Scalable Ensemble Diversification for OOD Generalization and Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Training a diverse ensemble of models has several practical application scenarios, such as model selection for out-of-distribution (OOD) generalization and the detection of OOD samples via Bayesian principles. Previous approaches to diverse ensemble training have relied on the framework of letting the models make the correct predictions for the given in-distribution (ID) data while letting them come up with different hypotheses for the OOD data. As such, they require well-separated ID and OOD datasets to ensure a performant and diverse ensemble and have only been verified in smaller-scale lab environments where such a separation is readily available. In this work, we propose a framework, Scalable Ensemble Diversification (SED), for scaling up existing diversification methods to large-scale datasets and tasks (e.g. ImageNet), where the ID-OOD separation may not be available. SED automatically identifies OOD samples within the large-scale ID dataset on the fly and encourages the ensemble to make diverse hypotheses on them. To make SED more suitable for large-scale applications, we propose an algorithm to speed up the expensive pairwise disagreement computation. We verify the resulting diversification of the ensemble on ImageNet and demonstrate the benefit of diversification on the OOD generalization and OOD detection tasks. In particular, for OOD detection, we propose a novel uncertainty score estimator based on the diversity of ensemble hypotheses, which lets SED surpass all the considered baselines in OOD detection task. Code will be available soon.

1 Introduction

Training a diverse ensemble of models is useful in multiple applications. Diverse ensembles are used to enhance out-of-distribution (OOD) generalization, where strong spurious features learned from the in-distribution (ID) training data hinder generalization [30, 31, 28, 23]. By learning multiple hypotheses, the ensemble is given a chance to learn causal features that are otherwise overshadowed by the prominent spurious features [39, 4]. In Bayesian machine learning, diversification of the posterior samples has been studied as a means to improve the precision and efficiency of sample uncertainty estimates [5, 37].

A common strategy to train a diverse ensemble is to introduce two objectives: one for the main task and one for diversification [29, 5, 28, 23]. The main task loss, such as the cross-entropy loss for classification, encourages the hypotheses to solve the task on the labeled ID training set. The diversification loss encourages the hypotheses to diversify the responses on an unlabelled OOD dataset [28, 23] (Figure 1). The datasets for the objectives are separated to avoid contradictory objectives: prediction diversification on the ID set will encourage wrong answers if there is only one correct label.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

This strategy, however, requires a separate OOD dataset where the hypotheses may make diverse predictions without harming the main task performance on the ID training samples. Previous work has thus been tested on hypothetical lab settings where the spurious and causal features can easily be controlled to secure separate ID and OOD datasets for diverse ensemble training. It is not clear yet how one could diversify an ensemble of models for realistic, uncontrolled, and large-scale applications (e.g. ImageNet scale) where collecting a separate OOD dataset can be very costly, if not impossible.

To address the scalability challenge, we propose a novel diversification framework, Scalable Ensemble Diversification (SED, Figure 1). We introduce three ingredients. (1) OOD samples are dynamically selected from the ID training samples, on which the models are trained to make different predictions. (2) At each iteration, a subset of model pairs are stochastically selected to construct the disagreement objective, rather than the full list of model pairs. (3) Deep networks are trained to diversify only a few layers at the end, rather than the full networks. This framework allows scaling up existing ensemble diversification methods. In this work, we focus on scaling up the Agree to Disagree (A2D) method [28].

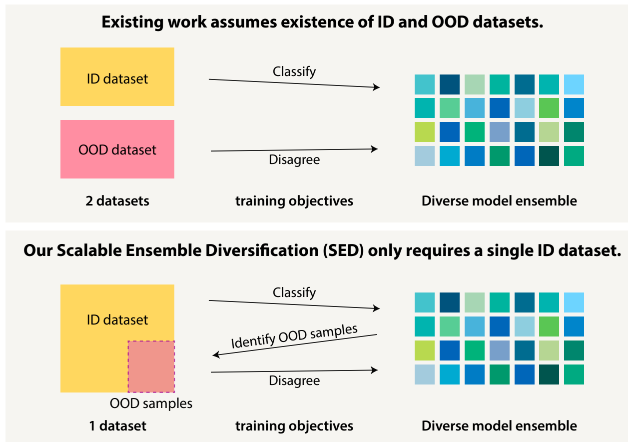


Figure 1: **Existing diversification work vs SED**. Unlike previous diversification approaches that require a separate OOD dataset on which the models are trained to diverge, our Scalable Ensemble Diversification (SED) operates on a single ID dataset where OOD samples are dynamically identified and are used to let the ensemble members diverge.

We verify that SEDdiversifies a model ensemble trained on ImageNet. We demonstrate the benefit of diversification on OOD generalization and OOD detection tasks. For the former, we showcase the usage of SED-diversified ensemble in three variants: (a) *vanilla ensemble* of prediction probabilities [22], (b) an average of the model weights through *model soup* [38], and (c) the *oracle selection* of the individual models for each OOD test set [23, 30]. In all three cases, SEDachieves a superior generalization to OOD datasets like ImageNet-A/R/C, OpenImages, and iNaturalist.

For OOD detection, we seek multiple ways to use the SED-diversified ensemble: (a) treating them as samples of the Bayesian posterior and (b) using our novel OODness estimate of Predictive Diversity Score (PDS) that measures the diversity of predictions from an ensemble. We show that PDS provides a superior detection of OOD samples like ImageNet-A/R/C, OpenImages, and iNaturalist.

Our contributions are

1. Scalable Ensemble Diversification (SED) framework that scales up existing ensemble methods;
2. Predictive Diversity Score (PDS) that computes the OODness score for samples based on ensemble prediction diversity;
3. First demonstration of the ensemble diversification and its application to OOD generalization and detection at ImageNet level.

The code will be released with the next versions of the manuscript.

2 Related work

In this section, we give a short overview of ensembling methods. At first, we speak about ensembles in general and the role of diversity in them (§ 2.1), then we focus on ensembling methods for neural networks and separate them into two big groups. The first group includes algorithms that use loss regularizers (§ 2.2) and the second group covers works that do not modify the training loss (§ 2.3).

2.1 Ensembles as a technique

Ensembling is a powerful technique of aggregating the outputs of multiple models to make more accurate predictions and it has been around for decades [12, 21, 18, 2, 3]. It is well known that diversity in ensemble members' outputs leads to better performance of the ensemble compared to the performance of a single model [21] because ensemble members make independent errors [12, 11]. Therefore, one way to reduce DNNs' reliance on spurious correlations is to train multiple models on the same task and make them diverse in terms of errors they make so that their ensemble is less dependent on such correlations.

2.2 Neural network ensembles that promote diversity through loss regularizers

Diversity in models can be induced by supplying training loss with a suitable regularizer.

Such regularizers can diversify models' weights [5, 7, 34, 6], features [39, 4], input gradients [29, 30, 31, 33] and outputs [25, 5, 28, 23].

Notably, in [5] authors showed that regularizer of a certain structure that repulses ensemble members' weights or outputs leads to ensembles that provide a better approximation of Bayesian Model Averaging. This idea was later extended by works that repulse ensemble members' features [39] and input gradients [33].

Since the ensemble performs better due to the diversity of errors that ensemble members make [21] we want those members to give pairwise different outputs for the same inputs. Unfortunately, diversity in weights space, input gradient space, or features space does not guarantee such property without additional assumptions due to functional symmetry which means that models can be different in terms of their weights or feature maps and input gradients they produce but still give the same outputs for a given input. That is why we are focused on methods that diversify models' outputs, specifically [28, 23] which are state-of-the-art according to [1] and use regularizer of repulsive nature conceptually similar to [5].

2.3 Neural network ensembles that promote diversity without modifying loss

In addition to loss regularizers, there were an uncountable number of different ways to induce diversity in ensembles of neural networks that did not modify the training loss. The most straightforward approach of independently training multiple models of the same architecture by changing only random seeds is called Deep Ensemble [22] which was extended from the Bayesian perspective in [37]. Another solution is to construct an ensemble from models trained with different hyperparameters [36], augmentations [24], or architectures [40]. More computationally efficient direction allows training only one base model inducing diversity by ensembling either checkpoints saved in different local minima along the training trajectory of this base model [19] or models produced by the base model after applying dropout [10] or masking [9] to it. The mixture of experts paradigm can also be viewed as an ensemble diversification technique [41] where diversification happens due to assigning different training samples to different ensemble members.

Despite their conceptual simplicity Deep Ensembles [22] and ensembles of models trained with different hyperparameters [36] are strong baselines for OOD detection [27] and OOD generalization tasks, especially when combined with model souping techniques [38]. That is why we selected them as baselines for our experiments.

3 Method

We present our main technical contributions, Scalable Ensemble Diversification (SED, §3.2) and the Predictive Diversity Score (PDS, §3.3).

3.1 Preliminaries

We cover background materials before introducing our main technical contributions. We work with a training set $\mathcal{D} := \{x_n, y_n\}_{n=1}^{N}$, which we refer to as the in-distribution (ID) dataset. For prior diversification methods, we also assume the existence of a separate, unlabeled out-of-distribution

(OOD) dataset $\mathcal{D}^{\text{ood}} := \{x_n^{\text{ood}}\}_{n=1}^{N^{\text{ood}}}$. We write $f(\cdot, \theta)$ for a deep neural network classifier parametrized by θ. $f(x;\theta) \in \mathbb{R}^C$ indicates the logit outputs for C classes for input x. We write $p(x) := \mathrm{Softmax}(f(x)) \in [0,1]^C$ for the probability outputs. We consider an ensemble $\{f^1, \cdots, f^M\}$ of M models.

3.1.1 Existing ensemble diversification approach

We introduce an existing approach for diversifying an ensemble of models [28, 23]. Two objectives are imposed upon the ensemble of models: the main task loss and the diversification regularization.

For the main task, the community has focused on the classification task. The cross-entropy loss $-\log p_y(x;\theta)$ is used to train the model ensemble $\{f^1, \cdots, f^M\}$ on the ID dataset $\mathcal{D}$:

$$\mathcal{L}_{\text{main}} = \frac{1}{MN} \sum_n \sum_m -\log p^m_{y_n}(x_n;\theta). \tag{1}$$

This encourages each member of the ensemble to behave similarly on the ID dataset.

Different diversification schemes use different diversification regularization loss $\mathcal{L}_{\text{div}}$ applied on pairs (f^m, f^l) of ensemble members. The diversification objective is commonly optimized on the OOD dataset $\mathcal{D}^{\text{ood}}$ to encourage the training of multiple hypotheses on the OOD samples while avoiding clashes with the main task objective. In this work, we focus on the Agree to Disagree [28] method. The diversification loss for a pair (p^m, p^l) is defined as:

$$\text{A2D}(p^m(x), p^l(x)) = -\log\left[p^m_{\hat{y}}(x) \cdot (1 - p^l_{\hat{y}}(x)) + (1 - p^m_{\hat{y}}(x)) \cdot p^l_{\hat{y}}(x)\right] \tag{2}$$

where $\hat{y} := \arg\max_c p^m_c(x)$ is the predicted class for the first model p^m. One may symmetrically define $\hat{y}$ to be the prediction for the second model p^l; in practice, it does not make a difference [28]. Note that the diversification loss favors p^l to predict a lower likelihood for the prediction by p^m, $p^l_{\hat{y}}(x)$, and vice versa. For M models in an ensemble, A2D is applied on the OOD dataset $\mathcal{D}^{\text{ood}}$ for every pair of models (p^m, p^l):

$$\mathcal{L}_{\text{div}} = \frac{1}{N^{\text{ood}} \cdot M(M-1)} \sum_n \sum_{m<l} \text{A2D}(p^m(x_n^{\text{ood}}), p^l(x_n^{\text{ood}})). \tag{3}$$

3.2 Scalable Ensemble Diversification (SED)

We present Scalable Ensemble Diversification (SED) that addresses the limitation of the existing ensemble diversification framework that requires a separate OOD dataset. We introduce two main components of SED: dynamic selection of OOD samples within the ID dataset (§3.2.1) and the stochastic selection of pairs to diverge in the optimization iterations (§3.2.2).

3.2.1 Dynamic selection of OOD samples

If only the ID training dataset is present, it is difficult to induce diversity in ensemble members, as they are uniformly incentivized to solve the main task objective: given x, predict y. Hence, previous approaches have introduced a qualitatively disjoint unlabeled set, which we refer to as the OOD dataset, where the ensemble members are encouraged to disagree with each other. The clear separation of ID and OOD datasets for the two objectives matters for ensuring a good balance between the main task performance and the diversity of hypotheses.

Previous works like Pagliardini et al. [28], Lee et al. [23] have performed experiments on small-scale datasets where factors are well-controlled and clean versions of OOD datasets are readily available. Examples include Waterbirds, Camelyon17, CelebA, MultiNLI, C-MNIST, and the Office-Home datasets. For example, for Waterbirds, the ID dataset is set as the cases where the bird's habitat matches with the visual background and the OOD dataset corresponds to the complementary case.

While conceptually desirable, collecting a separate OOD dataset can be highly cumbersome and expensive. For a large-scale dataset like ImageNet, it is highly non-obvious how one could build a corresponding OOD dataset where the underlying feature-label correlations are different from the ID training dataset.

To address this challenge, we consider dynamically identifying an OOD subset of the ID dataset and letting the ensemble diverge on this subset. The desiderata for the identification of OOD samples within the ID dataset are twofold: (a) we wish to discriminate samples where the ensemble members make mistakes and (b) we only trust the ensemble prediction for the OOD sample identification when the ensemble is sufficiently trained.

We define the sample-wise weight α_n on each ID sample $(x_n, y_n) \in \mathcal{D}$ that satisfy the two conditions:

$$\alpha_n := \frac{\mathrm{CE}(f^1, \cdots, f^M; x_n, y_n)}{\left(\frac{1}{|B|}\sum_{b\in B}\mathrm{CE}(f^1, \cdots, f^M; x_b, y_b)\right)^2} \tag{4}$$

where $\mathrm{CE}(f^1, \cdots, f^M; x_n, y_n) := \mathrm{CE}(\frac{1}{M}\sum_m f^m(x_n), y_n)$ is the loss on the logit-averaged prediction and B is a minibatch that contains the sample (x_n, y_n). α_n is a weight proportional to the ensemble loss on the sample; we thus meet the condition (a). The normalization is designed to handle the condition (b). To see this, consider the batch-wise weight

$$\alpha_B := \frac{1}{|B|}\sum_{b\in B}\alpha_b = \frac{1}{\frac{1}{|B|}\sum_b \mathrm{CE}(f^1, \cdots, f^M; x_b, y_b)}. \tag{5}$$

Note that α_B is now *inversely proportional* to the average cross-entropy loss of the ensemble on the batch B. Thus, the overall level of α_n for $n \in B$ is lower for earlier iterations of the ensemble training, where the predictions from the models are not trustworthy yet.

With this definition of sample-wise weight α_n for the diversification objective, we define the SED objective with the A2D loss for the diversification kernel:

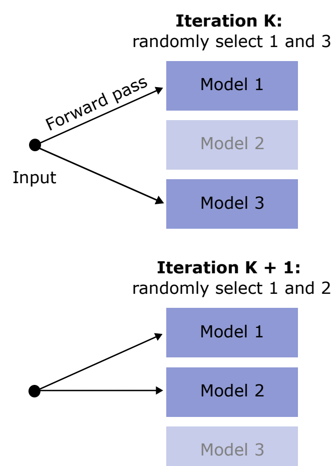
$$\mathcal{L}_{\text{SED}} := \mathcal{L}_{\text{main}} + \frac{\lambda}{NM(M-1)}\sum_n\sum_{m<l}\mathrm{stopgrad}(\alpha_n)\cdot \mathrm{A2D}(p^m(x_n), p^l(x_n)), \tag{6}$$

where $\lambda > 0$ controls the overall weight of the diversification term. Note that, compared to Equation 3, this formulation does not rely on the OOD dataset $\mathcal{D}^{\text{ood}}$. Instead, all ID samples are treated as potential OOD samples, where their OODness is softly determined via α_n. This enables a seamless adaptation of existing ensemble diversification methods to a relaxed setting where a separate OOD dataset is unavailable.

3.2.2 Further tricks for scalability

Ensemble diversification algorithms are often based on pairwise similarities of the members. Pairwise similarity computation scales quadratically with the size of the ensemble M. The second term of Equation 6 is an example of this. This is potentially a hurdle when ensemble diversification is to be applied to $M \geq 10$, and the data and parameter sizes are in the order of millions (e.g. ImageNet).

We address this computational challenge by computing the summation of pairwise distances as a stochastic sum. For every minibatch B of SGD iterations, we uniformly-iid sample a subset $\mathcal{I}$ of $\{1, \cdots, M\}$ to compute the diversification term in Equation 6. The procedure is illustrated in the figure on the right.



To further speed up the SED training, we consider diversifying only a subset of layers, while freezing the other layers. In our experiments, ensemble members share the same frozen feature extractor of Deit3b [32] pretrained on ImageNet-21k [8] and we diversify only the last two layers of the models.

3.3 Predictive Diversity Score (PDS) for OOD Detection

We demonstrate several benefits of the diversified ensembles in §4. One of them is the possibility of using them for detecting OOD samples through the notion of epistemic uncertainty [13]. Given an ensemble of models, a simple baseline for OOD detection is to compute the predictive uncertainty of the Bayesian Model Averaging (BMA) by treating the ensemble members as samples of the posterior $p(\theta|\mathcal{D})$ [22, 37]:

$$\eta_{\text{BMA}} := \max_c \frac{1}{M}\sum_m p_c^m(x). \tag{7}$$

This notion of epistemic uncertainty does not directly exploit the potential diversity in individual models of the ensemble because it averages out the predictions along the model index m.

We propose a novel measure for epistemic uncertainty, Predictive Diversity Score (PDS), that directly measures the prediction diversity of the individual members. The formulation is given below:

$$\eta_{\text{PDS}} := \frac{1}{C} \sum_{c} \max_{m} p_c^m(x). \tag{8}$$

PDS is a continuous relaxation of the number of unique argmax predictions within an ensemble of models. To see this, consider the special case where $p^m \in \{0, 1\}$ are one-hot vectors. Then, $\max_m p_c^m(x)$ is 1 if any of m predicts c and 0 otherwise. Thus, $\sum_c \max_m p_c^m(x)$ computes the number of classes that at least one of the ensemble members predicts. We show that, with our diverse ensembles, PDS outperforms the DE baseline for the OOD detection task (§4.4).

4 Experiments

We verify our contributions, Scalable Ensemble Diversification (SED, §3.2) and Predictive Diversity Score (PDS, §3.3), on ImageNet-scale tasks and datasets. We first verify that SED diversifies the ensemble (§4.2). Then, we demonstrate the application of diversified ensemble to OOD generalization (§4.3) and OOD detection (§4.4) tasks.

4.1 Experimental setup

We task the ensemble with the OOD generalization and OOD detection tasks.

Training settings. For both tasks, we train an ensemble of models with the SED framework with the A2D [28] diversity regularization using AdamW optimizer [26]. We use the default settings of a batch size of 16, learning rate 10^{-3}, weight decay 0.01, and the number of epochs 10. The overall diversity weight λ is set to 0.1 and the stochastic pairing is done for $|\mathcal{I}| = 2$ models for each SGD batch. We use Deit3b [32] network pretrained on ImageNet21k [8] for all the experiments. Following the speed-up trick in §3.2.2, we use only the last 2 layers of the network. For the in-distribution (ID) dataset where the ensemble is trained to diversify, we use the training split of ImageNet with $|\mathcal{D}| = 1,281,167$. All experiments were ran on RTX2080Ti GPUs with 12GB vRAM and 40GB RAM, each experiment took from 2 to 12 hours depending on the complexity of the training.

Baselines. For naive ensemble training, we consider the *deep ensemble* [22] where each ensemble member independently with different random seeds that control the weight initialization and SGD batch shuffling. To match the resource usage of our SED, where we diversify only the last 2 layers of the network, we consider the *shallow ensemble* variant, which is the deep ensemble where only the last 2 layers are trained. We further consider a viable diversification scheme that performs deep ensemble with *varying hyperparameters* [36]. In addition to that, we reimplement A2D [28] and DivDis [23] algorithms and apply them without stochastic model sampling to do classification on labeled samples from ImageNet-Train and disagreement on unlabeled samples from ImageNet-R. For A2D we use frozen feature extractor and a parallel variant of their method which means that all ensemble members are trained simultaneously and not sequentially. The computational complexity of both these approaches scales quadratically with ensemble size which is why they are called Naive A2D and Naive DivDis respectively.

Evaluation benchmarks. The generalization performances of the ImageNet-trained ensembles are measured on multiple test datasets, ranging from the in-distribution validation split of ImageNet with 50,000 samples to OOD datasets like ImageNet-A (*A* [17], 7.5k images & 200 classes), ImageNet-R (*A* [16], 30k images, 200 classes), ImageNet-C (*C-i* for corruption strength i [14], 50k images, 1k classes). OpenImages-O (*OI* [35], 17k images, unlabeled), and iNaturalist (*iNat* [20], 10k images, unlabeled). For OOD detection, we task the ensemble with the detection of the above OOD datasets against the ImageNet validation split.

Evaluation metrics. For OOD generalization, we use the accuracy. For OOD detection, we use the area under the ROC curve, following [15].

					
GT	Cowboy hat	Sea lion	Scuba diver	Great shark	Weimaraner
SED	**Cowboy hat** Comic book	**Sea lion** Otter	**Scuba diver** Jellyfish	**Great shark** Killer whale	**Weimaraner** Vizsla
PDS	0.300	0.300	0.294	0.292	0.292
					
GT	Pomegranate	Zebra	Pomegranate	Pomegranate	Hummingbird
SED	**Pomegranate**	**Zebra**	**Pomegranate**	**Pomegranate**	**Hummingbird**
PDS	0.216	0.216	0.216	0.216	0.216

Figure 2: **ImageNet-R examples leading to the greatest and least disagreement**. We show the 5 most divergent and 5 least divergent samples according to the SED ensemble. We measure the prediction diversity with the Prediction Diversity Score (PDS) in §3.3. GT refers to the ground truth category. Ensemble predictions are shown in bold, in cases when ensemble members predict classes different from the ensemble prediction we provide them on the next line with standard font.

4.2 Diversification

We start with the question of whether Scalable Ensemble Diversification (SED) truly diversify the ensemble at the ImageNet scale. To measure the diversity of the ensemble, we compute the number of unique predictions for each sample for the committee of models (#unique).

Table 1 shows the #unique values for the IN-Val as well as multiple OOD datasets. We observe that the deep ensemble baseline does not increase the diversity dramatically (e.g. 1.09 for C-1) beyond no-diversity values (1.0). Diversification tricks like hyperparameter diversification (1.11 for C-1) or Naive A2D (1.04 for C-1) and DivDis (1.04 for C-1) do not improve the prediction diversity dramatically. On the other hand, our SED increases the prediction diversity across the board (e.g. 5.00 for C-1).

Method	C-1	C-5	iNat	OI
Deep ensemble	1.09	1.19	1.31	1.23
+Diverse hyperparams	1.11	1.32	1.48	1.33
Naive DivDis	1.04	1.14	1.19	1.16
Naive A2D	1.04	1.15	1.19	1.91
SED-A2D	**5.00**	**5.00**	**4.68**	**4.11**

Table 1: **#unique for ensembles.** We report the #unique on OOD datasets (see §4.1 for the datasets). The ensemble size M is 5 for all methods; it is the max possible #unique value.

Qualitative results on ImageNet-R further verify the ability of SED to diversify the ensemble (Figure 2). As a measure for diversity, we use the Predictive Diversity Score (PDS) in §3.3. We observe that the samples inducing the highest diversity (high PDS scores) are indeed ambiguous: for the first image, where the "cowboy hat" is the ground truth category, we observe that "comic book" is also a valid label for the image style. On the other hand, samples with low PDS exhibit clearer image-to-category relationship.

4.3 OOD Generalization

We examine the first application of diversified ensembles: OOD generalization. We hypothesize that the superior diversification ability verified in §4.2 leads to greater OOD generalization due to the consideration of more robust hypotheses that do not rely on obvious spurious correlations.

Ensemble aggregation for OOD generalization. As a means to exploit such robust hypotheses, we consider 3 aggregation strategies. (1) *Oracle selection*: the best-performing individual model is chosen from an ensemble [28, 30]. Final prediction is given by $f(x;\theta^{m^\star})$ where

		Oracle selection					Prediction ensemble					Uniform soup				
Method	M	Val	IN-A	IN-R	C-1	C-5	Val	IN-A	IN-R	C-1	C-5	Val	IN-A	IN-R	C-1	C-5
Single model	1	85.4	37.9	44.7	75.6	38.5	85.4	37.9	44.7	75.6	38.5	85.4	37.9	44.7	75.6	38.5
Deep ensemble	5	**85.4**	37.9	44.9	75.7	38.6	**85.4**	39.9	46.3	75.7	38.6	**85.3**	36.7	44.6	75.5	38.3
+Diverse HPs	5	**85.4**	**38.5**	**45.4**	**77.4**	**40.7**	**85.4**	39.9	46.5	76.0	39.0	**85.3**	35.3	44.1	75.9	38.7
Naive DivDis	5	85.2	35.8	40.8	77.2	40.2	85.1	36.3	41.8	77.2	40.2	84.8	**40.7**	42.5	76.2	38.9
Naive A2D	5	85.2	36.6	44.3	77.3	40.4	85.1	37.8	45.2	77.2	40.3	84.5	39.3	45.1	75.5	39.1
SED-A2D	5	85.1	38.3	45.3	77.2	40.4	85.3	**42.4**	**48.1**	**77.3**	**40.6**	**85.3**	40.3	**46.1**	**77.3**	**40.6**
Deep ensemble	50	**85.5**	38.1	45.2	75.7	38.6	**85.5**	38.8	45.8	75.6	38.5	**85.4**	37.5	45.0	75.5	38.4
+Diverse HPs	50	85.5	38.5	45.6	**77.5**	**40.8**	85.5	42.5	48.5	**76.0**	39.0	**85.4**	36.4	44.8	**75.9**	38.8
SED-A2D	50	82.6	**39.0**	**45.8**	74.4	38.3	83.5	**50.9**	**54.4**	75.8	**39.3**	83.5	**39.2**	**46.5**	75.8	**39.3**

Table 2: **OOD generalization of ensembles.** Models are trained on the ImageNet training split. M is the ensemble size. For Naive DivDis and A2D, we use the ImageNet-R as the OOD datasets where the respective diversification objectives are applied.

$m^\star := \arg\max_m \mathrm{Acc}(f^m, \mathcal{D}^{\mathrm{ood}})$. (2) *Prediction ensemble* is a vanilla prediction ensemble where the logit values are averaged: $\frac{1}{M}\sum_m f^m(x)$ [38]. (3) *Uniform soup* [38] averages the weights themselves. Final prediction is given by $f(x; \frac{1}{M}\sum_m \theta^m)$.

SED improves OOD generalization for ensembles. We show the OOD generalization performances of ensembles in Table 2, for the three ensemble prediction aggregation strategies described above. We observe that our SED framework (SED-A2D) results in superior OOD generalization performances for all three strategies. SED-A2D is particularly strong in prediction ensemble (e.g. 48.1% for $M = 5$ and 54.4% for $M = 50$ on ImageNet-R) and uniform soup (e.g. 46.1% for $M = 5$ and 46.5% for $M = 50$ on ImageNet-R). We contend that the increased ensemble diversity contributes to the improvements in OOD generalization. We also remark that the SED framework (SED-A2D) envelops the performance of Naive A2D in this ImageNet-scale experiment. Together with the superiority of computational efficiency (as discussed at the end of § 4.4) of SED-A2D over the Naive A2D, this demonstrates that SED fulfills its purpose of scaling up ensemble diversification methods like A2D.

Deep ensemble is a strong baseline. We also note that deep ensemble, particularly with diverse hyperparameters, provides a strong baseline, outperforming dedicated diversification methodologies under the oracle selection strategy when $M = 5$. It also provides a good balance between ID (ImageNet validation split) and OOD generalization.

4.4 OOD Detection

We study the impact of ensemble diversification on OOD detection capabilities of an ensemble. Once an ensemble is trained, we compute the epistemic uncertainty, or likelihood of the sample being OOD, following two schemes, η_{BMA} and η_{PDS} introduced in §3.3.

SED and PDS together lead to superior OOD detection performances. We show the OOD detection results in Table 3. For the BMA scores, deep ensemble remains a strong baseline. In particular, when the hyperparameters are varied ("+Diverse HPs"), the detection AUROC reaches the maximal performances among the ensembles using the BMA scores. The quality of PDS is more sensitive to the ensemble diversity, as seen in the jump from the deep ensemble (e.g. 0.589 for OI) to the diverse-HP variant (0.889). However, when the ensemble

Method	η	C-1	C-5	iNat	OI
Single model	BMA	0.615	0.833	0.958	0.909
Deep Ensemble	BMA	0.619	0.835	0.958	0.911
+Diverse HPs	BMA	**0.642**	**0.861**	**0.969**	**0.923**
Naive DivDis	BMA	0.598	0.843	0.966	0.922
Naive A2D	BMA	0.594	0.835	0.966	0.916
SED-A2D	BMA	0.641	0.845	0.960	0.915
Deep Ensemble	PDS	0.565	0.625	0.592	0.589
+Diverse HPs	PDS	0.643	0.849	0.926	0.889
Naive DivDis	PDS	0.600	0.851	0.969	0.939
Naive A2D	PDS	0.599	0.850	0.971	0.939
SED-A2D	PDS	**0.686**	**0.896**	**0.977**	**0.941**

Table 3: **OOD detection via ensembles.** For each OOD dataset (C-1, C-5, iNat, and OI), the ensembles are tasked to detect the respective OOD samples among ID samples (ImageNet validation split). We show the AUROC scores for the OOD detection task. Ensemble size is fixed at $M = 5$. η refers to the epistemic uncertainty computation framework discussed in §3.3.

is sufficiently diverse, such as when trained with SED-A2D, the PDS leads to high-quality OODness scores. SED-A2D with PDS achieves the best AUROC across the board, including the BMA variants.

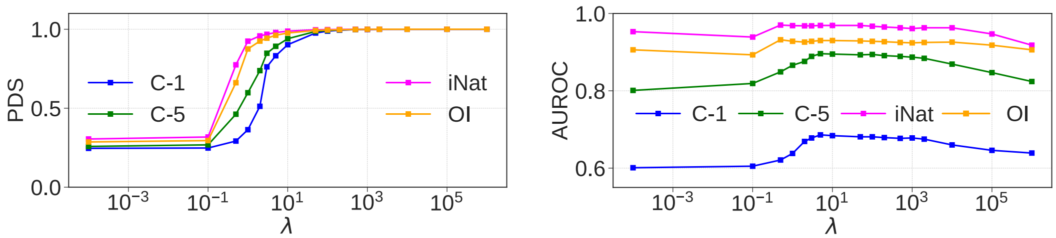


Figure 3: **Impact of diversity regulariser on OOD detection**. We show the model answer diversity, measured by PDS, and the OOD detection performance, measured by AUROC, against λ values, the loss weight for the disagreement regularizer term.

Impact of diversification parameter λ. We further study the impact of ensemble diversification on the OOD detection with the PDS estimator. In Figure 3, we observe that strengthening the diversification objective (higher λ) indeed leads to greater diversity (higher PDS), with a jump at around $\lambda \in [10^{-1}, 10^{1}]$. This range corresponds to the jump in the OOD detection performance (higher AUROC).

Influence of ensemble size. How ensemble size influences performance of our method? We can see that increasing ensemble size helps to improve AUROC for OOD detection on C-1 (Figure 4). Increasing ensemble size marginally helps, but using 5 models provides already a significant improvement over the smallest possible ensemble of size 2. It is also important to mention, that SED framework is computationally more efficient w.r.t. ensemble size M than Naive A2D and Naive DivDis: since we train ensembles for the fixed number of epochs, training complexity for SED is $O(1)$ thanks to stochastic model pairs selection, while for Naive A2D and Naive DivDis it is $O(M^2)$.

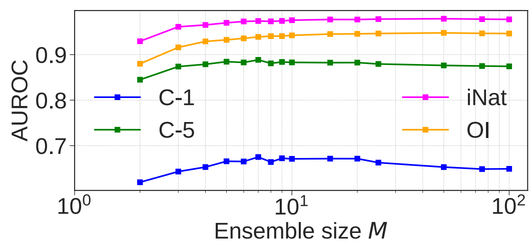


Figure 4: **Impact of ensemble size on OOD detection**.

5 Conclusion

Ensemble diversification has many implications for treating one of the ultimate goals of machine learning, handling out-of-distribution (OOD) samples. By training a large number of plausible hypotheses on an in-distribution (ID) dataset, an OOD-generalizable hypothesis may appear. Moreover, the diversity of hypotheses lets us distinguish ID samples from OOD samples by measuring the degree of divergence in ensemble members' predictions. Despite conceptual benefits, diverse-ensemble training has previously remained a lab-bound concept for several reasons. First, previous approaches required a separate OOD dataset that may nurture diverse hypotheses. Second, computational complexities of previous pairwise diversification objectives increase quadratically with the ensemble size.

We have addressed the challenges through the novel Scalable Ensemble Diversification (SED) framework. SED identifies the OOD-like samples from a single dataset, bypassing the need to prepare a separate OOD dataset. SED also employs a stochastic pair selection algorithm which reduces the quadratic complexity of previous approaches to a constant cost per SGD iteration. We have demonstrated good performances by SED on the OOD generalization and detection tasks, both at the ImageNet scale, a largely underexplored regime in the ensemble diversification community. In particular, for OOD detection, our novel diversity measure of Predictive Diversity Score (PDS) amplifies the benefits of diverse ensembles for OOD detection. The code to reproduce the results of our experiments will provided with the next revision of the manuscript.

Limitations

We do not provide theoretical justification for the method. Our experiments were conducted on models with a frozen feature extractor.

References

[1] H. L. Benoit, L. Jiang, A. Atanov, O. F. Kar, M. Rigotti, and A. Zamir. Unraveling the key components of OOD generalization via diversification. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Lvf7GnaLru`.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996. ISSN 1573-0565. doi: 10.1007/BF00058655. URL `https://doi.org/10.1007/BF00058655`.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

[4] A. S. Chen, Y. Lee, A. Setlur, S. Levine, and C. Finn. Project and probe: Sample-efficient domain adaptation by interpolating orthogonal features. *arXiv preprint arXiv:2302.05441*, 2023.

[5] F. D'Angelo and V. Fortuin. Repulsive deep ensembles are bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465, 2021.

[6] A. de Mathelin, F. Deheeger, M. Mougeot, and N. Vayatis. Maximum weight entropy. *arXiv preprint arXiv:2309.15704*, 2023.

[7] A. de Mathelin, F. Deheeger, M. Mougeot, and N. Vayatis. Deep anti-regularized ensembles provide reliable out-of-distribution uncertainty quantification, 2023.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[9] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021.

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

[11] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[12] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. doi: 10.1109/34.58871.

[13] J. C. Helton, J. D. Johnson, and W. L. Oberkampf. An exploration of alternative approaches to the representation of uncertainty in model predictions. *Reliability Engineering & System Safety*, 85(1-3):39–71, 2004.

[14] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

[15] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Hkg4TI9xl`.

[16] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[17] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.

[18] T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

[19] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[20] R. Huang and Y. Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[21] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL `https://proceedings.neurips.cc/paper_files/paper/1994/file/b8c37e33defde51cf91e1e03e51657da-Paper.pdf`.

[22] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf`.

[23] Y. Lee, H. Yao, and C. Finn. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RVTOp3MwT3n.

[24] Z. Li, I. Evtimov, A. Gordo, C. Hazirbas, T. Hassner, C. C. Ferrer, C. Xu, and M. Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023.

[25] Y. Liu and X. Yao. Simultaneous training of negatively correlated neural networks in an ensemble. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(6):716–725, 1999.

[26] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[27] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[28] M. Pagliardini, M. Jaggi, F. Fleuret, and S. P. Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=K7CbYQbyYhY.

[29] A. Ross, W. Pan, L. Celi, and F. Doshi-Velez. Ensembles of locally independent prediction models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5527–5536, 2020.

[30] D. Teney, E. Abbasnejad, S. Lucey, and A. van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16761–16772, June 2022.

[31] D. Teney, M. Peyrard, and E. Abbasnejad. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 458–476, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20050-2.

[32] H. Touvron, M. Cord, and H. Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022.

[33] T. Trinh, M. Heinonen, L. Acerbi, and S. Kaski. Input-gradient space particle inference for neural network ensembles. In *International Conference on Learning Representations*, 2024.

[34] H. Wang and Q. Ji. Diversity-enhanced probabilistic ensemble for uncertainty estimation. In R. J. Evans and I. Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2214–2225. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/wang23c.html.

[35] H. Wang, Z. Li, L. Feng, and W. Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[36] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:6514–6527, 2020.

[37] A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

[38] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

[39] S. Yashima, T. Suzuki, K. Ishikawa, I. Sato, and R. Kawakami. Feature space particle inference for neural network ensembles. In *International Conference on Machine Learning*, pages 25452–25468. PMLR, 2022.

[40] S. Zaidi, A. Zela, T. Elsken, C. C. Holmes, F. Hutter, and Y. Teh. Neural ensemble search for uncertainty estimation and dataset shift. *Advances in Neural Information Processing Systems*, 34:7898–7911, 2021.

[41] T. Zhou, S. Wang, and J. A. Bilmes. Diverse ensemble evolution: Curriculum data-model marriage. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3070e6addcd702cb58de5d7897bfdae1-Paper.pdf.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to § 4

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to § 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper contains no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to § 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code will be available soon, please refer to § 4.1.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: please refer to § 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because their magnitude was below the rounding error or roughly around it for the majority of experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: please refer to § 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: we followed the Code to the best of our knowledge.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We believe that this work has no societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We believe that our paper does not pose such risks as we train models for ImageNet classification.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: we were unable to find the license for the dataset we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.