Assessing the Impact of Diversity on the Resilience of Deep Learning Ensembles: A Comparative Study on Model Architecture, Output, Activation, and Attribution

Rafael Rosales, Pablo Munoz, Michael Paulitsch
Intel Labs, Intel Corporation
{rafael.rosales, pablo.munoz, michael.paulitsch}@intel.com

Abstract

We investigate the relationship between different diversity metrics, accuracy, and resiliency to natural image corruptions of Deep Learning (DL) image classifier ensembles. We evaluate existing diversity dimensions such as model architecture, model prediction, and neuron activations, as well as a novel diversity dimension of input attribution.

Using ResNet50 as a comparison baseline, we evaluate the resiliency of multiple individual DL model architectures against dataset distribution shifts corresponding to natural image corruptions. We compare ensembles created with diverse model architectures trained either independently or through a Neural Architecture Search technique and evaluate the correlation of prediction-based and attribution-based diversity to the final ensemble accuracy.

Finally, we evaluate a set of diversity enforcement heuristics for training based on negative correlation learning (NCL) and compare how effective they are to achieve independent failure behavior.

Our key observations are: 1) model architecture is more important for individual resiliency than model size or model accuracy but architecture diversity in an ensemble is typically not more resilient, 2) attribution-based diversity is less negatively correlated to the ensemble accuracy than prediction-based diversity, 3) a balanced loss function of individual and ensemble accuracy creates more resilient ensembles for image natural corruptions, 4) architecture diversity produces more diversity than NCL in all explored diversity metrics: predictions, attributions, and activations.

1. Introduction

In the context of Deep Learning (DL), it has been empirically discovered that the use of ensembles can improve the model's accuracy in tasks such as regression and classification. It has been speculated [15] that the main reason behind these improvements is the implicit diversity in the solutions found that when aggregated as an ensemble obtain better predictions. In this work, we evaluate the resiliency of diverse deep learning classifiers to out-of-distribution data due to natural image corruptions and the role that the different kinds of diversity play in improving it.

1.1. The case for design diversity

Design diversity [25] is a technique to increase the resilience of safety critical systems. It is established as a best practice in standards such as in vehicle functional safety [20] to prevent dependent failures, safety of the intended functionality [21] to address system limitations of machine-learning-based components, and avionic software [14].

A common pitfall is to misunderstand *independent development* as *design diversity*. In [49] the designers of a safety-critical system preferred to let multiple teams collaborate, although the purpose of having multiple teams is to produce multiple designs of a single specification. This was justified with the claim that specification problems can be better mitigated with such collaboration but at the cost of the sought *independence*.

The key problem is, that independent development can (and will) produce designs with common failures mainly due to the fact that independent designs do not enforce diverse design choices. In fact, it has been statistically proven that independently developed software results in dependent failure behavior on randomly selected inputs [13].

In [24], it has been shown that what is needed to reduce dependent failure behavior is diversity in design *choices*. If the choices are made satisfying certain properties, it can be expected (in the average case) to obtain negatively correlated failure behavior, i.e., better than independent.

In DL, however, design choices are not made by the human designers explicitly but are a result of the architecture, data, and optimization approach. Furthermore, existing diversity metrics are not directly related to the DL model's *design choices* and are known to have a diversity-accuracy

trade-off [23]. In [1], the resiliency benefit of ensembles in out-of-distribution inference is challenged, but the study only considers independently created members, so the members are expected to show dependent failure behavior.

1.2. Main research questions

To understand to what extent different kinds of diversity in a DL ensemble impact its accuracy and generalization to out-of-distribution data (natural image corruptions), we aim to answer the following research questions:
RQ1: For a single model, is model accuracy, size, or architecture the main explanatory factor of resilience against natural image corruptions?
RQ2: In an ensemble, can an attribution-based diversity metric improve the known accuracy-diversity trade-off?
RQ3: In an ensemble, which diversity enforcement heuristic produces the most resilient models?
RQ4: How diverse are the predictions, activations, and input feature attributions of models created with a diversity enforcement learning approach?

The rest of the paper is structured as follows: Section 2 provides a brief overview of the current state-of-the-art in diversity enforcement and measurement. Next, the methodology is stated in Section 3. Our experiments are then presented in Section 4, followed by a discussion of the outcomes and conclusions in Section 5.

2. Related work

2.1. Ensemble creation techniques

The most relevant techniques for ensemble creation are: a) Ensembles of independently trained models where diversity originates from the training process randomness, e.g., seed. Each ensemble member loss is a function notated as:

$$l(h^i, y) \tag{1}$$

where y is the ground-truth label, and h^i is the output of the ith single ensemble member. [7] presents an analysis of the resilience of independently trained ensembles. b) Bagging [2], reduces the variance of multiple models by averaging the outcomes of models created with different training data subsets. c) Boosting [39] sequentially trains models to reduce bias by sampling incorrectly classified inputs more often in the next model. d) Negative Correlation Learning (NCL) [26] trains models *in parallel* with a shared penalty term in their loss function to enforce prediction diversity. Generalized NCL (GNCL) [4] proposes two extensions for NCL: i) a generalized loss function for each member:

$$\sum_{i=1}^{M} l(h^i, y) - \frac{\lambda}{2M} \sum_{i=1}^{M} {d_i}^T \mathfrak{D} d_i \tag{2}$$

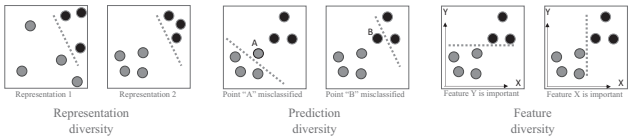


Figure 1: Model behavior diversity. a) Invariant decision boundary & diverse sample representation. b) Diverse decision boundary and measured by prediction errors. c) Diverse decision boundary and measured by feature relevance.

where M is the total number of members in the ensemble, d is the difference $h^i - f$ with f as the ensemble prediction, $\mathfrak{D}$ is the 2nd derivative of the loss function, and λ is a weighting hyper-parameter. ii) An implicit enforcement of diversity by balancing the ensemble and the individual loss:

$$\frac{1}{M} \sum_{i=1}^{M} \big(\lambda l(f, y) + \frac{1-\lambda}{M} \sum_{i=1}^{M} l(h^i, y)\big) \tag{3}$$

2.2. Diversity metrics in DL

There are many proposed metrics for diversity. [3] presents a survey and taxonomy for diversity metrics and [16] presents a survey of diversity for ML. In this work, we focus on behavior diversity metrics of a DL model. Input data diversity such as different modalities or implementation aspects such as the number of layers is not considered.

Prediction (output, failures) diversity Multiple prediction-based diversity metrics have been proposed. [23] presents a comprehensive evaluation of this class of metrics. Pair-wise measures based on the correct and incorrect statistics of two models include the Q-statistics, correlation coefficient ρ, and the disagreement measure:

$$D_{p,q} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}} \tag{4}$$

where the indexes a, b of the binary vectors N^{ab} indicate the correctness of the classifiers, e.g., N^{10} means first model is correct, second is incorrect. Non-pairwise measures that evaluate non-binary diversity include entropy, coincident failure diversity, cosine similarity, Kullback–Leibler divergence [12, 33], and the the Shannon equitability index [34, 7]:

$$E = -\sum_{i=1}^{S} p_i \ln p_i / \ln S \tag{5}$$

where S is the total number of prediction species/classes and p_i is the proportion of observed species i.

Representation (activation) diversity The intermediate representations (IR) can also be used to measure diversity. [15] compares the diversity in representation space from independently created ensembles and ensembles from variational approaches. Measuring IR diversity is challenging

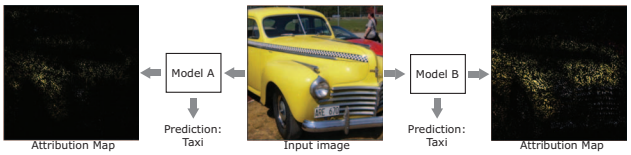


Figure 2: Attribution map diversity: Two models may predict the same outcome but based on different *evidence*.

due to space size and semantic ambiguity, i.e., the same semantic concept can be represented in many different ways. A naive use of any diversity metric such as cosine similarity could give semantically irrelevant diversity scores. In [22], the Centered Kernel Alignment (CKA) metric is proposed to obtain a statistical measure across a dataset on the similarity of any two layers of a DL model:

$$\text{CKA}(K, L) = \frac{\text{HSIC}(K, L)}{\sqrt{\text{HSIC}(K, K)\text{HSIC}(L, L)}} \tag{6}$$

where K and L are similarity matrices of the two feature maps being compared and HSIC is the Hilbert-Schmidt Independence Criterion which measures statistical independence. The feature maps may be layer activations or attention maps such as Saliency, Integrated Gradients, and Grad-CAM [42, 41, 43].

[36] proposed the use of the pull-away loss term from generative adversarial networks to induce diversity of such activations. Self-attention [47] (not related to attention maps) is one of the key techniques in the transformer architecture. In [35] the embeddings used to feed attention heads are masked in such a way as to enforce diversity of activations. In zero-shot learning, the *attribute* concept is used to enable training models that can, later on, predict unseen labels. These attributes can be considered for IR diversity, as well [50]. Closely related to NCL, the self-supervised approach of contrastive learning [40, 8] trains two models to produce latent features that are diverse for false positive cases and similar in true positives through a loss function such as the triplet loss that enforces the models to learn the similarity metric.

Input feature attribution diversity The importance of input features can be used to measure behavior diversity that to the best of our knowledge has not been explored. Figure 1 shows the relationship of an attribution-based metric w.r.t. prediction and representation diversity. Note, that attribution is not the same as attention or attributes in the context of zero-shot learning. Attention maps, such as those obtained from the activation of intermediate layers of CNNs, reflect the excitation of a network given an input. This activation however is not necessarily correlated with the final prediction, e.g., it could be an inhibitory factor. Attribution, on the other hand, indicates the importance of a feature to the final decision. See Figure 2 where *Saliency* is used to display the original pixels masked by the attribution scores from each model. A change to a pixel with high attribution (brighter) will have a stronger influence on the model prediction than a change to a pixel with low attribution.

2.3. Other diversity-based resilience approaches

Augmenting the training data by applying affine transformations such as rotations and scaling, geometric distortions such as blurring, and texture transfer help DL models to generalize better with a limited training data [29]. Adversarial training increases the robustness to intended attacks with adversarial samples to limit the model vulnerability to input perturbations [17, 9]. Such training data approaches are effective and complementary to the design diversity approaches of this study that address the model diversity.

Modality and point of view diversity [37] is an approach to address the failure modes of sensors such as cameras, radar, and lidar. The design diversity of DL models explored in this study is orthogonal to this approach, as model diversity can be applied to every single modality.

3. Methodology

We perform three main set of experiments: Before evaluating the impact of diversity in an ensemble, in our first experiment we asses the resiliency of individual models resulting from diverse architectures and training approaches. Second, we create three-member ensembles from independently created models of diverse architectures and evaluate the correlation between diversity metrics and resiliency. Finally, we create ensembles by enforcing diversity with NCL, and evaluate the resulting robustness.

3.1. Attribution-based diversity

Beyond the typical metrics of diversity in the literature, we propose to explore attribution-based diversity. In Equation 7 we propose a straight-forward measure of attribution-diversity based on the variance of pixel-level attribution.

$$A = \sum_{c=1}^{C} \sum_{p=1}^{P} Var(a_{c,p}) \tag{7}$$

where a is the input attribution score of a model at color channel c and pixel coordinate p. The computation of the input attribution scores a is performed with an attribution method, such as Saliency.

Our hypothesis is that attribution-based diversity can be positively correlated with ensemble resiliency with a better accuracy trade-off compared to prediction-based diversity.

This hypothesis is inspired by the theoretical result of the Littlewood and Miller (LM) model [24] that diverse design choices can produce less common failures. Diverse attribution maps of correct classifications imply that the models make predictions based on independent evidence.

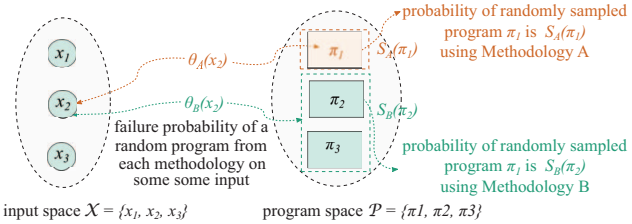


Figure 3: Relation between diverse design methodologies and the difficulty function in the LM Model [24]

3.1.1 Probability model for design diversity (LM model)

The Littlewood and Miller (LM) model [24] defines a probabilistic framework to analyze the impact of methodological diversity in the expected failure behavior. The model defines: 1) an input space $\mathcal{X} = \{x_1, x_2, ...\}$, representing all possible inputs x to a program and 2) a program space $\mathcal{P} = \{(\pi_1, \pi_2)\}$, for all possible programs π that could implement a program specification. A given design methodology will determine the probability to come up with a program π and is denoted as $S_A(\pi)$. Another design methodology $S_B(\pi)$ will assign a different probability to the same program. The model uses the concept of a difficulty function $\theta_M(x)$ that measures the probability that a randomly chosen program π from a given methodology distribution $S_M(\pi)$ will fail on a particular input $x \in \mathcal{X}$. The key insight consists in noticing that $\theta_A(x)$ can be different for a different methodology $\theta_B(x)$, i.e., for some methodology, a certain input may be difficult, but for another, it may be easy. See Figure 3 for a visual representation of these spaces. An analysis of this model concludes that if the design methodologies produce different difficulty functions θ, then the expected failure behavior on a random input will be negatively correlated due to the fact that the covariance of the θ's can be negative.

With this model, it is finally shown that a design methodology with diverse design choices that satisfy the following three properties will result in less common failures: 1) logically unrelated (one decision is independent of the other), 2) common failures of a decision are due to different factors, and 3) indifference to the selection of each methodology (no methodology is superior).

3.1.2 Loss function to enforce attribution diversity

We perform a first attempt to enforce attribution diversity with the following loss:

$$\frac{1}{M}\sum_{i=1}^{M} l(h^i, y) - \lambda A \tag{8}$$

This loss computes attribution scores variance in an ensemble and uses it as a penalty term weighted by λ.

Table 1: Resiliency of architectures to natural image corruption (ImageNet-$\overline{\text{C}}$ Lines(1.6 strength))

Training approach	Arch. class	DL model	# Params	ImageNet		ImageNet-$\overline{\text{C}}$ Lines1.6	
				top1	top5	top1	top5
Self-sup.	CNN	DINO ResNet50 [6]	25.6M	75.30	92.61	21.80	40.66
Superv.	CNN	ResNet50 [18]	25.6M	75.85	92.88	34.95	55.88
Superv.	CNN	ResNext50_32.4d [48]	25.0M	77.49	93.57	39.46	60.61
Superv.	CNN	ResNet152 [18]	60.2M	78.25	93.96	40.17	62.09
Superv.	NAS	EfficientNet_b7 [45]	66.3M	74.82	92.13	51.11	73.81
Superv.	CNN	ResNext101_64x4d [48]	83.4M	82.90	96.22	51.89	72.46
Superv.	ViT	Swin tiny [28]	28.3M	81.34	95.612	55.29	78.45
Self-sup.	ViT	DINO ViT_b_8 [6]	87.3M	80.06	95.02	55.75	78.37
Superv.	ViT	Swin base [28]	87.8M	83.30	96.46	58.71	81.05
Superv.	ViT	ViT base p16 [11]	86.6M	80.88	95.28	61.28	82.26
K. distill.	ViT	DeiT-base [46]	87.3M	83.34	96.50	64.07	84.66
Self-sup.	ViT	SwinV2 [27]	87.9M	83.34	96.44	59.976	81.70

3.2. Type of out-of-distribution data addressed

In this study, we evaluate resilience to covariate dataset distribution shifts, i.e., when the distribution of input features of the test dataset does not match the distribution of the training dataset. We use four natural image perturbations from the ImageNet-$\overline{\text{C}}$ dataset [30] that are sensible to occur in vision application domains, such as obstructions or liquid contaminants. Our scope is not to evaluate robustness against adversarial attacks, label shift variations, or resiliency to noise variations such as Gaussian, brown, etc.

4. Experimental results

4.1. Single model resiliency to data corruptions

Architecture, model complexity, and training approach can have an impact on the resiliency of a single DL model to natural image corruptions. To understand how diversity in these dimensions impact the model resiliency we evaluate models of different accuracy and complexity, different architecture, such as CNNs, transformers, and subnetworks from neural architecture search (NAS), as well as supervised and self-supervised training approaches on both the ImageNet validation dataset and on the corrupted version ImageNet-$\overline{\text{C}}$ "Lines" (strength of 1.6). See Table 1.

Observations to Table 1: Although the model size is highly correlated with the final accuracy and resilience in the corrupted dataset, the architecture seems to play a more determining factor. The smallest transformer with only 28M parameters is superior to other CNNs with 2 or 3x more parameters. Self-supervision slightly decreases both metrics as appreciated in the comparison of ResNet50 and ViT models using supervised learning. SwinV2 is an exception but this model introduced more architectural innovations too. Knowledge distillation from a CNN teacher shows a slight improvement over supervised ViTs.

Answer to RQ1: In this experiment, it is observed that model architecture is more important to resiliency than model accuracy or size.

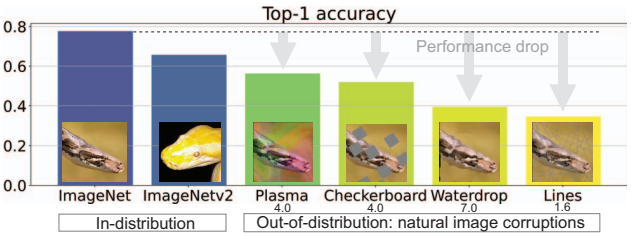


Figure 4: Top-1 accuracy of ResNet50 on in-distribution data sets (ImageNet and ImageNetv2[38]) and out-of-distribution datasets (ImageNet-$\overline{\text{C}}$). The resilience of ResNet50 drops significantly against natural corruptions.

To understand the effect of other corruptions, we evaluate a ResNet50 model[1] on six different data sets: ImageNet [10], ImageNetv2 [38], and four corruptions on a fixed perturbation strength from the ImageNet-$\overline{\text{C}}$ dataset [30]: Plasma (4.0), Checkerboard(4.0), Waterdrop(7.0) and Lines (1.6). See Figure 4. The first two are in-distribution, i.e., the covariates (input features) and labels of the validation set follow a similar distribution to the training data set. The last four are out-of-distribution, as the model has never seen such corruptions of the input images during training.

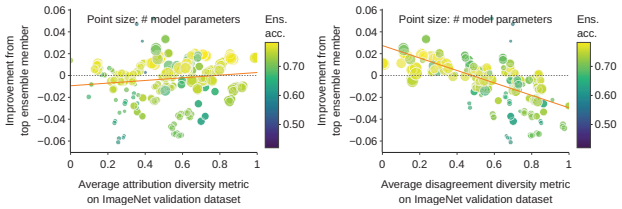
Observations to Figure 4: A "good" classifier with accuracy close to 80% can have a tremendous performance decrease in the presence of moderate natural corruptions where a human would probably not.

4.2. Diversity of ensembles from heterogeneous architectures

To understand the diversity/accuracy trade-off of the attribution-based metric in comparison to the established prediction-based diversity approach we perform two different experiments: First, we create multiple ensembles of independently trained models with a wide diversity in architecture. Second, we create multiple ensembles using models discovered in a weight-sharing super-network [5], i.e., models whose architecture has been found using neural architecture search (NAS) and not by manual design.

The architectures explored here are CNNs (ResNext [48] & SqueezeNet [19]), Vision transformers (DeiT [46]) and NAS (MNASNET [44] & BootstrapNAS [31]) using supervised or self-supervised training[2]. In total 14 models were trained with different hyperparameters to create three-member ensembles [3] of all possible combinations.

Figure 5 shows the ensemble performance of all 364 ensembles created from these 14 models using an averaging consensus mechanism, i.e., the logit output of all ensemble members is averaged first and then the highest score is used to make the prediction. The left-hand side shows on the X-axis the proposed attribution-based diversity metric (Eq. 7) using *Saliency* as attribution method. The right-hand side shows the disagreement prediction diversity metric (Eq. 4). Each point is an ensemble evaluated on the entire validation dataset of ImageNet. The color indicates the final average accuracy of the ensemble. The Y-axis indicates the average benefit of creating an ensemble: $Y = A_{ens} - A_{top}$, where A_{ens} is the ensemble accuracy and A_{top} is the accuracy of its most accurate member, i.e., how much accuracy improvement was obtained in comparison to a single model (the most accurate in the ensemble). In this way, it can be appreciated when an ensemble makes sense: it has to lay above the zero line (dashed). The ensemble cost is measured by the number of parameters which has a direct influence on the memory and the number of operations required. The ideal ensemble is one with the brightest color, smallest radius, and residing above zero.



(a) Attribution diversity (b) Prediction disagreement

Figure 5: Evaluation on ImageNet val. dataset of 364 three-member ensembles from heterogeneous architectures using **averaging as consensus mechanism**. Y-axis: Improvement of the ensemble against its own top ensemble member. X-axis: Normalized diversity metric. Color: absolute ensemble accuracy. Bubble size: Model parameter size. The attribution diversity metric is not negatively correlated with the ensemble improvement as disagreement diversity is.

Observations to Figure 5: In contrast to prediction-based diversity (b) which is negatively correlated with the ensemble improvement [23], Figure 5 (a) shows how the proposed attribution diversity is positively correlated.

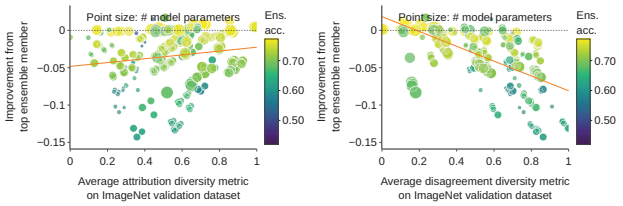
Figure 6 shows the same ensemble combinations, but this time using a majority voting consensus mechanism, i.e., the prediction with the highest number of votes wins.

Observations to Figure 6: The same correlation trends can be observed with majority voting. However, the most interesting aspect is that the vast majority of the ensembles here reside under the zero line. This means that majority voting with three ensembles tends on average to produce less accurate models. This corroborates the findings of [23].

[1] In the remainder of this paper, we select ResNet50 as the baseline architecture due to its common use as a reference.

[2] Details on the architecture and optimization hyper-parameters can be found in Section A of the supplementary material.

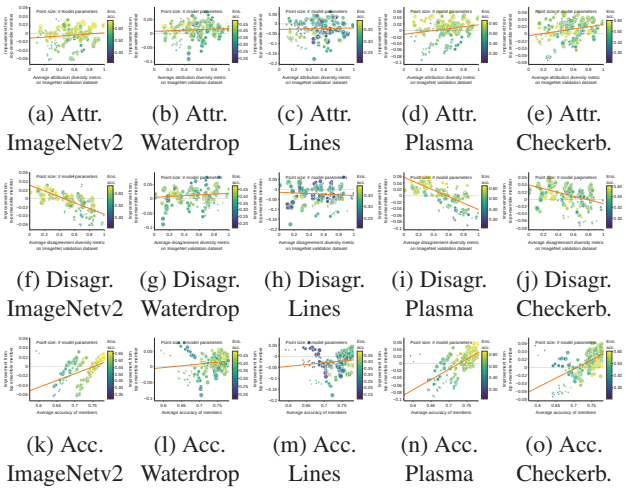
[3] We restricted to 3 member-ensembles to keep the number of possible ensemble combinations manageable.



(a) Attribution diversity (b) Prediction disagreement

Figure 6: Evaluation on ImageNet validation dataset of the same ensembles as in Figure 5 but using **voting as consensus mechanism**. In contrast to averaging, voting produces mostly ensembles that decrease the final performance instead of improving it.

We evaluate the same ensembles on five more validation datasets and verify that the observed trend in the validation dataset applies to natural corruptions. In addition, we compare the two diversity metrics to a simple validation accuracy metric. See Figure 7.



(a) Attr. ImageNetv2 (b) Attr. Waterdrop (c) Attr. Lines (d) Attr. Plasma (e) Attr. Checkerb.

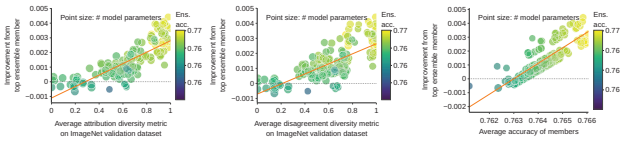
(f) Disagr. ImageNetv2 (g) Disagr. Waterdrop (h) Disagr. Lines (i) Disagr. Plasma (j) Disagr. Checkerb.

(k) Acc. ImageNetv2 (l) Acc. Waterdrop (m) Acc. Lines (n) Acc. Plasma (o) Acc. Checkerb.

Figure 7: Comparison of **trend lines**, i.e., correlation of improvement to three different metrics on five validation datasets (consensus: averaging). Columns: Datasets. Rows: Metrics. The attribution metric (a to e) has a much lower accuracy trade-off compared to the disagreement metric (f to j) but is higher than the average accuracy metric (k to o). Bigger plots to appreciate individual ensembles are presented in Section B of the supplemental material.

Observations to Figure 7: Attribution-based diversity is better correlated as well. These results serve as evidence to confirm that the diversity-accuracy trade-off is better for attribution than for prediction diversity. However, the metric of averaging the individual accuracies of the ensemble members is more strongly correlated with the ensemble improvement in corruptions.

Next, Figure 8 presents the results of the second experiment on architectures created with NAS. We used the open-source framework BootstrapNAS [31, 32] to create a weight-sharing super-network. The super-network is trained from an initial ResNet50 model. We then sample 11 subnetworks with different configurations but similar complexity by varying the width and depth of the CNN.



(a) Attribution (b) Disagreement (c) Accuracy

Figure 8: Comparison of 165 heterogeneous ensembles of **architectures automatically created with weight-sharing neural architecture search (NAS)**. 11 models with different architectures were selected to be very close in complexity, i.e., number of parameters. The correlation between the two diversity metrics is highly positive but the increment in performance is very small.

Observations to Figure 8: Although the correlations seem strong for all metrics, the actual ensemble improvement is very low, i.e., less than 0.04%.

Answer to RQ2, we consistently observed that attribution-based diversity is more positively correlated with accuracy than prediction-based disagreement diversity.

To assess if a more complex attribution method could provide different metric results, we evaluated the pair-wise diversity on the entire validation set on six subnetworks using *Saliency* and *Integrated Gradients* attribution methods with 1, 2, 10, and 50 backpropagation passes. See Figure 9.

Observations to Figure 9: The average correlation coefficient of the normalized diversity scores of all methods is 0.998. Using Saliency-based attribution is then justified as it provides the lowest performance penalty with a very similar performance to more complex methods.

4.3. Enforcing diversity in homogeneous ensembles

We perform a set of training experiments to enforce diversity into the ensembles through the loss function via the Negative Correlation Learning paradigm. We use ResNet50 for all ensemble members, and evaluate different heuristics: a) Independently trained members using cross-entropy as loss in Equation 1. Four different consensus approaches in GNCL using Equation 2: b) average, c) median, d) geometric mean, e) majority vote, f) GNCL and averaging consensus but masking the penalty term for incorrect classifications, i.e., $(h^i \neq y) \Rightarrow (\lambda = 0)$, and g) Balancing a loss function between the team and individual members

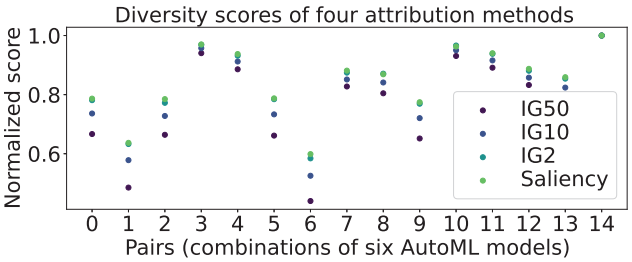


Figure 9: Comparison of Saliency and Integrated Gradients with 2, 10 & 50 samples. The X-axis represents different pairs of models created with BootstrapNAS. The Y-axis is the normalized diversity score of each method.

(Equation 3). The optimization method in all cases was AdaBelief [51] for 100 epochs with a learning rate of 1e-3 decaying 10% every 30 epochs, epsilon of 1e-8, betas: (0.9,0.999), batch size of 64 and a λ factor of 0.2. In ImageNet classification, we empirically observed that bigger λ values in Equation 2 fail to learn. Results for the six heuristics are presented in Figure 10.

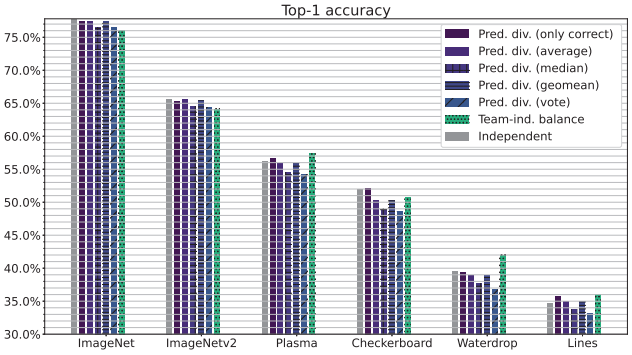


Figure 10: The resulting accuracy and **resiliency to natural corruptions of diversity enforcement** based on Negative Correlation Learning with different consensus mechanisms. Each ensemble has three members with a ResNet50 architecture. The heuristic of balancing the loss of the ensemble and the loss of the individual member produced the most resilient ensemble to corruptions.

Observations to Figure 10: Explicit enforcement of prediction diversity does not result in improved resilience. However the balanced loss (Eq. 3) provides a significant advantage in 3 out of 4 natural corruptions.

Answer to RQ3: balancing the loss of the individual members and the ensemble provided a significant advantage in 3 out of 4 natural corruptions when compared to the prediction diversity enforcement variants.

Table 2: **Diversity of predictions** from all members of three ensembles as measured by the Shanon equitability index H. H_{corr} and H_{inco} indicate the metric computed on all samples that were correctly or incorrectly classified. The six subcolumns correspond to the six validation datasets.

	H_{corr}						H_{inco}					
	IN	I2	WD	LI	PL	CB	IN	I2	WD	LI	PL	CB
Att. div.	0.13	0.16	0.29	0.32	0.23	0.23	0.46	0.49	0.62	0.60	0.57	0.58
Pred. div.	0.10	0.13	0.29	0.31	0.21	0.23	0.46	0.49	0.67	0.69	0.59	0.63
Hetero.	0.12	0.17	0.35	0.37	0.25	0.29	0.51	0.55	0.74	0.77	0.67	0.71

4.3.1 First attempt at enforcing attribution diversity

We perform a first attempt to enforce attribution diversity using the loss of Equation 8 and the same optimization parameters used in GNCL. The computational overhead to calculate the attributions is 2x using the Saliency method. Empirically, we tried five different lambda weights values: $\{10, 1, 0.1, 0.01, 0.001\}$ but found training instabilities. The smallest λ value resulted in convergence up to epoch 21 for 63.7% top1 accuracy. We believe that the penalty term of Eq. 8 is in conflict with the original loss and it would be more appropriate to investigate a better penalty term than to optimize this hyper-parameter in future work.

4.3.2 Diversity of NCL ensembles

Figures 11, 12 together with Table 2 show three types of diversity (attribution, prediction, and intermediate representation) for three models created through independently created heterogeneous architectures, prediction diversity enforcement and attribution diversity enforcement with the following top1 accuracies on the ImageNet validation dataset: 78.2%, 76.1%, and 63.7%.

Prediction diversity. In Table 2, the Shannon equitability index metric (Eq. 5) is shown for correctly and incorrectly classified samples for three ensembles: attribution diversity (Eq. 8), prediction diversity (Eq. 4) and heterogeneous architectures on all six datasets. The heterogeneous ensemble produces more diverse predictions in general. High values indicate that the individual predictions are more often in disagreement in the cases of OOD samples.

Attribution diversity. We present a few resulting attribution maps in Figure 11 for the NCL-based prediction-diversity enforcement, attribution-diversity enforcement (at epoch 21), and independently trained architectures.

Observations to Figure 11: Independently trained heterogeneous architectures and attribution-diversity enforcement produce more diverse attribution maps than homogeneous models trained to have diverse prediction outcomes.

Representation diversity. In Figure 12, we investigate the resulting diversity/similarity of the internal layers via CKA (Equation 6) of two ensemble members for three different diversity enforcing techniques.

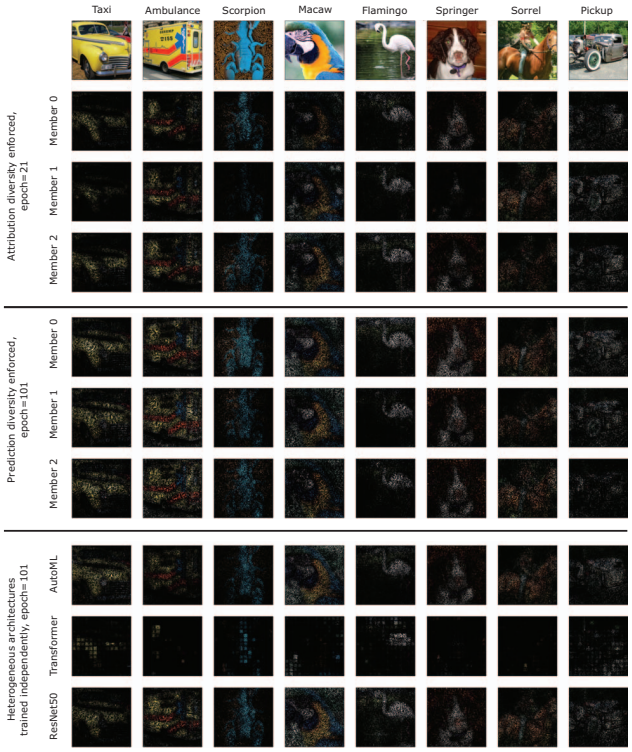
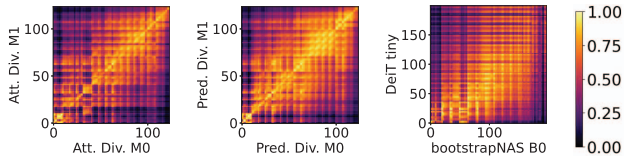


Figure 11: **Attribution map diversity** of three diversity-inducing techniques on 8 ImageNet val. set samples.



(a) Attribution diversity enf. (b) Prediction diversity enf. (c) Heterog. architectures

Figure 12: Centered Kernel Alignment maps to visualize the resulting **similarity of model layers** on different diversity enforcing techniques.

Observations to Figure 12: The CKA visualization reflects that the enforcement of attribution diversity produces less similarity in the layers than output diversity enforcement or by independent heterogeneous architectures.

Answer to RQ4: Prediction diversity was higher for heterogeneous architectures trained independently than NCL on prediction or attribution diversity. Attribution diversity is significantly lower when enforcing prediction diversity compared to heterogeneous architectures trained independently. Activation diversity is low at the last layers for both prediction and attribution diversity enforcement, while for heterogeneous architectures trained independently, the middle layers showed less diversity.

5. Discussion and conclusions

In our experiments, we could observe how ensembles are capable to improve the resiliency of the best single models but it is not always the case.

A model created out of a superior architecture such as a vision transformer can benefit in the context of an ensemble as long as the other members are diverse in the *right sense*, i.e., their predictions complement each other and demonstrate low common failure behavior.

Models extracted from NAS are observed to be too similar to provide any meaningful advantage in an ensemble.

If the models are created independently from multiple architectures, ensembling will be more successful by using individual accuracy as the member selection criteria. While the proposed attribution-based diversity improves the accuracy-diversity trade-off in comparison to prediction diversity, the correlation of individual accuracy is stronger to the final resiliency. However, mixing different architectures does not consistently produce good ensembles as observed in the many ensembles under the zero line in Figure 7.

If the models are trained to be negatively correlated in their output predictions, a balanced approach of individual and ensemble accuracy produces the most resilient ensembles against out-of-distribution samples.

5.1. Conclusions and next steps

In this study, we explored different approaches to measure and enforce diversity in ensembles and evaluated their impact on natural data corruption resiliency. The key takeaways are: 1) model architecture is more important for individual resiliency than model size or model accuracy, but architectural diversity in an ensemble is usually not more resilient, 2) attribution-based diversity is less negatively correlated to the ensemble accuracy than prediction-based diversity, 3) a balanced loss function of individual and ensemble accuracy creates more resilient ensembles for image natural corruptions, and 4) architecture diversity produces more diversity in all explored diversity metrics: predictions, attributions, and activations.

In addition, other valuable findings are: a) Saliency attribution can be sufficient to measure input attribution diversity, b) Ensembles created from models of similar complexity that were discovered by weight-sharing Neural Architecture Search for our experiments barely provide any accuracy improvement, and c) Enforcing attribution-based diversity during training through a variance-based penalty term is not stable and needs further research.

In future work, several experiments could be done to understand the complexity-resiliency trade-off, e.g., through knowledge distillation, improved heuristics to enforce attribution diversity, and compare diversity approaches in tasks beyond image classification.

Acknowledgements This work was partially funded by the Federal Ministry for Economic Affairs and Climate Action of Germany, as part of the research project SafeWahr (Grant Number: 19A21026C).

We would also like to thank Professors Lorenzo Strigini and Peter Popov for the fruitful conversations and feedback on this work.

References

[1] Taiga Abe, Estefany Kelly Buchanan, Geoff Pleiss, Richard S. Zemel, and John P. Cunningham. Deep ensembles work, but are they necessary? In *NeurIPS*, 2022.

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

[3] Gavin Brown, Jeremy L. Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Inf. Fusion*, 6(1):5–20, 2005.

[4] Sebastian Buschjäger, Lukas Pfahler, and Katharina Morik. Generalized negative correlation learning for deep ensembling. *CoRR*, abs/2011.02952, 2020.

[5] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

[7] Abraham Chan, Niranjhana Narayanan, Arpan Gujarati, Karthik Pattabiraman, and Sathish Gopalakrishnan. Understanding the resilience of neural network ensembles against faulty training data. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, pages 1100–1111, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[9] Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Jeremy M. Dawson, and Nasser M. Nasrabadi. Exploiting joint robustness to adversarial perturbations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1119–1128. Computer Vision Foundation / IEEE, 2020.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[12] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Diversity with cooperation: Ensemble methods for few-shot classification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3722–3730. IEEE, 2019.

[13] DE Eckhardt and LD Lee. A theoretical basis for the analysis of redundant software subiect to coincident errors. *NASA Tech. Memo*, 86369:151, 1985.

[14] EUROCAE. Eurocae ed-12c - software considerations in airborne systems and equipment certification. Standard ED-12C, 2012.

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

[16] Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *IEEE Access*, 7:64323–64350, 2019.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[19] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[20] ISO. Iso26262 road vehicles – functional safety. Standard 26262, 2011.

[21] ISO/PAS. Iso/pas 21448:2019 safety of the intended functionality. Standard 21448, 2019.

[22] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

[23] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, 2003.

[24] Bev Littlewood and Douglas R. Miller. Conceptual modeling of coincident failures in multiversion software. *IEEE Trans. Software Eng.*, 15(12):1596–1614, 1989.

[25] Bev Littlewood, Peter Popov, and Lorenzo Strigini. Modeling software design diversity: a review. *ACM Computing Surveys (CSUR)*, 33(2):177–208, 2001.

[26] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999.

[27] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11999–12009. IEEE, 2022.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[29] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122, 2018.

[30] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 3571–3583, 2021.

[31] J. Pablo Muñoz, Nikolay Lyalyushkin, Yash Akhauri, Anastasia Senina, Alexander Kozlov, and Nilesh Jain. Enabling nas with automated super-network generation. *Association for the Advancement of Artificial Intelligence*, 2022.

[32] J. Pablo Muñoz, Nikolay Lyalyushkin, Chaunte Lacewell, Anastasia Senina, Daniel Cummings, Anthony Sarah, Alexander Kozlov, and Nilesh Jain. Automated super-network generation for scalable neural architecture search. In *International Conference on Automated Machine Learning*, pages 5–1. PMLR, 2022.

[33] Giung Nam, Jongmin Yoon, Yoonho Lee, and Juho Lee. Diversity matters when learning from ensembles. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8367–8377, 2021.

[34] Robert K Peet. Relative diversity indices. *Ecology*, 56(2):496–498, 1975.

[35] Viraj Prabhu, Sriram Yenamandra, Aaditya Singh, and Judy Hoffman. Adapting self-supervised vision transformers by probing attention-conditioned masking consistency. *CoRR*, abs/2206.08222, 2022.

[36] Zhongang Qi, Saeed Khorram, and Fuxin Li. Embedding deep networks into visual explanations. *Artif. Intell.*, 292:103435, 2021.

[37] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.

[39] Robert E Schapire. A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406. Citeseer, 1999.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

[41] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017.

[42] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[44] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2820–2828. Computer Vision Foundation / IEEE, 2019.

[45] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[48] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995. IEEE Computer Society, 2017.

[49] Y.C. Yeh. Design considerations in boeing 777 fly-by-wire computers. In *Proceedings Third IEEE International High-Assurance Systems Engineering Symposium (Cat. No.98EX231)*, pages 64–72, 1998.

[50] Xiaojie Zhao, Yuming Shen, Shidong Wang, and Haofeng Zhang. Boosting generative zero-shot learning by synthesizing diverse features with attribute augmentation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3454–3462. AAAI Press, 2022.

[51] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha C. Dvornek, Xenophon Papademetris, and James S. Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.