Guardian of the Ensembles: Introducing Pairwise Adversarially Robust Loss for Resisting Adversarial Attacks in DNN Ensembles

Shubhi Shukla[1], Subhadeep Dalui[2], Manaar Alam[3], Shubhajit Datta[4],
Arijit Mondal[5], Debdeep Mukhopadhyay[2], Partha Pratim Chakrabarti[2]
[1]Centre for Computational and Data Sciences, IIT Kharagpur, India
[2]Computer Science and Engineering Department, IIT Kharagpur, India
[3]Center for Cyber Security, New York University Abu Dhabi, UAE
[4]Department of Artificial Intelligence, IIT Kharagpur, India
[5]Computer Science and Engineering Department, IIT Patna, India
{shubhishukla, csesubhadeep2022}@kgpian.iitkgp.ac.in, alam.manaar@nyu.edu,
shubhajitdatta1988@gmail.com, arijit@iitp.ac.in, {debdeep, ppchak}@cse.iitkgp.ac.in

Abstract

Adversarial attacks rely on transferability, where an adversarial example (AE) crafted on a surrogate classifier tends to mislead a target classifier. Recent ensemble methods demonstrate that AEs are less likely to mislead multiple classifiers in an ensemble. This paper proposes a new ensemble training using a Pairwise Adversarially Robust Loss (PARL) that by construction produces an ensemble of classifiers with diverse decision boundaries. PARL utilizes outputs and gradients of each layer with respect to network parameters in every classifier within the ensemble simultaneously. PARL is demonstrated to achieve higher robustness against black-box transfer attacks than previous ensemble methods as well as adversarial training without adversely affecting clean example accuracy. Extensive experiments using standard Resnet20, WideResnet28-10 classifiers demonstrate the robustness of PARL against state-of-the-art adversarial attacks. While maintaining similar clean accuracy and lesser training time, the proposed architecture has a 24.8% *increase in robust accuracy (*$\epsilon = 0.07$*) from the state-of-the art method. Code is available at:* https://github.com/shubhishukla10/PARL

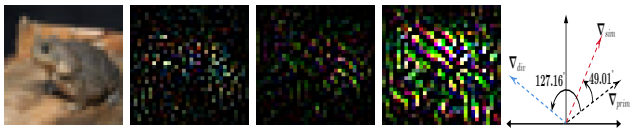
1. Introduction

While Deep Learning (DL) models are extremely efficient in solving complicated decision-making tasks, they are vulnerable to well-crafted Adversarial Examples (AE) [18]. The widely-studied phenomenon of AE has produced numerous attacks with varied complexities and effective deceiving strategies [6]. An extensive spectrum of defenses against such attacks has also been proposed in the literature [6], which generally falls into two categories. The first category enhances the training strategy of DL models to make them less vulnerable to AE [9, 16]. However, it has been demonstrated that these defenses are not generalized for all varieties of AE but are constrained to specific categories [2, 5]. The second category intends to detect AE by simply flagging them [12, 14]. However, it has been illustrated with several experiments that these detection techniques could be efficiently bypassed by a strong adversary having partial or complete knowledge of the internal working procedure [4].

While the approaches mentioned above deal with standalone models, in this paper, *we utilize an ensemble of models to resist adversarial examples (AE)*. The notion of using *diverse ensembles* to increase robustness against AE has recently gained popularity [25]. The primary motivation for using an ensemble-based defense with diverse decision boundaries is that if multiple models with similar *decision boundaries* perform the same task, the *transferability property* of deep learning models makes it easier for an adversary to misclassify all the models simultaneously using AE crafted on any of the models. However, it will be difficult for an adversary to misclassify multiple models simultaneously if they have diverse decision boundaries.

Related Works: Ensemble-based Adversarial Defense:

[17] introduced ensemble-based defense against AE using various ad-hoc techniques such as different random initializations, different neural network structures, bagging the input data, and adding Gaussian noise while training. [19] proposed *Ensemble Adversarial Training* that incorporates perturbed inputs transferred from other pre-trained models during *adversarial training* to decouple AE gen-



(a) (b) (c) (d) (e)

Figure 1. (a) Input image; (b) ∇_{prim}: Gradient of loss in the primary model; (c) ∇_{sim}: Gradient of loss in another model with *similar* decision boundaries; (d) ∇_{div}: Gradient of loss in a model with *not so similar* decision boundaries but comparable accuracy; (e) Symbolic directions of all the gradients in higher dimensions. Gradients are computed with respect to the image shown in (a).

eration from the parameters of primary model. However, these methods do not explicitly focus on incorporating diversity in the decision boundaries of the models within an ensemble. [10] proposed *Diversity Training* of an ensemble of models with uncorrelated loss functions using *Gradient Alignment Loss* (GAL) to reduce the dimension of adversarial sub-space shared between different models and increase the robustness of the classification task. [15] proposed *Adaptive Diversity Promoting* (ADP) regularizer to train an ensemble of models that encourages the non-maximal predictions in each member in the ensemble to be mutually orthogonal, degenerating the transferability that aids in resisting AE. [22] proposed a methodology, called DVERGE, that isolates the adversarial vulnerability in each model of an ensemble by distilling non-robust input features. [23] proposed *Transferability Reduced Smooth* (TRS) ensemble that enforces diversity among the models within an ensemble by simultaneously reducing loss gradient and smoothing decision regions using support instances as regularizers. Recently, [3] proposed *Ensemble-in-One* (EIO), method which works by using a random gated network to exponentially increase the number of paths for ensemble learning within a single model, resulting in better adversarial robustness.

The methods mentioned above either do not inherently enforce diversity on decision boundaries of the models or are less robust against stronger adversaries, significantly impacting clean example accuracy. *In this work, we propose a systematic approach that incorporates diversity among decision boundaries of submodels in an ensemble to enhance robustness against adversarial examples (AE). This diversity is achieved by considering mutual dissimilarity in gradients of each layer with respect to intermediate network parameters and the output of intermediate convolution layers during training. Such diversity reduces the transferability of AE within the ensemble.*

Intuition behind the Proposed Approach:

The first part our proposed ensemble loss method aims to diversify classifiers by making their gradients dissimilar/orthogonal. We illustrate this with an example of image classifiers trained on CIFAR-10, shown in Fig. 1.

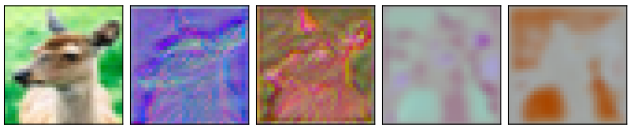
Fig. 1a shows an input image of a 'frog' which we use to demonstrate how gradient of loss with respect to intermediate convolution layer parameters is visualized in classifiers $\mathcal{M}_{prim}$ and $\mathcal{M}_{sim}$, which have similar decision boundaries, and in $\mathcal{M}_{div}$, which has a distinctly different decision boundary from $\mathcal{M}_{prim}$. Although $\mathcal{M}_{prim}$ and $\mathcal{M}_{sim}$ are trained under the same settings but with different initializations, their gradients, ∇_{prim} (Fig. 1b) and ∇_{sim} (Fig. 1c), tend to point in nearly the same directions as shown in Fig. 1e. In contrast, the gradient of $\mathcal{M}_{div}$, ∇_{div} (Fig. 1d), significantly diverges. This indicates that adversarial examples crafted for $\mathcal{M}_{prim}$ can easily fool $\mathcal{M}_{sim}$ but are less likely to affect $\mathcal{M}_{div}$, highlighting the impact of decision boundary diversity on adversarial example transferability. Focusing on gradients relative to intermediate layer weights rather than just the input delves deeper into the neural network's learning process. This helps in targeting the core of the classifiers' feature extraction and representation mechanisms. Early and intermediate convolutional layers are where raw input data begins its transformation into a hierarchy of features, which are then used for classification. By promoting orthogonal gradients in these layers, we ensure that each classifier within the ensemble develops a unique approach to processing and interpreting the input data, leading to diverse feature representations.

Building on the gradient diversity objective, our ensemble loss method aims to further diversify internal representations by minimizing the correlation between outputs of intermediate convolutional layers across classifiers. Unlike cosine similarity, which encourages gradient orthogonality, correlation is used here to assess similarity between intermediate layer outputs. This is because convolutional layers capture spatial hierarchies of features where both activation patterns (direction) and intensities (magnitude) are crucial. Correlation accounts for both aspects, providing a comprehensive measure of similarity and ensuring each classifier uniquely contributes to the ensemble, thus increasing robustness against adversarial attacks. Fig.2 illustrates this with an input image labeled as 'Deer' from CIFAR-10, processed by four CNN classifiers. The first two classifiers, trained identically but with different initializations, produce similar outputs (Fig.2b and Fig.2c). The latter two classifiers, specifically trained for distinct representations, show significantly different outputs (Fig.2d and Fig. 2e), demonstrating enforced diversity.

Our method is among the first to encourage diversity at both the decision boundary and intermediate representation levels. This novel approach ensures a more robust model compared to previous adversarial ensemble defenses by diversifying the potential paths an adversarial input might take, making it harder for such attacks to succeed.

Contributions: We summarize our contributions below:

- We propose a method that, *by construction*, increases diversity in the decision boundaries among all the models within an ensemble to degrade the transferabil-



(a) (b) (c) (d) (e)

Figure 2. (a) Input image; (b) and (c) Similar outputs of an intermediate convolution layer of two classifiers with similar internal representations; (d) and (e) Contrasting outputs of an intermediate convolution layer of two classifiers which are trained simultaneously to have distinct internal representations

ity of AE.

- We propose a *Pairwise Adversarially Robust Loss* (PARL) function by utilizing outputs and gradients of each layer of every model within the ensemble simultaneously while training to produce such varying decision boundaries.
- PARL significantly improves the overall robustness of an ensemble against black-box transfer attacks without substantially impacting the clean example accuracy.

We evaluated PARL extensively using CIFAR-10, CIFAR-100, and Tiny Imagenet datasets with *Resnet20* and *WideResnet28-10* architectures against state-of-the-art adversarial attacks such as *PGD* [13], *M-DI*^{2}*-FGSM* [21], *SGM* [20], and *Square* [1]. We compared PARL with previous ensemble adversarial defenses and the recent adversarial training method *TRADES*. At the highest perturbation strength of 0.07, PARL achieved a robust accuracy surpassing the state-of-the-art ensemble defense by 24.8%, with nearly one-third of the training time and similar clean accuracy for the CIFAR-10 Resnet20 model. Additionally, compared to TRADES, PARL showed similar robust accuracy and 3.68% higher clean accuracy, offering a better balance of security and utility.

2. Building Ensemble Networks using PARL

Threat Model: We define our threat model for generating AE in the context of a *Zero Knowledge Adversary*. This adversary lacks access to the target ensemble $\mathcal{M}_T$, but possesses knowledge of a surrogate ensemble $\mathcal{M}_S$ that has been trained using the same dataset. Also known as a *black-box* adversary, the adversary formulates AE on the source model $\mathcal{M}_S$ and subsequently transfers them to the target model $\mathcal{M}_T$.

Overview of PARL: Consider an ensemble $\mathcal{M}_\mathcal{T}$ consisting of $\mathcal{N}$ neural networks and denoted as $\mathcal{M}_\mathcal{T} = \bigcup_{i=1}^{\mathcal{N}} \mathcal{M}_i$, where $\mathcal{M}_i$ is the i^{th} network in the ensemble. All $\mathcal{M}_i$'s are trained simultaneously using PARL, which we subsequently discuss in detail. The final decision for an input on $\mathcal{M}_\mathcal{T}$ is decided over *majority voting* among all $\mathcal{M}_i$'s. Formally, assume a test set of t inputs $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$ with respective ground truth labels as $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\}$. The final decision of $\mathcal{M}_\mathcal{T}$ for an input $\mathbf{x}_j$ is defined as

$$\mathcal{C}(\mathcal{M}_\mathcal{T}, \mathbf{x}_j) = majority\{\mathcal{M}_1(\mathbf{x}_j), \mathcal{M}_2(\mathbf{x}_j), \cdots, \mathcal{M}_\mathcal{N}(\mathbf{x}_j)\}$$

$\mathcal{C}(\mathcal{M}_\mathcal{T}, \mathbf{x}_j) = \mathbf{y}_j$ for most $\mathbf{x}_j$'s in an appropriately trained $\mathcal{M}_\mathcal{T}$. The primary argument behind PARL is that all $\mathcal{M}_i$'s have dissimilar decision boundaries but not significantly different accuracies. Hence, a clean example classified as class $\mathcal{C}_\mathbf{x}$ in $\mathcal{M}_i$ will also be classified as $\mathcal{C}_\mathbf{x}$ in most other $\mathcal{M}_j$'s (where $j = 1 \ldots \mathcal{N}$, $j \neq i$) with a high probability. Consequently, due to the diversity in decision boundaries between $\mathcal{M}_i$ and $\mathcal{M}_j$ (for $i, j = 1 \ldots \mathcal{N}$ and $i \neq j$), the AE generated for a surrogate ensemble $\mathcal{M}_\mathcal{S}$ will have a different impact on each classifiers within $\mathcal{M}_\mathcal{T}$, i.e., the *transferability* of AE will be challenging within the ensemble. The adversary can also generate AE for $\mathcal{M}_\mathcal{T}$. However, the input image perturbation will be in different directions due to the diversity in decision boundaries among all $\mathcal{M}_i$'s. The collective disparity in perturbation directions makes it challenging to craft AE for the ensemble.

Basic Terminologies used in PARL: We assume that each $\mathcal{M}_i$ in $\mathcal{M}_\mathcal{T}$ has the same architecture with $\mathcal{H}$ hidden layers. Let $\mathcal{J}_{\mathcal{M}_i}(\mathbf{x}, \mathbf{y})$ be the loss function for the network $\mathcal{M}_i$ considering a data point $\mathbf{x}$, where $\mathbf{y}$ is the ground-truth label for $\mathbf{x}$. Let $\mathcal{F}_i^k(\mathbf{x})$ be the output of k^{th} hidden layer of $\mathcal{M}_i$ for input $\mathbf{x}$ and let w_i^k denote all network parameters up to k^{th} hidden layer which are involved in computation of $\mathcal{F}_i^k(\mathbf{x})$. Let us consider $\mathcal{F}_i^k(\mathbf{x})$ has $\mathcal{D}_k$ number of output features. Let $\nabla_{w_i^k}\mathcal{F}_i^k(\mathbf{x})$ denote the sum of gradients over each output feature of k^{th} hidden layer with respect to the parameters represented by w_i^k on the network $\mathcal{M}_i$ for data point $\mathbf{x}$. Hence, $\nabla_{w^k i}\mathcal{F}^k i(\mathbf{x}) = \sum_{f=1}^{\mathcal{D}k} \nabla w^k i[\mathcal{F}^k i(\mathbf{x})]_f$ where $\nabla_{w_i^k}[\mathcal{F}_i^k(\cdot)]_f$ is gradient of f^{th} output feature of k^{th} hidden layer on network $\mathcal{M}_i$ with respect to the parameters represented by w_i^k for data point $\mathbf{x}$. Let $\mathcal{X}$ be the training dataset with $|\mathcal{X}|$ examples.

PARL Construction: The main idea behind PARL is to train an ensemble of neural networks with diverse decision boundaries. To achieve such diversity the parameters in each network which are learned during training must be dissimilar across the ensemble. In this paper, we introduce PARL to train an ensemble so that the gradients of loss with respect to the network parameters lead to different directions in different networks for the same input. The gradients guide training of any neural network by giving an idea of the direction in which the parameters should be updated. Hence, the fundamental strategy is to make these gradients as dissimilar as possible while training all the networks simultaneously. The loss is computed using the output of the last layer and ground truth label. However, the last layer output depends on all intermediate layers' outputs. Therefore, loss and so its gradient both depend on the intermediate layers' outputs. As a result, employing diversity in intermediate layers will also enforce diversity in the model

decision boundary. Hence, instead of loss, we considered the output of intermediate layers for implementing diversity with a higher degree of control and better flexibility in employing constraints. Recognizing that gradient computation hinges on all intermediate parameters within a network, we bring into play a strategy to not only make the gradients dissimilar but also to influence the intermediate layers of all networks within the ensemble. We aim to minimize the correlation between the outputs of the hidden layers, instigating enhanced diversity at each layer. This twofold approach of reducing correlation and enhancing diversity not only ensures different gradient paths but also fortifies the robustness of the ensemble to adversarial perturbations. Consequently, the PARL framework presents a more potent defense against adversarial examples.

The pairwise similarity of gradients of the output of the k^{th} hidden layer with respect to the parameters between $\mathcal{M}_i$ and $\mathcal{M}_j$ for a particular data point $\mathbf{x}$ can be represented as

$$\mathcal{G}_k^{(i,j)}(\mathbf{x}) = \cos\theta_{i,j}(\mathbf{x}) = \frac{\langle \nabla_{w_i^k} \mathcal{F}_i^k(\mathbf{x}), \nabla_{w_j^k} \mathcal{F}_j^k(\mathbf{x}) \rangle}{|\nabla_{w_i^k} \mathcal{F}_i^k(\mathbf{x})| \cdot |\nabla_{w_j^k} \mathcal{F}_j^k(\mathbf{x})|}$$

where $\langle a, b \rangle$ represents the dot product between two vectors a and b, and $\cos\theta_{i,j}(\mathbf{x})$ represents the cosine of the angle between two vectors. The overall pairwise similarity between $\mathcal{M}_i$ and $\mathcal{M}_j$ for $\mathbf{x}$ considering $\mathcal{H}$ hidden layers is given as $\mathcal{G}^{(i,j)}(\mathbf{x}) = \sum_{k=1}^{\mathcal{H}} \mathcal{G}_k^{(i,j)}(\mathbf{x})$

Additionally, we utilize intermediate outputs of the convolution layers to further diversify the decision boundaries. For this we define the pairwise similarity of outputs of the k^{th} hidden layer with respect to input between $\mathcal{M}_i$ and $\mathcal{M}_j$ for a particular data point $\mathbf{x}$ as

$$\mathcal{L}_k^{(i,j)}(\mathbf{x}) = \rho(\mathcal{F}_i^k(\mathbf{x}), \mathcal{F}_j^k(\mathbf{x})) = \frac{\text{cov}(\mathcal{F}_i^k(\mathbf{x}), \mathcal{F}_j^k(\mathbf{x}))}{\sigma_{\mathcal{F}_i^k(\mathbf{x})} \sigma_{\mathcal{F}_j^k(\mathbf{x})}}$$

where $\mathcal{L}_k^{(i,j)}(\mathbf{x})$ is the Pearson correlation between $\mathcal{F}_i^k(\mathbf{x})$ and $\mathcal{F}_j^k(\mathbf{x})$, ρ is the Pearson correlation function, $\text{cov}(\mathcal{F}_i^k(\mathbf{x}), \mathcal{F}_j^k(\mathbf{x}))$ denotes the covariance between outputs of $\mathcal{F}_i^k(x)$ and $\mathcal{F}_j^k(x)$, $\sigma_{\mathcal{F}_i^k(\mathbf{x})}$ and $\sigma_{\mathcal{F}_j^k(\mathbf{x})}$ denote the standard deviations of the sub-model outputs respectively. We define: $\mathcal{L}^{(i,j)}(\mathbf{x}) = \sum_{k=1}^{\mathcal{H}} \mathcal{L}_k^{(i,j)}(\mathbf{x})$
Next, we define a penalty term $\mathcal{R}(\mathcal{M}_i, \mathcal{M}_j)$ for all training examples in $\mathcal{X}$ to pairwise train $\mathcal{M}_i$ and $\mathcal{M}_j$ as

$$\mathcal{R}(\mathcal{M}_i, \mathcal{M}_j) = \frac{1}{|\mathcal{X}| \cdot |\mathcal{H}|} \sum_{\mathbf{x} \in \mathcal{X}} \left(\mathcal{G}^{(i,j)}(\mathbf{x}) \cdot \mathcal{L}^{(i,j)}(\mathbf{x}) \right)$$

Here, the rationale behind multiplying the terms $\mathcal{G}$ and $\mathcal{L}$ lies in creating an interdependent relationship between the diversity of learning trajectories (as encouraged by gradient orthogonality) and the diversity of internal feature representations across the classifiers. By this design, a classifier's contribution to the ensemble's robustness is maximized only when it exhibits both gradient diversity and diverse feature representation simultaneously. This further ensures that the classifiers do not lean towards optimizing one aspect of diversity at the expense of the other.

We add $\mathcal{R}$ to training loss as a penalty parameter to penalize training for a large $\mathcal{R}$. $\mathcal{R}$ computes average pairwise similarity for all training examples. $\mathcal{R}$ will gradually decrease as relative angles between the pair of gradients increases in higher dimension. Hence, the objective of PARL is to reduce $\mathcal{R}$. Thus, we add $\mathcal{R}$ to training loss as a penalty parameter to penalize training for a large $\mathcal{R}$.

In ensemble $\mathcal{M}_\mathcal{T}$, we compute $\mathcal{R}$ for each distinct pair of $\mathcal{M}_i$ and $\mathcal{M}_j$ in order to enforce diversity between each pair of classifiers. We define PARL to train $\mathcal{M}_\mathcal{T}$ as

$$\begin{aligned} PARL(\mathcal{M}_\mathcal{T}) = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{i=1}^{\mathcal{N}} J_{\mathcal{M}_i}(\mathbf{x}, \mathbf{y}) \\ + \gamma \cdot \sum_{1 \le i < j \le \mathcal{N}} \mathcal{R}(\mathcal{M}_i, \mathcal{M}_j) \end{aligned} \tag{1}$$

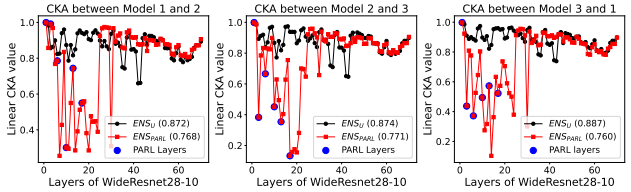
where γ is a hyperparameter controlling the accuracy-robustness trade-off. A lower γ enhances clean accuracy during ensemble training but reduces AE robustness. Conversely, a higher γ boosts AE robustness while potentially sacrificing overall accuracy.

One may note that including a penalty for each distinct pair of classifiers within $\mathcal{M}_\mathcal{T}$ to compute PARL has one fundamental advantage. If we omit the pair $(\mathcal{M}_a, \mathcal{M}_b)$ in PARL computation, training will continue without any diversity restrictions between them. Consequently, producing similar decision boundaries that increase the likelihood of adversarial transferability between them, affecting the overall robustness of $\mathcal{M}_\mathcal{T}$. One may also note that in an efficient implementation of PARL one needs a single forward pass to get all the hidden layer outputs. Additionally, the number of gradient computations (backward pass) is directly proportional to the number of classifiers in $\mathcal{M}_\mathcal{T}$. The gradients for each classifier are computed once and are reused to compute $\mathcal{R}$ for each pair of classifiers. Reusing gradients protects the implementation from exponential computational overhead. Moreover, the complexity of gradient computation of PARL mostly depends on the architectural depth of neural networks. Given a fixed architecture, the training complexity of PARL grows linearly with the number of training examples. Hence, PARL is applicable to any dataset without adversely affecting the original training time.

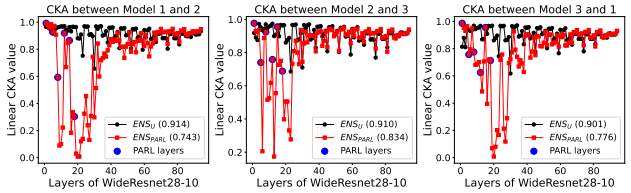
3. Experimental Evaluation

Evaluation Configurations: We consider two standard architectures Resnet20 and WideResnet28-10 for creating our ensembles. Each ensemble is a set of three sub-models of the same architecure[1]. We consider CIFAR-10 and CIFAR-100, standard image classification datasets for our evaluation. We consider four previously proposed

[1]We select 3 sub-models to compare PARL with related methods, most of which use 3 sub-models. PARL is scalable for larger ensemble sizes.



(a)



(b)

Figure 3. Layer-wise linear CKA values between each pair of models trained with CIFAR-10 on (a) Resnet20 and (b) WideResnet28-10 showing the similarities at each layer.

countermeasures to compare the performance of PARL. We denote ENS_{ADP}, ENS_{GAL}, ENS_{TRS}, ENS_{DVERGE}, and ENS_{EIO} to be the ensembles trained with the methods proposed in [15], [10], [22], [23], and [3], respectively. The ensemble trained with PARL is denoted as ENS_{PARL}. ENS_{U} is the baseline ensemble model.

We use *Adam optimization* to train all the ensembles with adaptive learning rate starting from 0.001. We dynamically generate an augmented dataset using random shifts, flips and crops to train both CIFAR-10 and CIFAR-100. We use $\gamma = 0.25$ as it provided best clean and robust accuracy trade-off (cf. Sec. 3.2), and *categorical crossentropy* loss for $\mathcal{J}_{\mathcal{M}_i}(\cdot)$ for ENS_{PARL} (ref. Equation (1)). All ensembles are trained using two GPU servers: Intel Xeon CPU@2.30GHz with 16GB NVIDIA Tesla P100 GPU and Intel Xeon CPU@2.40GHz with 48GB NVIDIA A40.

We evaluate PARL, considering the same attacks as most recent defense EIO [3]. For black-box transfer attack, we use the following attacks: (1) PGD with momentum and three random starts [13]; (2) M-DI^2-FGSM [21]; and (3) SGM [20]. The iterative steps are set to 100 with step size of $\epsilon/5$. We use $\epsilon = \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07\}$ for generating AE of different strengths[2]. We report the robust accuracy in all-or-nothing manner, meaning a sample is said to be correctly classified if all of its adversarial samples using different attack methods are correctly classified.

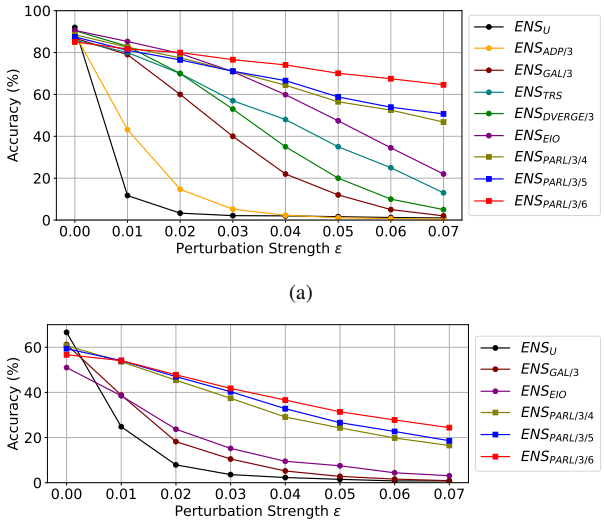
Analysing the Diversity: PARL aims to increase the diversity among all classifiers within an ensemble. To analyze the diversity of different classifiers trained using PARL, we use Linear Central Kernel Alignment (CKA) analysis [11]. The CKA metric $\in [0, 1]$ measures similarity between decision boundaries represented by a pair of neural networks. A higher CKA indicates a significant similarity in decision boundary representations, which implies good transferability of AE. We present an analysis on layer-wise CKA for each pair of classifiers within $ENS_{\mathcal{U}}$ and ENS_{PARL} trained with CIFAR-10 on Resnet20 and WideResnet28-10 architectures in Fig. 3 to show the effect of PARL on diversity. We enforce diversity among all classifiers in ENS_{PARL} for the first six convolution layers. We chose six layers as it provides better robust accuracy compared to other number of layers as shown in next subsection. We highlight the selected convolution layers for PARL in blue. There are intermediate layers between convolution layers as well such as batch-normalization and activation function.

We observe that each pair of models in $ENS_{\mathcal{U}}$ show a significant similarity at each layer. However, since ENS_{PARL} restricts the first six convolution layers, we observe a notable decline in CKA values at initial layers. The observation is expected as PARL imposes layer-wise diversity in its formulation. The overall average Linear CKA values between each pair of models in Fig. 3 are mentioned inside brackets within corresponding figure legends, which signifies that the classifiers within an ensemble trained using PARL shows a higher overall dissimilarity than the unprotected baseline ensemble. Next, we analyze the effect of observed diversity on the performance of ENS_{PARL}.

3.1. Robustness Evaluation

The attacker cannot access the model parameters and rely on surrogate models to generate transferable AE. Under such a black-box scenario, we use one hold-out ensemble with three Resnet20 architectures as the surrogate model. We randomly select 1000 test samples and evaluate the performance of black-box transfer attacks for all ensembles across a wide range of attack strength ϵ. We give a detailed performance evaluation considering multiple attack strengths for CIFAR-10 and CIFAR-100 dataset in Fig. 4a and Fig. 4b respectively. To avoid confusion with nomenclature of other ensemble defenses, $ENS_{PARL/3/4}$, $ENS_{PARL/3/5}$ and $ENS_{PARL/3/6}$ indicates an ensemble of 3 classifiers with 4, 5 and 6 initial convolution layers modified with the PARL loss respectively. On the CIFAR-10 dataset, we note some key observations. The model $ENS_{PARL/3/6}$ with a clean accuracy of 85.09% performs the best among all the previous ensemble defense methods. It is to be noted, for $\epsilon = 0.07$, $ENS_{PARL/3/6}$ has a robust accuracy of 64.6%, which is 42.6% higher than the previous state-of-the-art defense ENS_{EIO}, but has drop of 5% clean accuracy. On the other hand $ENS_{PARL/3/5}$ and $ENS_{PARL/3/4}$ show an increase of 28.7% and 24.8% robust accuracy for $\epsilon = 0.07$ with just 3.3% and 2% drop in clean accuracy compare to ENS_{EIO}. For the CIFAR-100 dataset, PARL surpasses state-of-the-art method both in terms of robust as well as clean accuracy. These results suggest that, based on the desired robustness level, a model architect can adjust the number of layers influenced by the PARL loss function. This fine-tuning enables

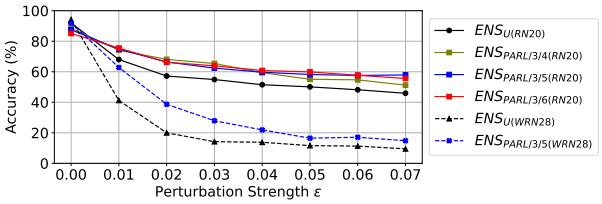
[2]We use majority voting as final ensemble decision for PARL. Hence, we use majority attack [8] for all adversarial attacks on PARL ensembles.



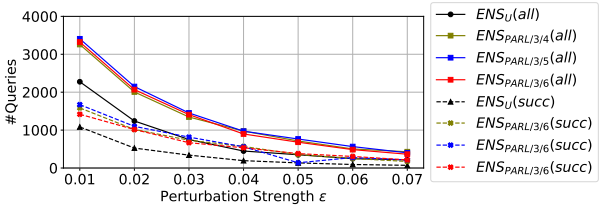
(a)

(b)

Figure 4. Resnet20 Ensemble classification accuracy (%) vs. Attack Strength (ϵ) against black-box transfer attacks generated from surrogate ensemble with (a) CIFAR-10 and (b) CIFAR-100 dataset



(a)



(b)

Figure 5. (a) Resnet20 (RN20) and WideResnet28-10 (WRN28-10) Ensemble classification accuracy (%) vs. Attack Strength (ϵ) against Square attack for CIFAR-10 (b) Average number of queries required for square attack for all samples as well as successful (succ) attack samples with RN20

achieving the desired defense against adversarial attacks, with only minor trade-offs in clean accuracy. PARL primarily focuses on defending against black-box transfer attacks. These adversaries can craft adversarial examples using the complete network parameters, although such scenarios are rarely practical in real-world applications where only API based query access is provided for the target model.

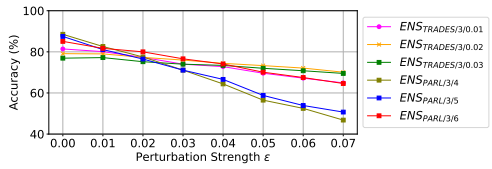
Performance evaluation against Query-based black-box attack: In a query-based adversarial attack, the attacker crafts adversarial examples by analyzing the responses from the target machine learning model, to which they have only black-box access. Our assessment of PARL involves its performance against a specific query-based black-box adversarial attack known as the Square Attack [1]). This attack method is notable for its ability to efficiently modify a minimal number of pixels within a square area of an image, effectively deceiving machine learning models while maintaining the overall visual integrity of the image. The Square Attack is unique in the AutoAttack suite [7], which comprises four different methods of adversarial attacks used for thorough model evaluation, with Square Attack being the sole black-box method.

In Fig. 5a, we illustrate the outcomes of using the Square Attack on PARL. For this experiment, we used the default maximum queries and square size settings, which are 5000 and 0.8, respectively and considered 1000 test images. We found that ensembles of Resnet20 and WideResnet28-10 trained with PARL outperformed the surrogate model. Additionally, we calculated the average number of queries utilized across all test images and specifically for those images which were successfully misclassified by the ensemble. Our findings, displayed in Fig. 5b, indicate that all models trained with PARL required over twice the number of queries compared to the surrogate model for all levels of perturbation, except at 0.01, where the requirement was approximately 1.5 times higher. PARL's ability to require significantly more queries for successful adversarial attacks, especially in comparison to the surrogate model, demonstrates its robustness in less query-restrictive environments.

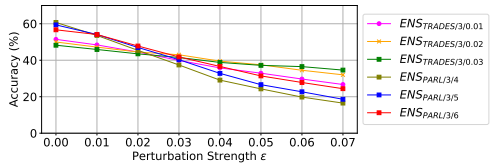
Performance comparison with Adversarial Training: We evaluate the performance of the PARL model in comparison with the adversarial training method TRADES [24]. The TRADES loss function is defined as:

$$\mathcal{L}_{\text{TRADES}}(x, y) = \mathcal{L}(x, y) + \beta \cdot \mathcal{L}(x + \delta, y) \qquad (2)$$

where, x is the natural input and y is its corresponding label. $\mathcal{L}(x, y)$ is the natural loss representing the model's prediction error on the clean data. δ represents the adversarial perturbation, typically computed using methods like PGD. $\mathcal{L}(x + \delta, y)$ is the adversarial loss, emphasizing correct classification of adversarial examples. β is a hyperparameter that balances the contributions of the two terms. This approach differs from standard adversarial training which often seeks to minimize the adversarial loss $\mathcal{L}(x + \delta, y)$ alone. By combining both the natural and adversarial losses, TRADES ensures robustness against adversarial attacks while maintaining performance on clean examples. We assume the default β value of 6 for our experiments. In Fig. 6a and Fig. 6b we show the comparison of three adversarially trained TRADES Resnet20 ensemble models $ENS_{TRADES/3/0.01}$, $ENS_{TRADES/3/0.02}$ and $ENS_{TRADES/3/0.03}$ (trained with different PGD attack perturbations $\epsilon = 0.01, 0.02, 0.03$) against $ENS_{PARL/3/4}$, $ENS_{PARL/3/5}$ and $ENS_{PARL/3/6}$ for CIFAR-10 and CIFAR-100 respectively. For CIFAR-10, we observe that at



(a)



(b)

Figure 6. Resnet20 Ensemble classification accuracy (%) vs. Attack Strength (ϵ) against black-box transfer attacks generated from surrogate ensemble for (a) CIFAR-10 and (b) CIFAR-100

$\epsilon = 0.07$ $ENS_{PARL/3/6}$ gives only 5.5% less robust accuracy than $ENS_{TRADES/3/0.02}$ and $ENS_{TRADES/3/0.03}$ with clean accuracy 5.97% and 8.17% higher than them respectively. Additionally, it gives same robust accuracy as $ENS_{TRADES/3/0.01}$ with 3.68% higher clean accuracy. We observe similar trends in the results for the CIFAR-100 dataset as well. In conclusion, PARL stands out as a preferred defense strategy against adversarial attacks on ensembles, offering similar or slightly lower robust accuracy compared to TRADES but with significantly higher clean accuracy, all achieved in less than one-third of TRADES's training time (cf. Sec. 3.2).

Table 1. Resnet20 Ensemble clean and robust ($\epsilon = 0.01$) classification accuracy on Tiny Imagenet Dataset

Model	Clean Accuracy	Robust Accuracy
ENS_U	60.85%	17.8%
ENS_{EIO}	57.33%	8.6%
$ENS_{TRADES/3/0.01}$	44.32%	27.7%
$ENS_{PARL/3/4}$	55.95%	28.3%
$ENS_{PARL+TRADES}$	42%	35.6%

Performance evaluation on Tiny Imagenet Dataset: We evaluated PARL on the Tiny Imagenet dataset, which contains 200 classes. Table 1 presents the clean and robust accuracy results for EIO, TRADES, PARL, and a combination of PARL and TRADES. The state-of-the-art method ENS_{EIO} underperformed, with both robust and clean accuracy falling below the baseline ensemble[3]. While $ENS_{TRADES/3/0.01}$ and $ENS_{PARL/3/4}$ showed similar robust accuracy, $ENS_{PARL/3/4}$ achieved better clean accuracy. Additionally, we also combined PARL and TRADES losses and observed an improvement in robust accuracy by nearly 7%, but clean accuracy dropped to 42%. Hence overall, in terms of clean accuracy and robust accuracy trade-off PARL performs the best among all methods. Overall, PARL offers the best trade-off between clean and robust accuracy among all methods tested.

[3]We used the open-source code provided with the EIO paper and reported all results based on runs using the default configuration described in the paper.

3.2. Ablation Study

In our previous evaluations, we train ENS_{PARL} by enforcing diversity in the first four, five and six convolution layers for all classifiers. Next, we provide an ablation study by analyzing a varying number of convolution layers considered for diversity training. We consider three ensembles, $ENS_{PARL/3/4}$, $ENS_{PARL/3/5}$, and $ENS_{PARL/3/6}$, for this study. Accuracies of all ensembles on clean examples for Resnet20 and WideResnet28-10 are mentioned in Table 2. We observe that as fewer restrictions are imposed, overall ensemble accuracy increases, which is expected and can be followed from Equation (1). We also present a layer-wise CKA analysis for each pair of classifiers within $ENS_{PARL/3/5}$ and $ENS_{PARL/3/6}$, trained with CIFAR-10. The layer-wise CKA values are shown in Fig. 7 to exhibit the effect of PARL on diversity. We observe a decline in the CKA values for more layers in case of $ENS_{PARL/3/6}$ compared to $ENS_{PARL/3/5}$, which is expected as $ENS_{PARL/3/6}$ is trained by restricting more convolution layers. We also observe that each pair of classifiers show more overall diversity in $ENS_{PARL/3/6}$ than in $ENS_{PARL/3/5}$. The overall average of CKA values are mentioned inside braces within figure legends.

Table 2. Ensemble classification accuracy (%) for Resnet20 and WideResnet28-10 on CIFAR-10 and CIFAR-100 clean examples.

Model	Dataset	$ENS_{PARL/3/4}$	$ENS_{PARL/3/5}$	$ENS_{PARL/3/6}$
Resnet20	CIFAR-10	88.53	87.49	85.79
	CIFAR-100	60.73	59.37	56.68
WideResnet28-10	CIFAR-10	92.86	92.31	91.98

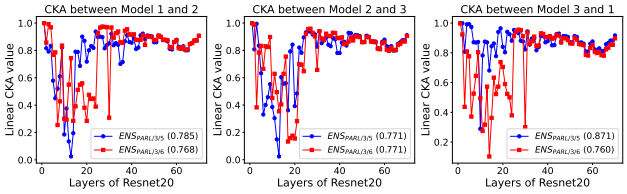
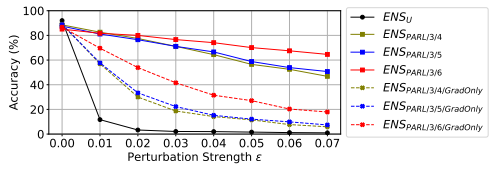


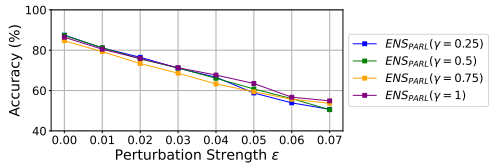
Figure 7. Layer-wise linear CKA values between each pair of PARL/3/5 and PARL/3/6 models trained with CIFAR-10 on Resnet20 showing the diversity at each layer.

Contribution of correlation term: In the defined PARL loss function (see Equation 1), we incorporate a penalty term that combines the cosine similarities of gradients with the correlation of outputs from distinct sub-model pairs at specific convolution layers. Fig. 8a illustrates the variance in PARL's robustness when the penalty term solely relies on the cosine similarities between gradients, excluding the output correlations ($ENS_{PARL/3/N/GradOnly}$). While $ENS_{PARL/3/N/GradOnly}$ models do show improvements over ENS_U, it's evident that $ENS_{PARL/3/N}$ models are superior, emphasizing the critical role of both gradient similarity and output correlation in the penalty term.

Selection of γ: In PARL Equation 1, γ is used for the purpose of regulating the PARL penalty term. We experimented with varying γ values, focusing on their effect on



(a)



(b)

Figure 8. (a) Comparing PARL robustness against PARL loss with penalty term that only uses gradient similarity. (b) Comparing clean and robust accuracies of Resnet20 PARL ensemble trained with different γ

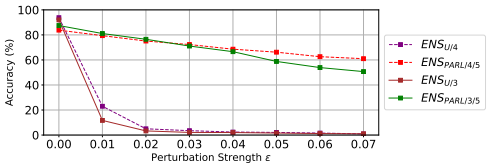


Figure 9. Resnet20 Ensemble classification accuracy (%) vs. Attack Strength (ϵ) for CIFAR-10 with increased number or classifiers

the model's clean and robust accuracies. The findings for Resnet-20 $ENS_{PARL/3/5}$, illustrated in Fig. 8b, reveal that while accuracy fluctuations are minimal across different perturbations, lower γ values tend to enhance clean accuracy, whereas higher γ values improve robust accuracy. For $\gamma = 1$, we obtain a clean accuracy of 86.42% and robust accuracy of 54.9% at $\epsilon = 0.07$, whereas we for $\gamma = 0.25$ we obtain increased clean accuracy of 87.49%, and decreased robust accuracy of 50.7%. For our experiments, we selected $\gamma = 0.25$ as it offers the best trade-off between clean and robust accuracy.

Increased number of classifiers: In Fig. 9, we present results obtained by increasing the number of classifiers from three to four. We observe that robust accuracy improves by 10.3% for $ENS_{PARL/4/5}$ compared to $ENS_{PARL/3/5}$ with $\epsilon = 0.07$, with 28.8% more training time, as discussed next. We opt for three classifiers throughout the paper as a trade-off between robust accuracy and train time.

Training time overhead: In Table 3 we give training time per epoch for the Resnet20 surrogate, PARL, EIO and TRADES models. Earlier we compared PARL with TRADES and observed that PARL gives similar or slightly lesser robust accuracy than TRADES but in trade-off provides much higher clean accuracy as well. TRADES also takes 3x more training time than the most computationally expensive $ENS_{PARL/3/6}$ among all PARL models and 12x compared to the surrogate model. Additionally, ENS_{EIO} takes 7x training time compared to surrogate model and 3x compared to $ENS_{PARL/3/4}$ which gives similar clean accuracy and much higher robust accuracy compared to

Table 3. Training time (sec/epoch) for CIFAR-10.

Model	Training Time	Model	Training Time
ENS_U	30	$ENS_{PARL/3/5}$	90
ENS_{EIO}	210	$ENS_{PARL/3/6}$	120
ENS_{TRADES}	370	$ENS_{PARL/4/5}$	116
$ENS_{PARL/3/4}$	65		

ENS_{EIO}. Lastly, we observe that increasing the number of classifiers from three to four has minimal impact on training time, as shown for $ENS_{PARL/4/5}$ and $ENS_{PARL/4/6}$.

Discussion: ADP forces different models in an ensemble to have mutually orthogonal non-maximal predictions. GAL reduces the dimension of adversarial sub-space shared between different models using uncorrelated loss functions. These methods do not inherently enforce diversity on decision boundaries learned by the models. DVERGE diversifies non-robust input features of models by performing adversarial training, making it more robust against weak attack strength and less robust against strong attack strength. EIO leverages random gated networks to enhance adversarial robustness by diversifying vulnerabilities across multiple paths of CNNs but again has a higher training overhead. In contrast, PARL, by construction, forces the models to have high diversity in decision boundaries using all intermediate feature space. The diversity of models attained through intermediate feature space (not limited to only non-robust input features) makes PARL more robust, even for strong attack strength. In addition, PARL produces robust ensembles without substantially impacting clean example accuracy and training time.

4. Conclusion

This paper proposes a new approach that, by construction, produces an ensemble of neural networks with diverse decision boundaries, making it robust against adversarial attacks. The diversity is obtained through the proposed Pairwise Adversarially Robust Loss (PARL) function utilizing the gradients and outputs of each layer in all the networks simultaneously. Experimental results show that PARL can significantly improve the overall robustness of an ensemble in comparison to previous approaches against state-of-the-art black-box transfer attacks as well as query-based black-box attacks without substantially impacting clean example accuracy. In particular, PARL achieves a 24.8% improvement in robust accuracy over the leading ensemble defense method EIO with highest perturbation strength. Furthermore, when compared to TRADES, PARL demonstrates robust accuracy of similar order with a 3.68% increase in clean accuracy. PARL also takes lesser training time compared to both EIO and TRADES method.

Acknowledgment

This work was supported in part by Prime Minister's Research Fellowship, granted by Government of India.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, volume 12368 of *Lecture Notes in Computer Science*, pages 484–501. Springer, 2020. 3, 6

[2] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsm"assan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 2018. 1

[3] Yi Cai, Xuefei Ning, Huazhong Yang, and Yu Wang. Ensemble-in-one: Ensemble learning within random gated networks for enhanced adversarial robustness. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14738–14747. AAAI Press, 2023. 2, 5

[4] Nicholas Carlini and David A. Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, TX, USA*. 1

[5] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA*. 1

[6] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Trans. Intell. Technol.*, 6(1):25–45, 2021. 1

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. 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 2206–2216. PMLR, 2020. 6

[8] Devvrit, Minhao Cheng, Cho-Jui Hsieh, and Inderjit S. Dhillon. Voting based ensemble improves robustness of defensive models. *CoRR*, abs/2011.14031, 2020. 5

[9] Shixiang Gu and Luca Rigazio. Towards Deep Neural Network Architectures Robust to Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*. 1

[10] Sanjay Kariyappa and Moinuddin K. Qureshi. Improving Adversarial Robustness of Ensembles with Diversity Training. *CoRR*, abs/1901.09981, 2019. 2, 5

[11] 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. 5

[12] Xin Li and Fuxin Li. Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy*, 2017. 1

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2017. 3, 5

[14] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 1

[15] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. 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 4970–4979. PMLR, 2019. 2, 5

[16] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 582–597. IEEE Computer Society, 2016. 1

[17] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *CoRR*, abs/1709.03423, 2017. 1

[18] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[19] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1

[20] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 3, 5

[21] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 2730–2739. Computer Vision Foundation / IEEE, 2019. 3, 5

[22] Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 5

[23] Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin I. P. Rubinstein, Ce Zhang, and Bo Li. TRS: transferability reduced ensemble via promoting gradient diversity and model smoothness. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17642–17655, 2021. 2, 5

[24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. 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 7472–7482. PMLR, 2019. 6

[25] Shaofeng Zhang, Meng Liu, and Junchi Yan. The diversified ensemble neural network. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16001–16011. Curran Associates, Inc., 2020. 1