 we will also consider **heterogeneous ensembles** where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$ represent different architectures or training procedures, and **implicit ensembles**, where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$ are approximated by changes to a single model [18, 26, 61]. Throughout the paper, we will also represent these member networks as a discrete distribution of models: $p(\boldsymbol{f}) = \text{Unif.}[\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M]$. The ensemble prediction $\bar{\boldsymbol{f}}(\boldsymbol{x})$ is given by the arithmetic mean of the ensemble member *probabilities*:[1]

$$\bar{\boldsymbol{f}}(\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})] = \tfrac{1}{M} \textstyle\sum_{i=1}^{M} \boldsymbol{f}_i(\boldsymbol{x}) \tag{1}$$

Metrics for ensemble diversity. Two metrics of ensemble diversity are 1. variance [e.g. 31], and 2. Jensen-Shannon divergence [e.g. 37, 15]. Mathematically, they are (respectively) defined as:

$$\operatorname*{Var}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})] = \textstyle\sum_{i=1}^{C} \operatorname*{Var}_{p(\boldsymbol{f})}\left[f^{(i)}(\boldsymbol{x})\right], \quad \operatorname*{JSD}_{p(\boldsymbol{f})}[y \,|\, \boldsymbol{f}(\boldsymbol{x})] = \mathrm{H}\big[y \,|\, \bar{\boldsymbol{f}}(\boldsymbol{x})\big] - \operatorname*{\mathbb{E}}_{p(\boldsymbol{f})}\big[\mathrm{H}\,[y \,|\, \boldsymbol{f}(\boldsymbol{x})]\big] \tag{2}$$

where H is the entropy. Both metrics are always positive and minimized when ensemble members are the same, i.e. not diverse.

Models and training datasets. We reuse and train a variety of convolutional neural networks on two benchmark image classification datasets: **CIFAR10** [33] and **ImageNet** [11]. In particular, we include the 137 models trained on CIFAR10 by [43], corresponding to 32 different architectures each trained for 2-5 seeds; as well as the "standard" 78 models trained on ImageNet by [59], each corresponding to a different architecture trained for 1 seed. To form homogeneous ensembles, we additionally train ten types of network architectures for CIFAR10 and three for ImageNet. We train 5 independent seeds of each model architecture, where each seed differs only in terms of initialization and minibatch ordering. We form homogeneous deep ensembles by combining 4 out of the 5 random seeds. From this, we can consider 5 single model replicas and 5 ensemble replicas for each model architecture. Unless otherwise stated, ensembles are formed following Eq. (1).

OOD datasets. A majority of our analysis concerns comparing deep ensembles on InD versus OOD test data. To that end, we consider three different catagories of OOD datasets as suggested by [43]: *Shifted reproduction datasets.* This includes the **CIFAR10.1** and **ImageNetV2** datasets [53], both of which were collected and labeled following the same curation processes of the original CIFAR10 and ImageNet datasets. Subtle differences in the curation process create a slight but discernible difference between model performance between these new datasets and the original test sets. Consequentially,

[1] While it is also possible to average the logits (log probabilities) of each model, we note that probability averaging is far more common in the literature [e.g. 37].

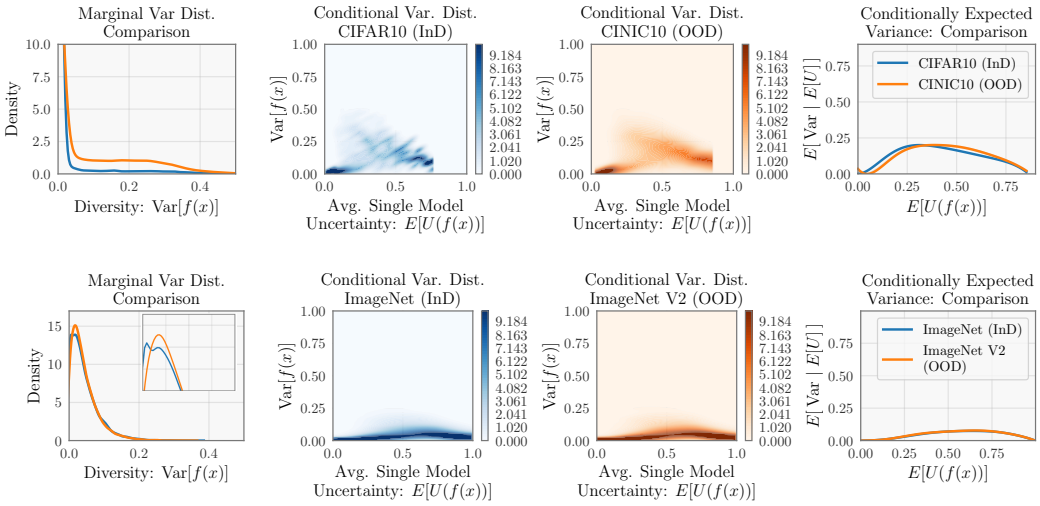

Figure 1: **Given average single model uncertainty, ensemble diversity does not capture dataset shift**. Panels compare ensemble variance ($\mathrm{Var}[\boldsymbol{f}(\boldsymbol{x})]$) on InD (blue) vs. OOD (orange) data. The top row represents the variance for ensembles composed of five WideResNet 28-10 [64] networks evaluated on CIFAR10 and CINIC10, and the bottom row represents the variance for ensembles of five AlexNets, evaluated on ImageNet and ImageNetV2. The ensembles are formed as described in Sec. 3. The left column shows that, consistent with previous results, deep ensembles express more diverse, higher variance predictions on OOD vs InD data, when comparing $p(\mathrm{Var})$ for InD vs. OOD datasets. The middle columns show $p(\mathrm{Var} \mid \mathbb{E}[U])$ (arguments suppressed for clarity) for InD (second column) and OOD data (third column). In the right columns, we show the similarity of these conditional distributions (InD and OOD) using the conditional expectation $\mathbb{E}[\mathrm{Var} \mid \mathbb{E}[U]]$, estimated with kernel ridge regression.

neural networks (trained on the original datasets) tend to achieve worse performance on these new test sets. *Alternative benchmark datasets.* The **CINIC10** dataset [9] shares the same classes as CIFAR10 but uses images drawn and downsampled from the ImageNet dataset. Because ImageNet and CIFAR10 images were collected using different curation procedures, models trained on CIFAR10 tend to achieve worse performance on CINIC10. *Synthetically corrupted datasets.* The **CIFAR10C** and **ImageNetC** datasets [28], apply synthetic perturbations to CIFAR10 and ImageNet images (e.g. Gaussian blur, fog effects, etc.). Due to their synthetic nature, these datasets offer shifts of various intensity (e.g. mild blur versus heavy blur). Models tend to perform worse on these data, especially those with intense corruptions. We relegate most of our analysis of these datasets to the Appendix.

4 Hypothesis: ensemble diversity is responsible for improved UQ

The ability of deep ensembles to produce higher estimates of uncertainty on OOD data has been attributed to ensemble diversity [37, 15, 62]. According to this hypothesis, one would expect that OOD predictions from individual ensemble members are less constrained by their shared training data, and reflect more of the variability introduced by independent initialization of ensemble members [37]. This hypothesis is attractive because it suggests that deep ensembles offer an additional mechanism for uncertainty quantification beyond what is afforded by any single model. In this section, we test this hypothesis by quantifying the contribution of ensemble diversity to a deep ensemble's total predictive uncertainty, and analyzing how diversity impacts InD versus OOD uncertainty estimates.

4.1 Metrics for ensemble diversity

Common metrics for deep ensemble diversity admit interpretable decompositions of the full ensemble uncertainty: *ensemble uncertainty* = *ensemble diversity* + *average single model uncertainty*. For example, if we use variance (Eq. 2 as a metric for ensemble diversity [31, 24], then we show (see

derivation in Appx. C) ensemble uncertainty can be decomposed as:

$$\overbrace{U\left(\bar{\boldsymbol{f}}(\boldsymbol{x})\right)}^{\text{ens. uncert.}} = \overbrace{\operatorname*{Var}_{p(\boldsymbol{f})}\left[\boldsymbol{f}(\boldsymbol{x})\right]}^{\text{ens. diversity}} + \overbrace{\operatorname*{E}_{p(\boldsymbol{f})}\left[U\left(\boldsymbol{f}(\boldsymbol{x})\right)\right]}^{\text{avg. single model uncert.}}. \tag{3}$$

where $U(\boldsymbol{f}(\boldsymbol{x}))$ is a quadratic notion of uncertainty:

$$U\left(\boldsymbol{f}(\boldsymbol{x})\right) \triangleq 1 - \sum_{i=1}^{C}\left[p(y=i \mid \boldsymbol{f}(\boldsymbol{x}))\right]^2.$$

Intuitively, U will be small when most probability is placed on a single class, and will be large when probability is distributed amongst classes (see Appx. C for analogous results with Jensen Shannon divergence (Eq. 2)). Throughout this section we will suppress arguments when clear from context. We use this decomposition to test the hypothesis that ensemble diversity is directly responsible for higher OOD uncertainty estimates. In particular, we expect that ensemble diversity (Var) should increase on OOD data *independently* of average single model uncertainty ($\mathbb{E}[U]$). In other words, given *any* level of $\mathbb{E}[U]$, we would expect more ensemble diversity for OOD data than InD data.

4.2 Experiment: InD vs OOD ensemble diversity

We test 10 different ensembles trained on CIFAR10, and three ensembles trained on ImageNet (ensemble size $M = 5$ for all ensembles, see Fig. 1 for details). We evaluate these ensembles on their respective InD (CIFAR10, Imagenet) and OOD (CIFAR10.1, CINIC10, CIFAR10C, ImageNet V2, ImageNetC) test sets. (See Appx. F for a complete set of results.) In Fig. 1, we analyze the variance of deep ensembles evaluated on CIFAR10 vs CINIC10 (top row) and ImageNet vs ImageNetV2 (bottom row). Fig. 1 (left) shows the distribution $p(\mathrm{Var})$. Ensembles tend to express higher variance on OOD data than InD data; which is consistent with previous work demonstrating that ensemble diversity is correlated with improved OOD data detection [37, 15]. However, we emphasize this result is not sufficient to directly attribute ensemble diversity as the source of UQ improvements.

Controlling for single model uncertainty. A different picture emerges when we control for single model uncertainty by working with the conditional distribution, $p(\mathrm{Var} \mid \mathbb{E}[U])$ (arguments suppressed for clarity). Fig. 1 (middle) shows histograms of ensemble variance *conditioned on* average single model uncertainty (as given by Eq. 3). Surprisingly, we see that the OOD and InD conditional histograms are very similar. We further study this similarity in Fig. 1 (right), which plots expected ensemble variance conditioned on average single model uncertainty: $\mathbb{E}[\,\mathrm{Var} \mid \mathbb{E}[U]\,]$. Far from demonstrating an increase across all levels of average single model uncertainty as we would expect given our hypothesis, we observe that the conditional expectation of ensemble diversity on InD vs OOD data is nearly identical. In Appx. F (Fig. 7-Fig. 11), we offer statistical validation of these observations, and further demonstrate that this phenomenon holds across various architectures, InD, and OOD datasets. In all cases, the difference between the InD and OOD expected variance is only a few percentage points, and/or not statistically significant.

Understanding the relationship between ensemble diversity and average single model uncertainty. By controlling for average single model uncertainty, we see that ensemble diversity does not differ significantly for InD versus OOD data. In turn, these results imply that the effective change in ensemble diversity we see in Fig. 1 (left) must be due in entirety to a change in the distribution of *average single model uncertainty*, $p(\mathbb{E}[U])$. Specifically, an increase in the number of points with given values of single model uncertainty is the only mechanism by which ensemble diversity can increase. From these results, we can conclude that surprisingly, the UQ benefits of ensemble diversity are dictated by properties of the corresponding average single model.

4.3 What does ensemble diversity actually measure?

Our analysis above shows that ensemble diversity is not responsible for the improved OOD uncertainty estimates offered by ensembles. To begin to understand why this might be the case, it is useful to consider the link between ensemble diversity and accuracy. Outside of neural networks, it has long been established that diversity amongst ensemble members is a necessary and sufficient condition for the superior performance of ensembles [e.g. 13]. To demonstrate this, consider any strictly convex loss function, such as negative log likelihood (NLL) or the multiclass Brier score (B) [6]:

$$\mathrm{NLL}(\boldsymbol{f}(\boldsymbol{x}), y) \triangleq -\log\left(f^{(y)}(\boldsymbol{x})\right), \quad \mathrm{B}(\boldsymbol{f}(\boldsymbol{x}), y) \triangleq \|\boldsymbol{f}(\boldsymbol{x}) - \mathbf{1}_y\|_2^2. \tag{4}$$

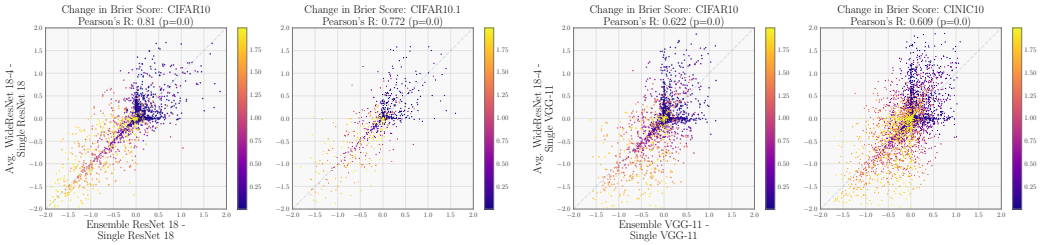

Figure 2: **Per-datapoint improvements from ensembling are meaningfully correlated with improvements from increasing model capacity**. Panels illustrate the per-datapoint gains in Brier score over a single model by either forming a deep ensemble (x-axis), or evaluating the average performance-matched single model (y-axis). (Left) For CIFAR10/CIFAR10.1, the base model is a Resnet 18, the ensemble in the x-axis is of size 4, and the second model in the y-axis is a WideResnet 18-4. Colors indicate the Brier score achieved by the single Resnet 18 model on each datapoint. Both InD and OOD performance improvements are strongly correlated. (Right) For CIFAR10/CINIC10, the base model is a VGG11 model, and the class of single model used for comparison is once again a WideResNet 18-4 model. Improvements are indistinguishable from relevant controls, Appx. I.

(Here, $\mathbf{1}_y$ represents a one-hot encoding of y.) Recall that the ensemble prediction $\bar{\boldsymbol{f}}(\boldsymbol{x})$ is the average of all model predictions (i.e. $\mathbb{E}_{p(\boldsymbol{f})[\boldsymbol{f}(\boldsymbol{x})]}$). By Jensen's inequality:

$$\mathrm{NLL}(\bar{\boldsymbol{f}}(\boldsymbol{x}), y) \leq \mathop{\mathbb{E}}_{p(\boldsymbol{f})}\left[\mathrm{NLL}(\boldsymbol{f}(\boldsymbol{x}), y)\right], \quad \mathrm{B}(\bar{\boldsymbol{f}}(\boldsymbol{x}), y) \leq \mathop{\mathbb{E}}_{p(\boldsymbol{f})}\left[\mathrm{B}(\boldsymbol{f}(\boldsymbol{x}), y)\right] \tag{5}$$

In other words, the performance of the ensemble (as measured by NLL or Brier score) must be better than the average performance of each ensemble member. Because both NLL and Brier score are strictly convex, the Jensen gap in Eq. (5) will grow as $p(\boldsymbol{f})$ becomes less constant, or more "diverse." In particular, the Jensen gap for Brier score is exactly equal to the ensemble variance (Eq. 2):

$$\mathrm{B}(\bar{\boldsymbol{f}}(\boldsymbol{x}), y) - \mathop{\mathbb{E}}_{p(\boldsymbol{f})}\left[\mathrm{B}(\boldsymbol{f}(\boldsymbol{x}), y)\right] = \mathop{\mathrm{Var}}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})]. \tag{6}$$

(Similar results are well known in the regression context- [e.g. 34, 42]- see Appx. D for a short derivation). In other words, $\mathrm{Var}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})]$ measures the expected predictive improvement we obtain through ensembling. If $\mathrm{Var}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})]$ were also responsible for improving UQ, this would imply that the performance gains from ensembling are somehow fundamentally different than the performance gains from increasing a single model's capacity, as the latter can hurt uncertainty estimates [23]. However, in the next section we will demonstrate that these two methods of increasing performance are in fact correlated.

4.4 Ensembling versus increasing model capacity

Inspired by the above, here we aim to quantify whether ensembling multiple models offers different performance improvements versus increasing the capacity of a single model. To measure this, we compare an ensemble of 4 CIFAR10 models (ResNet18 models) with a single large model (WideResNet-18-4). The ensemble and the large single model achieve comparable Brier Score: 0.084 ± 0.002 on the InD test dataset and 0.210 ± 0.002 on the CIFAR10.1 OOD dataset. In Fig. 2 (left), we plot the Brier score of the ensemble versus the large model on a per-datapoint level, depicting the *improvement correlation* across the dataset. Specifically, given a base ResNet18, the x-axis quantifies the per-datapoint improvement we obtain by ensembling it with other models to form a noisy estimate of Eq. (6), while the y-axis quantifies the improvement we obtain by increasing its model capacity (i.e. use a WideResNet-18-4).

Surprisingly, we find that increasing model capacity and ensembling yield very similar performance improvements *on most individual datapoints*. The ensemble improvements and large model improvements have a Pearson's correlation of 0.81 on the InD test set. Importantly, we see that this correlation is preserved even on OOD data (Pearson's correlation: 0.76). We replicate this result for a different ensemble/larger model pair (VGG-11 ensemble versus WideResNet-18-4) that again have nearly identical InD and OOD performance 0.093 ± 0.004 CINIC10 InD Brier Score; 0.48 ± 0.02

CINIC10 OOD Brier Score) Fig. 2, right. Importantly, we compare each improvement correlation in Fig. 2 to relevant controls (Appx. I). In all cases we find that improvements are as similar as we might expect if comparing two performance matched ensembles, or two single models. This result is unexpected, because the ensemble and the large model represent two distinct architectures (ResNet versus WideResNet) and two different modes of training (independent training of separate models versus training one large model). Recalling the relationship between ensemble diversity and relative performance gains, these results suggest that *ensemble diversity estimates the improvement we should expect by increasing model capacity.* We conclude that, with regards to UQ and performance improvements, ensemble diversity offers little benefit over what can be obtained with single models.

4.5 Implications for epistemic/aleatoric uncertainty estimation

Uncertainty is often broken down into two components. The **epistemic** component captures uncertainty due to a limited number of observations, or uncertainty that the model accurately and uniquely captures the ground truth labeling process. It can be reduced by collecting more data. In contrast, the **aleatoric** component reflects the inherent ambiguity in the data (e.g. a blurry image) and is considered to be irreducible noise. In decision making applications such as active learning [55, 19] or model-based reinforcement learning [36, 63], we can use estimates of these components to identify informative datapoints for our model to sample next [12]. Previous work has interpreted ensemble diversity (Eq. (2)) as epistemic uncertainty [41, 24, 63], with average single model uncertainty in Eqs. (3) and (7) identified as aleatoric uncertainty correspondingly [57]. Our results in Fig. 1 demonstrate that there is a limitation to this interpretation, as we would expect more epistemic uncertainty (ensemble variance) for OOD data than for InD data, independent of aleatoric (single model) uncertainty. We therefore suggest caution when using ensembles to differentiate sources of uncertainty in downstream applications.

5 Hypothesis: ensemble diversity is responsible for improved robustness

Beyond uncertainty quantification, ensembles have been shown to often achieve better predictive performance than single networks (as measured by 0-1 accuracy, NLL, or Brier score) on OOD or shifted datasets [37, 49, 24]. In this section, we test the hypothesis that ensemble diversity improves robustness over what single neural networks can offer.

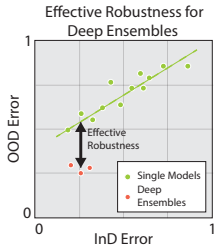

Figure 3: **Cartoon of effectively robust deep ensembles.**

5.1 Effective robustness

We use the concept of "effective robustness" as introduced by Taori et al. [59]. These authors note that there is often a deterministic relationship between a neural network's accuracy on InD data and its accuracy on an OOD dataset (green line in Fig. 3). While some neural networks obtain better OOD accuracy than others, this improved performance can often be predicted by the models' InD accuracy according to this deterministic relationship. A model is considered to be *effectively robust* only if it achieves better OOD accuracy than what is predicted by its InD accuracy. In general, there are very few neural networks or training procedures that exhibit effective robustness against any OOD dataset [59, 43]. To measure the role that ensemble diversity plays in robustness, we quantify to what extent deep ensemble OOD performance can be explained by InD performance (as measured by the deterministic relationship derived from single models). If ensemble performance follows the same deterministic relationship, then they are not effectively robust (i.e. multiple diverse predictors offer no additional robustness over what a single neural network provides).

5.2 Experiment: measuring effective robustness of deep ensembles across metrics

Ensembles are not effectively robust with respect to 0-1 accuracy. Following Miller et al. [43], we measure the InD and OOD error for all the models described in Sec. 3. The top left of Fig. 4 compares the error of models on CIFAR10 (InD) versus CINIC10 (OOD), and the bottom left plot compares the error of models on ImageNet (InD) versus ImageNetV2 (OOD). From these plots, we observe several trends. In agreement with Taori et al. [59] and Miller et al. [43], we observe that single models (green

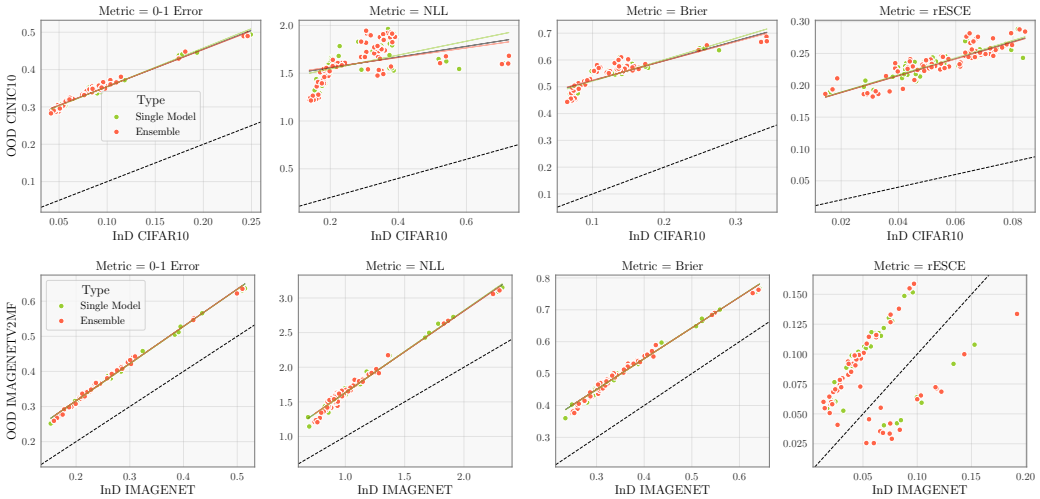

Figure 4: **Deep ensembles follow the same trend as individual models across commonly used InD/OOD performance metrics for a variety of datasets**. Panels illustrate InD vs OOD performance metrics, from left to right: 0-1 Error, NLL, Brier Score, and rESCE. The model types considered are single models, and ensembles, where each marker represents a model from each model type. Linear regressions are shown in solid lines, and black dotted lines indicate perfect robustness. See Appx. H for additional corruptions.

dots) follow a colinear relationship for InD versus OOD accuracy. Additionally, we plot the InD and OOD error of ensembles formed from these single models (orange dots). We find that *ensembles do not deviate from this colinear InD/OOD relationship.* In Appx. H.1, we evaluate the quality of these linear trends. In particular, we fit separate linear trend lines for individual models and deep ensembles. All trend lines achieve correlations of $R > 0.84$, and their coefficients only differ by 1% at most. This suggests that, after controlling for InD accuracy, the OOD accuracy of ensembles is nearly identical to that expected of single models. (See Appx. H for CIFAR10.1/CIFAR10C/ImageNetC results.)

Ensembles are not effectively robust with respect to NLL or Brier score. Although deep ensembles are not effectively robust in terms of predictive accuracy, many of their robustness benefits have been reported in terms of probabilistic metrics, such as NLL or Brier score [49]. We therefore extend our investigation of deep ensemble effective robustness to these metrics. Fig. 4 (middle left) plots the InD NLL and OOD NLL of various ensembles and single models. To the best of our knowledge, this is the first time that the effective robustness experiments of Taori et al. [59] and Miller et al. [43] have been extended to metrics other than 0-1 accuracy. We observe that the relationship between InD NLL and OOD NLL is not as linear as the accuracy trend. In particular, the CIFAR10/CINIC10 linear fit has a correlation coefficient of $R^2 = 0.184$, while the ImageNet/ImageNet-V2 fit has $R^2 = 0.985$. Nevertheless, we observe no discernible difference between the performance of single networks and ensembles (see Appx. H.1 for a quantitative analysis). We observe a similiar phenomenon when we plot InD versus OOD Brier score (Fig. 4, middle right)—ensembles and single models obtain similar OOD Brier score, after controlling for InD Brier score. Our key conclusion is that deep ensembles fail to demonstrate effective robustness when evaluated on probabilistic performance metrics, just as they do with 0-1 accuracy. (See Appx. H for CIFAR10.1/CIFAR10C/ImageNetC results.)

Ensembles do not offer effectively robust calibration. We compare InD and OOD calibration for various single models and ensembles. We consider various metrics for measuring and comparing calibration used throughout the literature. Expected Calibration Error (ECE) [47] is a standard metric for measuring calibration of neural networks. However, we observe few discernible trends relating the InD and OOD ECE of single models (see Appx. H.3 for ECE generalization plots for multiple OOD dataset). As a result, in Fig. 4 we measure calibration using the square root of the Expected *Squared* Calibration Error (ESCE) [10, 46] (see Appx. H.3 for a discussion), which appears in a common decomposition of the Brier score [7]. We therefore expect that any InD/OOD trend for this ESCE should be qualitatively similar to the InD/OOD trends observed for Brier score. Fig. 4 (right) displays the root ESCE (rESCE) for single models and ensembles on InD and OOD data. In the top

plot, we observe a linear trend relating the CIFAR10 (InD) and CINIC10 (OOD) rESCE of single models. The ImageNet models on the other hand follow a bimodal trend, where—depending on the model architecture—InD rESCE is correlated with either low or high OOD calibration. Nevertheless, for both datasets we find that ensembles tend to achieve no better OOD calibration that single models with similar InD calibration. (See Appx. H for CIFAR10.1/CIFAR10C/ImageNetC results.)

5.3 Heterogeneous and implicit ensembles

From the previous results, it is clear that—by many metrics—ensembling multiple copies of the same model architecture confers no additional robustness over single models. A natural question is whether we could achieve more robustness by ensembling different model architectures together. To test this hypothesis, we repeat the same robustness experiments with *heterogeneous ensembles*: ensembles that combine multiple architectures, and *implicit ensembles*: models which aim to approximate the performance of deep ensembles. To construct heterogeneous ensembles, we divide the 137 CIFAR10 models and 78 ImageNet models from Sec. 3 based on their InD accuracy. Ensembles are then formed by randomly selecting 4 models from each bin. This procedure ensures that all ensemble members will have similar accuracy, even though the ensemble members may represent different architectures and training regimens. Despite their additional diversity, these heterogeneous ensembles do not provide effective robustness as shown in Appx. H.4. Finally, we investigate if these results also follow for implicit ensembles, including models constructed with Monte Carlo Dropout [18], multiple-input-multiple-output (MIMO) configurations [26], or as Batch Ensembles [61]. We find that implicit ensembles are not effectively robust as well, as depicted in Appx. H.4.

5.4 Implications.

As discussed in Sec. 4.3, ensemble diversity is responsible for improved NLL and Brier score relative to constituent models. In this sense, ensemble diversity is responsible for improved OOD performance. However, these OOD improvements exactly follow the deterministic trends predicted by (standard) single models, and thus ensembling multiple diverse predictors does not yield any "effective robustness" over what could be achieved by a better performing single model. Unlike prior research [e.g. 49, 24], these results suggest that ensembles are a tool of convenience for obtaining better OOD performance, but not qualitatively any different than single models in this respect.

6 Discussion

In this work, we rigorously test common intuitions about the benefits of deep ensembles to UQ and robustness, and find these explanations wanting. We emphasize that our analysis only focuses on ensembles of neural networks, and does not necessarily apply to ensembling techniques in general (e.g. as applied to random forests or gradient boosted decision trees).

Neural network uncertainty quantification. In examining the conditional distribution of ensemble diversity in Fig. 1, we see that improvements to UQ on OOD datasets are not due to changes in ensemble diversity. In contrast, we find that ensemble diversity is constrained by properties of single model uncertainty estimates. These findings show the role of ensemble diversity in deep ensemble performance is far more limited than initially apparent.

Effective robustness. Our results in Figure 4 show that ensemble diversity does not yield improvements to robustness that cannot be explained by InD performance. This is in line with other results demonstrating that effective robustness is very difficult to achieve [1]. It is of interest to understand these results at the single datapoint level, to integrate with analyses in Sec. 4.

When should we use deep ensembles? We emphasize that our results concur with previous work showing that ensembling can be viewed as a reliable "black box" method of improving neural network performance on a variety of performance metrics, both InD and OOD [37]. Furthermore we agree that it is straightforward (though potentially expensive) to train more model instances, and training a single model that matches the performance of an ensemble is not always straightforward [32, 39, 60]. However we caution that deep ensembles are not a panacea for the issues faced by single models. In particular, it is dangerous to assume that deep ensembles mitigate the robustness issues of single models especially in contexts where we can expect dataset shift, or that ensemble uncertainty provides a reliable baseline for model uncertainty in the absence of ground truth.

References

[1] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[2] Luis Antonio Ortega Andrés, Rafael Cabañas, and Andres Masegosa. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11720–11743. PMLR, 2022.

[3] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

[4] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[7] Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

[8] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18, 2004.

[9] 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.

[10] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

[13] Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[14] Pedro M. Domingos. Why does bagging work? a Bayesian account and its implications. In *KDD*, 1997.

[15] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[16] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[19] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, 2017.

[20] Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *arXiv preprint arXiv:1809.11165*, 2018.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[22] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

[24] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[25] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *Transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[26] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OGg9XnKxFAH.

[27] 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.

[28] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. In *International Conference on Learning Representations*, 2019.

[29] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *International Conference on Learning Representations*, 2017.

[30] 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*, pages 4700–4708, 2017.

[31] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

[32] Dan Kondratyuk, Mingxing Tan, Matthew Brown, and Boqing Gong. When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570*, 2020.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[34] Vedelsby Neural Network Ensembles Krogh. Cross validation and active learning advances in neural information processing systems 7, 1995.

[35] Ananya Kumar, Aditi Raghunathan, Tengyu Ma, and Percy Liang. Calibrated ensembles: A simple way to mitigate ID-OOD accuracy tradeoffs. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[36] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations*, 2018.

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

[38] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

[39] Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, and Dmitry P Vetrov. On power laws in deep ensembles. In *Advances in Neural Information Processing Systems*, 2020.

[40] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, 2019.

[41] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[42] Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33:5479–5491, 2020.

[43] 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*, 2021.

[44] Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *British Machine Vision Conference*, 2016.

[45] Paul W Munro and Bambang Parmanto. Competition among networks improves committee performance. In *Advances in Neural Information Processing Systems*, 1997.

[46] Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(1):41–47, 1977.

[47] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

[48] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[49] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David 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*, 2019.

[50] Michael P Perrone and Leon N Cooper. When networks disagree: Ensemble methods for hybrid neural networks, 1992.

[51] Rahul Rahaman and Alexandre H Thiery. Uncertainty quantification and deep ensembles. In *Advances in Neural Information Processing Systems*, 2021.

[52] Carl Edward Rasmussen and Christopher K Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[53] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

[54] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[55] Burr Settles. Active learning literature survey, 2009.

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

[57] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *Conference on Uncertainty in Artificial Intelligence*, 2018.

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*, 2015.

[59] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, 2020.

[60] Abdul Wasay and Stratos Idreos. More or less: When and how to build convolutional neural network ensembles. In *International Conference on Learning Representations*, 2020.

[61] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

[62] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, 2020.

[63] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

[64] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[65] Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris Holmes, Frank Hutter, and Yee Whye Teh. Neural ensemble search for uncertainty estimation and dataset shift. In *Advances in Neural Information Processing Systems*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We relate our claims to two hypotheses in the introduction Sec. 1, and test each hypothesis in the corresponding results sections Secs. 4 and 5.
   (b) Did you describe the limitations of your work? [Yes] We specify in the discussion Sec. 6 that our results are limited to neural network ensembles, as opposed to more general ensembles.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We do so in Appx. A
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide a link to a repository in the supplemental material section Appx. B. This repository contains instructions to reproduce main figures and to download relevant data.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provide instructions in the code Appx. B which specify internally the data splits and hyperparameters we used. We further specify in Appx. G that we chose default hyperparameters as specified in a separate code repo.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In accordance with checklist guidelines, we report the fact that we ran statistical significance tests for our main results here- in particular, Appx. F describes tests for Fig. 1 and related results, Appx. I describes tests for Fig. 2 and related results, and Appx. H describes tests for Fig. 4 and related results.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We do so in Appx. B

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] In Sec. 3, we reference the origin of the models [43] and [59], and provide further details in Appx. B.

(b) Did you mention the license of the assets? [No] We provide links to publicly released assets with relevant licenses, but do not have a license for models that we were provided by the authors of [43].

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We do not provide new assets.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] In our acknowledgements we thank the authors of [43] for agreeing to share their data with us- all other data is released under a pubic license.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We do not believe this to apply to our data, which consists of deep network models trained using popular deep learning frameworks on benchmark datasets.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]