Measuring Ensemble Diversity and Its Effects on Model Robustness

Lena Heidemann[1] , **Adrian Schwaiger**[1] , **Karsten Roscher**[1]
[1]Fraunhofer IKS
{firstname.lastname}@iks.fraunhofer.de

Abstract

Deep ensembles have been shown to perform well on a variety of tasks in terms of accuracy, uncertainty estimation, and further robustness metrics. The diversity among ensemble members is often named as the main reason for this. Due to its complex and indefinite nature, diversity can be expressed by a multitude of metrics. In this paper, we aim to explore the relation of a selection of these diversity metrics among each other, as well as their link to different measures of robustness. Specifically, we address two questions: To what extent can ensembles with the same training conditions differ in their performance and robustness? And are diversity metrics suitable for selecting members to form a more robust ensemble? To this end, we independently train 20 models for each task and compare all possible ensembles of 5 members on several robustness metrics, including the performance on corrupted images, out-of-distribution detection, and quality of uncertainty estimation. Our findings reveal that ensembles trained with the same conditions can differ significantly in their robustness, especially regarding out-of-distribution detection capabilities. Across all setups, using different datasets and model architectures, we see that, in terms of robustness metrics, choosing ensemble members based on the considered diversity metrics seldom exceeds the baseline of a selection based on the accuracy. We conclude that there is significant potential to improve the formation of robust deep ensembles and that novel and more sophisticated diversity metrics could be beneficial in that regard.

1 Introduction

Deep Neural Networks (DNNs) are one of the key Machine Learning (ML) approaches in enabling high-impact applications such as autonomous driving or automated medical diagnoses. Convolutional Neural Networks (CNNs) are especially relevant for complex perception tasks and have shown impressive results. However, these networks have various insufficiencies that impede their use in safety-critical systems [Willers *et al.*, 2020]. For instance, it has been demonstrated that CNNs do not provide reliable uncertainty estimates for their predictions [Guo *et al.*, 2017; Henne *et al.*, 2020]. These are required for surrounding safety systems that dynamically decide if an ML component can be trusted in a given situation or if a safety action needs to be taken [Weiss *et al.*, 2018].

One widely used method to improve the reliability of uncertainty quantification are Deep Ensembles (DEs) [Lakshminarayanan *et al.*, 2017]. Their popularity stems from the ease of use — only multiple individual networks need to be trained and their predictions averaged — and their overall increase in performance and robustness, especially w.r.t. uncertainty quantification. This increase in robustness is due to the inherent randomness in the training of DNNs that causes the individual networks forming the DE to converge to different local minima [Fort *et al.*, 2020]. In other words, as with other ensemble learning approaches, e.g., gradient boosted trees, the diversity of the member models results in an overall increased performance and robustness [Beluch *et al.*, 2018]. However, measuring the diversity of DNNs is non-trivial, due to its indefinite specification, and as a consequence also the quantification of its effects on the robustness of DEs. In this paper, we therefore make the following contributions:

- We investigate the extent to which ensembles can differ w.r.t. performance and safety-relevant metrics when training their member models independently and with the same hyperparameters and architecture, but with different random initalizations.
- Furthermore, we investigate the correlation of safety-relevant metrics to different diversity metrics, in order to determine their applicability as indicators for the robustness of DEs.

2 Related Work

In the following, we present the current state of assuring ML systems from a safety engineering perspective, giving context how advances in the field of safe ML relate to it. Additionally, we illustrate current efforts in the field of uncertainty estimation for DNNs and elaborate especially on DE.

2.1 Safety Assurance for Machine Learning

Compared to other software-based systems, arguing the safety for ML systems is still an emerging field. For one, the lack of an exact specification of the function, the reason why ML is employed in the first place, impedes tractable testing and safety analyses. For the other, especially for DNNs, critical insufficiencies exist, most notably, unreliable confidence estimation, susceptibility to out-of-distribution data, and lack of generalizability and explainability [Willers *et al.*, 2020]. These insufficiencies require appropriate countermeasures, e.g., in the form of explicit uncertainty estimation, out-of-distribution detection, or methods enabling interpretability. Surrounding safety architectures can then incorporate these countermeasures to determine the trustworthiness of the ML function at runtime [Weiss *et al.*, 2018]. PURSS, for instance, integrates perceptual uncertainty quantification in the Responsibility-Sensitive Safety model, a safety approach based on formal rules and physical constraints [Salay *et al.*, 2020]. However, for complex autonomous systems such safety architectures do not only require the quantification of uncertainties for DNNs, but also of all other elements of the (perception) system [Kurzidem *et al.*, 2020]. This means that approaches improving the ML functions will not be sufficient on their own, it is also necessary to approach the problem from the safety engineering side. For instance, existing safety standards, such as ISO26262 in the automotive domain, must be analyzed towards their applicability to systems employing ML functions [Salay *et al.*, 2018]. Moreover, new assurance approaches and standards need to be formulated to take into account the specifics of ML. One such approach is to structure the assurance case in Goal Structuring Notation by performance claims, for which limitations are analyzed and concrete qualitative and quantitative evidences are gathered [Burton *et al.*, 2019]. This framework allows a structured analysis of ML-based systems and supports developing best practices to facilitate future assurance efforts.

2.2 Uncertainty Quantification for DNNs

DNNs are known to be overconfident in their predictions [Guo *et al.*, 2017] which impedes their use in safety-critical systems. To overcome this insufficiency, the most prominent approach is to include Bayesian principles in DNNs in the form of Bayesian neural networks [Bishop, 1997]. These model and learn distributions over the weights of a network. For inference, the weights are sampled from their respective distributions, resulting in a predictive mean and variance. Although, Bayesian neural networks provide more reliable confidence estimates, their training still remains a challenge, as computing the posterior parameter distribution for deep networks is currently intractable [Mullachery *et al.*, 2018]. An approach to approximate Bayesian neural networks is Monte Carlo dropout [Gal and Ghahramani, 2016]. Here, dropout is used at inference to sample from the weights, i.e., applying random dropout masks is interpreted as placing a Bernoulli distribution over the weights. Although the sampling yields a more reliable predictive mean and variance, the applicability in low-power domains is limited, as multiple (partial) forward passes are required. A more efficient approach is Evidential Deep Learning, a non-Bayesian approach based on the Dempster-Shafer theory, that learns to estimate the parameters of a predictive Dirichlet distribution [Sensoy *et al.*, 2018]. However, the training may suffer from instabilities, complicating arguments towards its robustness [Henne *et al.*, 2020]. Recently, another approach has been proposed, arguing that standard deterministic DNNs with appropriate inductive biases are able to outperform more complex uncertainty quantification approaches in active learning and out-of-distribution detection, requiring only minimal changes to the architecture and training procedure [Mukhoti *et al.*, 2021].

Despite recent advances in the field of uncertainty quantification for DNNs, DE is still a very popular approach, as it consistently improves the reliability of uncertainty estimates and robustness as well as performance in general. It works by forming an ensemble over multiple independently trained DNNs and averages the predictions of the individual members to capture the predictive mean and variance. As with other ensembling approaches in ML, the reason for the increased robustness and performance lies in the diversity of the member networks [Beluch *et al.*, 2018], a result of the randomness in the weight initialization and optimization process of DNNs [Fort *et al.*, 2020]. Therefore, a few approaches have been proposed to increase the diversity of DEs. For instance, Pang et al. [2019] introduce an adaptive diversity promoting regularizer to increase the robustness against adversarial attacks. Another example is the diversity-promoting adversarial loss proposed by Sinha et al. [2020] that improves the overall robustness of DEs. Although DE is a sampling-based approach, it requires significantly fewer forward passes than Monte Carlo dropout, while outperforming it [Henne *et al.*, 2020]. Furthermore, with techniques such as Ensemble Distribution Distillation [Malinin *et al.*, 2020], a DE can be distilled into a single model, while maintaining its predictive qualities.

3 Diversity Metrics for Ensembles of Deep Neural Networks

This section introduces the ensemble diversity metrics we used in the experiments for this paper. We focused on pairwise metrics, i.e., metrics which are evaluated on a pair of classifiers. The diversity of an ensemble of more than two members is then calculated as the mean pairwise diversity metric value. In the following, let $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$ denote the dataset, where $y_n \in \{1, \dots, K\}$ is the true label out of K classes for the input $\mathbf{x}_n \in \mathbb{R}^D$. Furthermore, let G_{θ_i} be a deep neural network with parameters $\theta_i \in \mathbb{R}^p$ which models a predictive distribution over the labels, denoted by $p_{\theta_i}(y|\mathbf{x})$. The predicted label is defined as $\hat{y}_i = \arg\max_k p_{\theta_i}(y = k|\mathbf{x})$.

3.1 Disagreement

Disagreement of predictions is a common and straightforward measure of diversity. It is defined as the fraction of samples in the dataset $\mathcal{D}$ for which two classifiers predict different labels. The disagreement between two classifiers G_{θ_i} and G_{θ_j} can be expressed as

$$D_{i,j} = \frac{N^{\hat{y}_i \neq \hat{y}_j}}{N}, \tag{1}$$

where $N^{\hat{y}_i \neq \hat{y}_j}$ is the number of samples in $\mathcal{D}$ on which the two classifiers disagree.

3.2 Normalized Disagreement

Taking into account that the disagreement could stem from making random predictions, Fort et al. [2020] have introduced a normalized version of the disagreement metric which includes the accuracy of the classifiers. This has been widely adopted as a measure of ensemble diversity [Wen *et al.*, 2020; Durasov *et al.*, 2020; Wenzel *et al.*, 2021]. The normalized disagreement between G_{θ_i} and G_{θ_j} is defined as

$$ND_{i,j} = \frac{D_{i,j}}{(1-a)} = \frac{N^{\hat{y}_i \neq \hat{y}_j}}{N(1-a)}, \tag{2}$$

where a denotes the accuracy of the ensemble of the two classifiers and $D_{i,j}$ is the disagreement as defined in Equation (1).

3.3 Double Fault Measure

Another example of incorporating accuracy or error into a diversity metric is the double fault measure. It is one of the classical measures of similarity, which has long existed in the literature and has already been used to select the most error-independent nets for ensembling [Giacinto and Roli, 2001; Kuncheva and Whitaker, 2003]. With N^{00} denoting the number of samples for which both classifiers make a wrong prediction, the double fault measure between G_{θ_i} and G_{θ_j} can be described as

$$DF_{i,j} = \frac{N^{00}}{N}. \tag{3}$$

3.4 Output Correlation

Besides using the model's predicted labels, diversity can also be measured in terms of the predictive distribution. Huang et al. [2017], for example, have used the pairwise correlation of softmax outputs for describing diversity in the activation space. The output correlation between two classifiers G_{θ_i} and G_{θ_j} is then defined as the Pearson correlation coefficient of $p_{\theta_i}(y|\mathbf{x})$ and $p_{\theta_j}(y|\mathbf{x})$ on dataset $\mathcal{D}$.

3.5 Cosine Similarity

The four preceding diversity metrics are calculated based on the output of the models and therefore are dependent on the dataset $\mathcal{D}$. Contrary to this, the cosine similarity is applied to the parameters θ_i and θ_j of the two classifiers, e.g., as it was used by Fort et al. [2020]. It is defined as

$$CS_{i,j} = \cos(\theta_i, \theta_j) = \frac{\theta_i^\top \theta_j}{\|\theta_i\| \|\theta_j\|}. \tag{4}$$

3.6 Other Diversity Metrics

The presented metrics are only a selection of the diversity metrics available for classifier ensembles. Kuncheva and Whitaker [2003] give an overview of metrics measuring output diversity. Out of those, we also used the Q statistic, the disagreement measure, and the Cohen's kappa coefficient (κ) in our evaluation. However, in accordance with the results of Kuncheva and Whitaker [2003], these were highly correlated among themselves and with disagreement. Especially κ has a strong negative correlation with disagreement. It is a measure of agreement which also takes into account the probability of random agreement. For classifiers with a high number of classes, this probability is negligible and κ reduces to a simple agreement metric. Due to these strong correlations and for the sake of brevity and clarity, we omitted these metrics in our results.

4 Evaluation

In the following, we discuss our results on the task of image classification, evaluating the significance of the diversity metrics regarding safety-relevant metrics.

4.1 Design of Experiments

In order to provide a broad analysis on ensemble diversity, we trained three different architectures for our experiments: A basic 4-layer CNN (BasicCNN), a MobileNetV2 [Sandler *et al.*, 2018] for its suitability for mobile applications, and a 34-layer ResNet [He *et al.*, 2016] as a standard architecture. Within a deep ensemble, we only use one type of architecture in accordance with Lakshminarayanan et al. [2017]. The datasets used for training are CIFAR-10, CINIC-10 [Darlow *et al.*, 2018], and German Traffic Sign Recognition Benchmark (GTSRB) [Stallkamp *et al.*, 2011]. The CIFAR-10 dataset consists of 60,000 32×32px images in 10 classes, e.g., automobile, truck or dog. CINIC-10 is an extension of CIFAR-10 with downsampled ImageNet images, adding up to 270,000 images in the same 10 classes. The more than 50,000 images of GTSRB show German traffic signs, which are to be classified into 43 different classes. For evaluating the ability to detect Out-of-Distribution (OOD) inputs, we used CIFAR-100 and the Street View House Numbers (SVHN) dataset [Netzer *et al.*, 2011]. CIFAR-100 is similar to CIFAR-10, but with 100 classes, while SVHN comprises real-world images of house numbers.

For each architecture and dataset, we independently trained 20 models on the respective training data. No augmentations were applied to the data. Each model's parameters were randomly initialized and optimized w.r.t. the negative log likelihood loss using the Adam optimizer. To prevent overfitting, we applied early stopping, when the validation loss did not decrease for several epochs, and chose the model with the lowest validation loss for our analysis.

4.2 Evaluation Metrics

For the evaluation we use accuracy, the fraction of correctly classified samples, as well as the Remaining Accuracy Rate (RAR) [Henne *et al.*, 2020], which describes the trade-off between performance and safety. Using a threshold of predicted probability for discarding predictions as uncertain, RAR describes the remaining accuracy after discarding predictions below that threshold. In a well-performing but safe system, RAR should be as high as possible, while the number of certain but incorrect samples should be kept to a minimum. In our analysis we use RAR at an error rate of 1%, i.e., at the threshold where 1% of the samples are labeled as certain, but the predictions are incorrect. We refer to this metric as Acc@1%. To further evaluate the quality of uncertainty estimation, we calculate the Expected Calibration Error (ECE). It

groups the probabilities of a set of predictions into bins, each covering an equally-sized interval of probabilities, and takes the average absolute difference between the accuracy and the predicted probability of each bin. Given that overconfidence has a higher relevance for safety-critical applications, we also report a calibration error which only includes the error due to overconfidence, referred to here as the Negative Expected Calibration Error (NECE).

For OOD detection the In-Distribution (ID) test dataset is combined with the OOD test dataset. The trained network labels samples as ID, when its predictive probability is above a certain threshold, while the remaining samples are labeled as OOD. For this task we evaluate the Area Under the Receiver Operating Characteristic (AUROC), as well as the False Positive Rate (FPR) at 95% True Positive Rate (TPR). This metric describes the fraction of OOD samples wrongly classified as ID at a threshold where 95% of ID samples are correctly classified as ID. Throughout the paper we refer to this metric as FPR95.

4.3 Results and Discussion

For each setup we evaluate all 15504 possible ensembles of a combination of 5 members out of the 20 trained networks on a separate test set. The number of ensemble members is set to 5 since larger ensembles start to show diminishing returns [Lakshminarayanan *et al.*, 2017], as well as reduce the number of possible ensemble combinations for evaluation. The diversity of an ensemble is computed by taking the average of the respective pairwise diversity metric over all 10 possible pairs within the 5 ensemble members. Similarly, we also compute the average accuracy of all 2-member combinations out of the 5 available members. Selecting ensemble members based on accuracy is the evident and simple approach. We therefore use this pairwise accuracy as a baseline metric for a selection of ensemble members and denote it as the baseline throughout the paper. The results are presented in three parts: an evaluation on ID data, on corrupted data, and on the detection of OOD data.

In-Distribution Data

We first show the extent to which the evaluation metrics of all possible ensembles may differ, when selecting 5 out of 20 trained networks. A large variance in these metrics indicates that the ensembles may vary considerably in their performance or robustness, although their members were trained under the same conditions except for random initialization. In these cases, the potential benefit of an informed selection of ensemble members is the largest.

Table 1 shows the minimum and maximum values of the accuracy (Acc), Acc@1%, and the ECE, for all configurations and corruptions. The distribution of the metrics in this table usually follows a bell-shaped curve, i.e., most values are close to the mean of the minimum and maximum values. In this part, we focus on the first row of the table, i.e., the ID data without any corruptions. Depending on the choice of ensemble members, the accuracy may vary up to 2.1 percentage points, Acc@1% up to 8.3, and ECE up to 3.7 percentage points. The metric values for CIFAR-10 and CINIC-10 generally have a higher variation than these for GTSRB, presumably due to the high level of accuracy for GTSRB to begin with. The differences between the datasets of the variations in ECE are not as pronounced. For all three metrics, there is varying but apparent potential for improvement by an aimed selection of ensemble members.

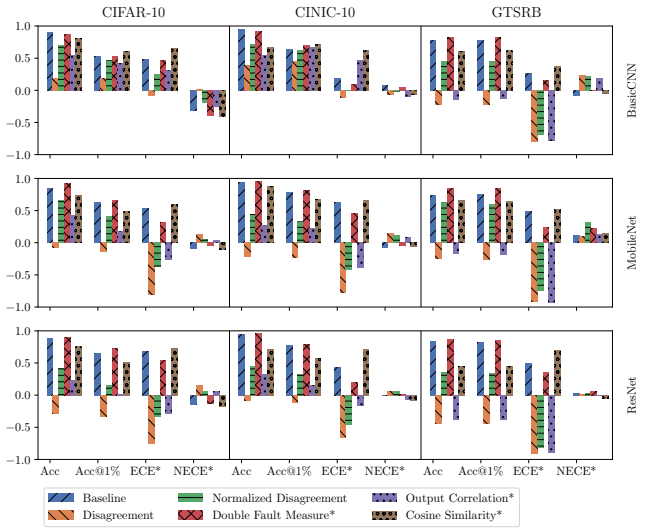

Figure 1: Correlation coefficients between diversity and evaluation metrics on the original ID dataset. An asterisk (*) indicates a reversed sign of the correlation coefficient for better readability. Thereby, to pose as a suitable indicator for ensemble robustness, a diversity metric should correlate most positively with Acc, Acc@1%, ECE* and NECE*.

Next, we evaluate if and which diversity metric can tap most of this potential by serving as selection metric. To this end, we also computed the diversity metric values for each possible ensemble and analyzed the correlation between them and the evaluation metrics. The resulting correlation coefficients are depicted in Figure 1. Since some diversity metrics are a measure of similarity, i.e., the lower the metric value, the more diverse the ensemble, we reversed the sign of these relations for the correlation plot. We did the same for correlation coefficients of evaluation metrics for which a lower value indicates better performance or robustness, like ECE and NECE. Therefore, a positive correlation in this figure uniformly describes that the more diverse the ensemble is according to this diversity metric, the better it performs in terms of the respective evaluation metric, and vice versa. Any sign reversal is indicated by an asterisk (*).

Except for disagreement and output correlation, most metrics positively correlate with both, accuracy and Acc@1%. The baseline, double fault measure, and cosine similarity stand out, as they additionally correlate positively with ECE* across all datasets and architectures. For NECE* the correlations are less strong and in some cases almost negligible. Only for BasicCNNs trained on CIFAR-10, the correlations are quite high and mostly negative. Considering the sign reversal, this means that the higher the diversity, or accuracy, for an ensemble according to that metric, the higher the NECE and therefore the stronger the overconfidence. In

		CIFAR-10			CINIC-10			GTSRB		
		BasicCNN	MobileNet	ResNet	BasicCNN	MobileNet	ResNet	BasicCNN	MobileNet	ResNet
None	Acc	81.75	85.98	**86.22**	70.64	77.09	77.56	98.78	98.39	98.20
		83.52	88.07	**88.35**	71.54	79.00	79.47	99.23	99.59	99.42
	Acc@1%	47.91	**54.50**	55.41	28.79	36.59	38.33	98.71	97.95	97.84
		53.05	**62.81**	63.23	30.76	42.64	43.36	99.23	99.59	99.42
	ECE	1.04	3.66	3.02	1.26	3.55	**2.11**	0.32	1.07	0.59
		2.81	6.87	6.41	2.92	6.31	**5.81**	0.94	3.26	2.89
Brightness	Acc	77.40	79.03	80.26	66.65	71.74	72.03	96.79	95.55	95.50
		79.53	83.85	83.66	68.02	74.91	75.14	97.71	97.60	97.55
	Acc@1%	37.78	39.60	41.29	23.05	29.43	29.25	94.40	91.84	91.35
		42.98	50.94	51.63	25.44	35.34	35.56	96.47	96.20	96.41
	ECE	1.01	3.40	2.86	1.75	4.17	2.74	0.47	2.04	0.56
		2.80	8.11	7.22	3.73	7.84	6.49	1.14	5.87	4.43
Contrast	Acc	61.23	64.02	**68.49**	50.69	61.00	64.02	97.21	94.81	95.54
		65.51	70.73	**76.62**	53.38	65.98	69.46	98.23	98.23	98.12
	Acc@1%	15.02	17.12	23.74	9.74	17.84	20.62	95.06	90.89	92.19
		21.35	27.10	35.18	12.13	22.88	27.26	97.53	97.66	97.52
	ECE	3.39	0.88	0.92	3.84	1.18	0.81	1.21	2.44	1.21
		7.97	4.48	5.45	6.49	5.68	5.88	2.32	6.34	6.32
Cutout	Acc	69.96	77.46	77.97	63.38	71.24	72.14	95.34	94.70	94.71
		73.08	81.72	82.64	64.90	74.07	75.56	96.07	96.52	96.38
	Acc@1%	25.94	**33.17**	36.61	20.41	28.38	29.37	91.27	89.68	89.64
		32.97	**49.21**	49.45	22.34	33.58	35.69	93.05	93.07	93.63
	ECE	1.56	2.28	2.11	0.49	3.39	1.59	0.44	1.37	0.68
		4.22	6.92	6.69	2.31	6.81	6.45	1.29	4.89	3.66

Table 1: The minimum and maximum values of accuracy (Acc), Acc@1%, and ECE on different types of corruptions for all ensembles of 5 out of 20 trained networks. Values mentioned in the text are highlighted in **bold**.

all other cases there is little correlation with NECE, which is presumably due to the already little variance of NECE between the ensembles.

Corruptions

We further evaluate all ensembles on corrupted images in order to test their robustness. We applied three types of corruptions: brightness and contrast, according to the framework by Hendrycks and Dietterich [2019], and cutout, which randomly cuts a patch of 8×8px from an image (see Figure 2).

We return to Table 1 for describing the variance of the evaluation metrics on corrupted input images between all 5-member ensembles. The range of values increases for all corruptions compared to the results on the original data. The highest variation in accuracy and Acc@1% can be observed for CIFAR-10. For ensembles trained on this dataset, accuracy may vary up to 8.1 percentage points, while Acc@1% shows a range of up to 16.0 percentage points. Although

Figure 2: Exemplary 32×32px image[1] with different types of corruption.

mostly to a lesser extent, we can also find an increase in variation for ECE values compared to no corruptions. Therefore, an informed selection of ensemble members could potentially lead to significant improvement of the performance on corrupted input data.

For our correlation analysis, we again consider the accuracy, Acc@1%, ECE, and NECE, but evaluated on the corrupted dataset. For brightness corruptions, the correlations between diversity and evaluation metrics take a very similar form to applying no corruptions. Mostly, the only difference is a generally weaker correlation. We can therefore assume that this type of corruption has a more or less evenly distributed effect on the performance of the ensembles and therefore their correlation with the diversity metrics.

Other types of corruptions, however, have a stronger influence on the correlations. Figure 3 shows the correlation coefficients of diversity with evaluation metrics on the contrast corrupted dataset. Similarly to brightness corruptions, we observe a weaker correlation in most cases. Additionally, for MobileNet and ResNet, also the sign of correlation is mostly in accordance with the results from the original dataset. For BasicCNN, however, the ECE* on the corrupted dataset is negatively correlated with almost all diversity metrics for CIFAR-10 and CINIC-10. Furthermore, there is almost no correlation with accuracy for CINIC-10 and GTSRB. Overall and especially for MobileNet and ResNet, we can

[1] Image source: https://commons.wikimedia.org/wiki/File:PH-ALW_Special_Air_Services_B.V.V.JPG

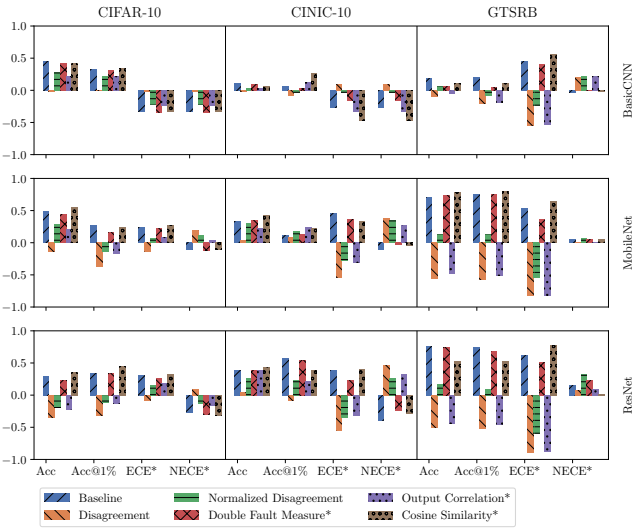

Figure 3: Correlation coefficients between diversity and evaluation metrics on the contrast corrupted dataset. An asterisk (*) indicates a reversed sign of the correlation coefficient for better readability.

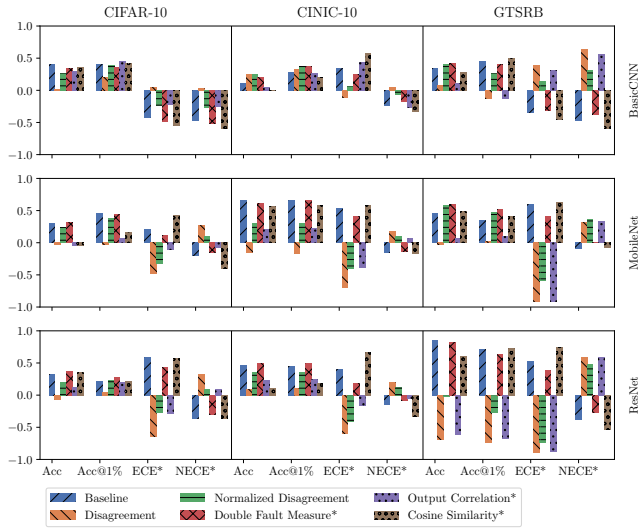

Figure 4: Correlation coefficients between diversity and evaluation metrics on the cutout corrupted dataset. An asterisk (*) indicates a reversed sign of the correlation coefficient for better readability.

still distinguish the baseline, double fault measure, and cosine similarity from the other diversity metrics. In most cases, they are positively correlated with accuracy, Acc@1%, and ECE*, and at most have a weak negative correlation with NECE*. On the other hand, while normalized disagreement was often close to the top three in the original dataset, its correlation with the evaluation metrics on contrast corrupted data is considerably weaker. In some cases, it is even negatively correlated with accuracy.

Lastly, the correlations for the evaluation metrics on the cutout corrupted dataset are depicted in Figure 4. For this type of corruption, the correlations are again weaker compared to the evaluation metrics on the original dataset, but also less consistent across all setups. In some cases for BasicCNN, the baseline, double fault measure, and cosine similarity are negatively correlated with ECE* and have a stronger negative correlation with NECE*. Accuracy and Acc@1% are still mostly positively correlated with these three selection metrics, but less distinguishable and to a lesser extent, especially for cosine similarity. Although the evaluation metrics on the corrupted datasets are mainly less correlated with the diversity metrics, the correlations for the baseline, double fault measure, and cosine similarity mostly point into the right direction in terms of accuracy, Acc@1%, and ECE.

Out-of-Distribution Data

Next to the robustness to corrupted data, we also look into robustness to OOD data and its link to ensemble diversity. We start again by inspecting the extent to which the evaluation metrics may vary across ensembles. Figure 5 shows the distributions of AUROC and FPR95 values for OOD detection on CIFAR-100 and SVHN for all possible ensembles of 5 members out of 20. For OOD detection on CIFAR-100, ResNets trained on GTSRB stand out for their large range of AUROC and especially FPR95 values. On the other hand, ensembles of models trained on GTSRB only vary little in AUROC and FPR95 for OOD detection on SVHN. MobileNets and ResNets trained on CIFAR-10 and CINIC-10, however, show a large variation. AUROC values can differ up to 15.4 percentage points, and FPR95 values show a range of up to 43.0 percentage points. Therefore, the most potential for improvement by an aimed selection of ensemble members can be observed for FPR95 on CIFAR-100 for ResNets trained on GTSRB, and for AUROC and FPR95 on SVHN for MobileNets and ResNets trained on CIFAR-10 as well as CINIC-10.

The correlation between these metrics and the diversity metrics are presented in Figure 6. The baseline, double fault measure, and cosine similarity mostly correlate positively with AUROC and FPR95*. Only for models trained on CINIC-10 and MobileNets trained on GTSRB, correlations are generally low and can have reversed orientations. While for models trained on CIFAR-10 other diversity metrics also show a positive correlation with AUROC and FPR95*, they are weaker than those of the baseline, double fault measure, and cosine similarity. The distinction is most clear for BasicCNNs and ResNets trained on GTSRB. For ResNets on GTSRB there is also a lot to gain from strong correlations, since their AUROC, but especially FPR95 values, may vary significantly, as shown in Figure 5. For instance, the most diverse ensemble according to the cosine similarity metric has an FPR95 of 0.78% on CIFAR-100, while the most similar one has an FPR95 of 15.81%.

For OOD detection, we can sometimes observe large variations between ensembles in AUROC and FPR95 values, although the ensemble members were trained under the same conditions except for random initialization. Furthermore, out of all diversity metrics, the baseline, double fault measure, and cosine similarity correlate particularly well with AUROC

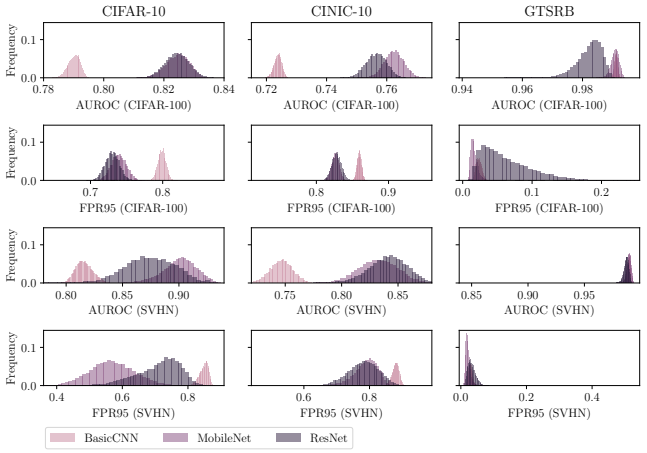

Figure 5: Distribution of evaluation metrics for OOD detection for all ensembles of 5 out of 20 trained networks. Each column indicates the ID dataset, the respective OOD dataset is displayed in parentheses.

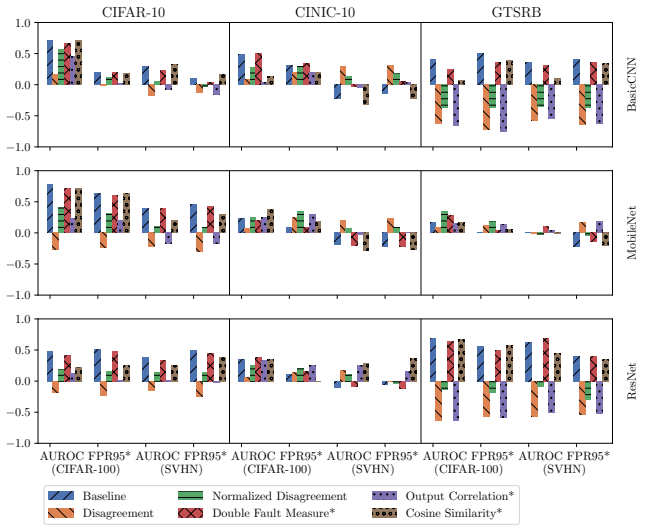

Figure 6: Correlation coefficients between diversity and evaluation metrics for OOD detection on CIFAR-100 and SVHN. An asterisk (*) indicates a reversed sign of the correlation coefficient for better readability. Ideally, a diversity metric correlates most positively with AUROC and FPR95*.

and FPR95, most evidently for ResNets trained on GTSRB.

Summary

We can see that there is varying potential benefit by selecting ensemble members based on diversity metrics. Disagreement mostly failed as a selection metric by oftentimes correlating with the evaluation metrics in the unfavorable direction. Output correlation performed similarly, albeit slightly better than disagreement. By taking accuracy into account, the normalized disagreement correlated more positively with accuracy, while still showing similar patterns to disagreement for ECE and NECE. The baseline, double fault measure, and cosine similarity performed reasonably well across all tasks and were most distinguishable from the other diversity metrics, but without a clear winner. Double fault measure is most similar to the baseline metric, the pairwise accuracy, as it is a measure of error. This is also reflected in the quite similar correlations with the evaluation metrics. Cosine similarity, being a measure of parameter diversity, has little in common with the baseline metric and double fault measure. Nevertheless, it correlates similarly with the evaluation metrics. This indicates the potential of parameter diversity measures, although the cosine similarity of all parameters is a rather simple form.

5 Conclusions and Future Work

In this paper, we investigated the variance in robustness of equally trained ensembles and if diversity metrics are suitable to indicate ensemble robustness. Our findings show, that ensembles trained under the same conditions can vary significantly w.r.t. performance and robustness. Especially for the task of OOD detection we observed differences between the individual ensembles of up to 43 percentage points for the false positive rate. Regarding the question whether diversity metrics are suitable indicators for ensemble robustness, we found that disagreement, normalized disagreement, and output correlation are not well suited. Cosine similarity and double fault measure, on the other hand, show a high correlation with the robustness metrics. However, they do not perform better than our baseline of selecting an ensemble based on the pairwise accuracy of the member networks.

As we see a significant potential in improving the formation of robust DEs, we suggest multiple directions for future work. Regarding diversity metrics, we suggest the design of metrics that consider the specifics of DNNs. For instance, existing metrics do not consider the semantic information and interplay of the individual neurons. From a safety perspective, diversity metrics should incorporate this information, as it is more important that the member networks base their decisions on different concepts than that they show a difference in their output, but based on the same evidences. Furthermore, this should be incorporated in approaches increasing the diversity of ensembles, ensuring that each member network is not susceptible to the same error patterns.

Acknowledgments

This work was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems.

References

[Beluch *et al.*, 2018] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The Power of Ensembles for Active Learning in Image Classification. In *Proc. CVPR 2018*, pages 9368–9377, 2018.

[Bishop, 1997] Christopher M. Bishop. Bayesian Neural Networks. *Journal of the Brazilian Computer Society*, 4(1), July 1997.

[Burton *et al.*, 2019] Simon Burton, Lydia Gauerhof, Bibhuti Bhusan Sethy, Ibrahim Habli, and Richard Hawkins. Confidence Arguments for Evidence of Performance in Machine Learning for Highly Automated Driving Functions. In Alexander Romanovsky, Elena Troubitsyna, Ilir Gashi, Erwin Schoitsch, and Friedemann Bitsch, editors, *Computer Safety, Reliability, and Security*, Lecture Notes in Computer Science, pages 365–377, Cham, 2019. Springer International Publishing.

[Darlow *et al.*, 2018] Luke N. Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. CINIC-10 is not ImageNet or CIFAR-10. *arXiv:1810.03505 [cs, stat]*, October 2018.

[Durasov *et al.*, 2020] Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for Uncertainty Estimation. *arXiv:2012.08334 [cs]*, December 2020.

[Fort *et al.*, 2020] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective. *arXiv:1912.02757 [cs, stat]*, June 2020.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML 2016*, volume 48, pages 1050–1059. PMLR, June 2016.

[Giacinto and Roli, 2001] Giorgio Giacinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, August 2001.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proc. ICML 2017*, pages 1321–1330. JMLR.org, August 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. CVPR 2016*, pages 770–778, 2016.

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations. *arXiv:1807.01697 [cs, stat]*, April 2019.

[Henne *et al.*, 2020] Maximilian Henne, Adrian Schwaiger, Karsten Roscher, and Gereon Weiss. Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics. In Huáscar Espinoza, José Hernández-Orallo, Xin Cynthia Chen, Seán S. ÓhÉigeartaigh, Xiaowei Huang, Mauricio Castillo-Effen, Richard Mallah, and John McDermid, editors, *Proc. SafeAI@AAAI 2020*, volume 2560 of *CEUR Workshop Proceedings*, pages 83–90, 2020.

[Huang *et al.*, 2017] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot Ensembles: Train 1, get M for free. In *Proc. ICLR 2017*, March 2017.

[Kuncheva and Whitaker, 2003] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2):181–207, May 2003.

[Kurzidem *et al.*, 2020] Iwo Kurzidem, Ahmad Saad, and Philipp Schleiß. A Systematic Approach to Analyzing Perception Architectures in Autonomous Vehicles. In *Proc. IMBSA 2020*, pages 149–162, 2020.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Proc. NeurIPS 2017*, 30:6402–6413, 2017.

[Malinin *et al.*, 2020] Andrey Malinin, Bruno Mlodozeniec, and Mark J. F. Gales. Ensemble distribution distillation. In *Proc. ICLR 2020*, 2020.

[Mukhoti *et al.*, 2021] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deterministic Neural Networks with Appropriate Inductive Biases Capture Epistemic and Aleatoric Uncertainty. *arXiv:2102.11582 [cs, stat]*, February 2021.

[Mullachery *et al.*, 2018] Vikram Mullachery, Aniruddh Khera, and Amir Husain. Bayesian Neural Networks. *arXiv:1801.07710 [cs, stat]*, January 2018.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[Pang *et al.*, 2019] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving Adversarial Robustness via Promoting Ensemble Diversity. In *Proc. ICLR 2019*, pages 4970–4979. PMLR, May 2019.

[Salay *et al.*, 2018] Rick Salay, Rodrigo Queiroz, and Krzysztof Czarnecki. An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software. SAE Technical Paper 2018-01-1075, SAE International, Warrendale, PA, April 2018.

[Salay *et al.*, 2020] Rick Salay, Krzysztof Czarnecki, Maria Soledad Elli, Ignacio J. Alvarez, Sean Sedwards, and Jack Weast. PURSS: Towards Perceptual Uncertainty Aware Responsibility Sensitive Safety with ML. In *Proc. SafeAI@AAAI 2020*, volume 2560 of *CEUR Workshop Proceedings*, pages 91–95, 2020.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proc. CVPR 2018*, pages 4510–4520, 2018.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty. In *Proc. NeurIPS 2018*, pages 3179–3189. Curran Associates, Inc., 2018.

[Sinha *et al.*, 2020] Samarth Sinha, Homanga Bharadhwaj, Anirudh Goyal, Hugo Larochelle, Animesh Garg, and Florian Shkurti. Diversity inducing Information Bottleneck in Model Ensembles. *arXiv:2003.04514 [cs, stat]*, December 2020.

[Stallkamp *et al.*, 2011] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification

competition. In *Proc. IJCAI 2011*, pages 1453–1460, July 2011.

[Weiss *et al.*, 2018] Gereon Weiss, Philipp Schleiss, Daniel Schneider, and Mario Trapp. Towards integrating undependable self-adaptive systems in safety-critical environments. In *Proc. SEAMS 2018*, SEAMS '18, pages 26–32, New York, NY, USA, May 2018. Association for Computing Machinery.

[Wen *et al.*, 2020] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning. In *Proc. ICLR 2020*, Addis Ababa, Ethiopia, April 2020.

[Wenzel *et al.*, 2021] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter Ensembles for Robustness and Uncertainty Quantification. In *Proc. NeurIPS 2020*, virtual, January 2021.

[Willers *et al.*, 2020] Oliver Willers, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks. In *Proc. SAFECOMP Workshops 2020*, Lisbon, Portugal, January 2020.