Improving Ensemble Robustness by Collaboratively Promoting and Demoting Adversarial Robustness

Anh Tuan Bui[1], **Trung Le**[1], **He Zhao**[1]
Paul Montague[2], **Olivier deVel**[2], **Tamas Abraham**[2], **Dinh Phung**[1]
[1]Monash University, Australia
[2]Defence Science and Technology Group, Australia
tuananh.bui@monash.edu

Abstract

Ensemble-based adversarial training is a principled approach to achieve robustness against adversarial attacks. An important technique of this approach is to control the transferability of adversarial examples among ensemble members. We propose in this work a simple yet effective strategy to collaborate among committee models of an ensemble model. This is achieved via the secure and insecure sets defined for each model member on a given sample, hence help us to quantify and regularize the transferability. Consequently, our proposed framework provides the flexibility to reduce the adversarial transferability as well as to promote the diversity of ensemble members, which are two crucial factors for better robustness in our ensemble approach. We conduct extensive and comprehensive experiments to demonstrate that our proposed method outperforms the state-of-the-art ensemble baselines, at the same time can detect a wide range of adversarial examples with a nearly perfect accuracy. Our code is available at: https://github.com/tuananhbui89/Crossing-Collaborative-Ensemble.

Introduction

Deep neural networks have experienced great success in many disciplines (Goodfellow, Bengio, and Courville 2016), such as computer vision (He et al. 2016), natural language processing and speech processing (Vaswani et al. 2017). However, even the state-of-the-art models are reported to be vulnerable to adversarial attacks (Biggio et al. 2013; Goodfellow, Shlens, and Szegedy 2015; Szegedy et al. 2014; Carlini and Wagner 2017; Madry et al. 2018; Athalye, Carlini, and Wagner 2018), which is of significant concern given the large number of applications of deep learning in real-world scenarios. It is thus urgent to develop deep learning models that are robust against different types of adversarial attacks. To this end, several adversarial defense methods have been developed but typically addressing the robustness within a single model (e.g., Papernot et al. 2016; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Madry et al. 2018; Qin et al. 2019; Shafahi et al. 2019). To cater for more diverse types of attacks, recent work, notably (He et al. 2017; Tramèr et al. 2018; Strauss et al. 2017; Liu et al. 2018; Pang et al. 2019), has shown that ensemble learning can strengthen robustness significantly.

Despite initial success, key principles for ensemble-based adversarial training (EAT) largely remain open. One crucial challenge is to achieve minimum 'transferability' between committee members to increase robustness for the overall ensemble model (Papernot, McDaniel, and Goodfellow 2016; Liu et al. 2016; Tramèr et al. 2018; Pang et al. 2019; Kariyappa and Qureshi 2019). In (Kariyappa and Qureshi 2019), robustness was achieved by aligning the gradient of committee members to be diametrically opposed, hence reducing the shared adversarial spaces (Tramèr et al. 2017), or the transferability. However, the method in (Kariyappa and Qureshi 2019) was designed for black-box attacks, thus still vulnerable to white-box attacks. Furthermore, attempting to achieve gradient alignment is unreliable for high-dimensional datasets and it is difficult to extend for ensemble with more than two committee members. More recently (Pang et al. 2019) proposed to promote the diversity of non-maximal predictions of the committee members (i.e., the diversity among softmax probabilities except the highest ones) to reduce the adversarial transferability among them. Nonetheless, the central concept of transferability has not been systematically addressed.

Our proposed work here will first make the concept of adversarial transferability concrete via the definitions of secure and insecure sets. To reduce the adversarial transferability and increase the model diversity, we aim to make the insecure sets of the committee models as disjoint as possible (i.e., lessening the overlapping of those regions) and challenge those committee members with divergent sets of adversarial examples. In addition, we observe that lessening the adversarial transferability alone is not sufficient to ensure accurate predictions of the ensemble model because the committee member that offers inaccurate predictions might dominate the final decisions. With this in mind, we propose to realize what we call a "transferring flow" by collaborating robustness promoting and demoting operations. Our key principle to coordinate the promoting and demoting operations is to promote the prediction of one model on a given adversarial example and to demote the prediction of another model on this example so as to maximally lessen the negative impact of the wrong predictions and ensure the correct predictions of the ensemble model. Moreover, different from

	$\mathbf{x}_a \in \mathcal{B}_{\text{se}}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon)$		$\mathbf{x}_a \in \mathcal{B}_{\text{in}}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon)$
$\mathbf{x}_a \in \mathcal{B}_{\text{se}}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon)$	S_{11}	$\Leftarrow$	S_{01}
	$\Uparrow$		$\Uparrow$
$\mathbf{x}_a \in \mathcal{B}_{\text{in}}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon)$	S_{10}	$\Leftarrow$	S_{00}

Table 1: Four subsets of the ensemble model and the transferring flow (arrows). $\mathcal{B}_{se}/\mathcal{B}_{\text{in}}$ represent for secure/insecure sets, respectively.

other works (Strauss et al. 2017; Pang et al. 2019; Kariyappa and Qureshi 2019) which only consider adversarial examples of the ensemble model, the committee members in our ensemble model are exposed to various divergent adversarial example sets, which inspire them to become gradually more divergent. Interestingly, by strengthening demoting operations, our method is capable to assist better detection of adversarial examples. In brief, our contributions in this work include:

- We propose a simple but efficient collaboration strategy to reduce the transferability among ensemble members.
- We propose two variants of our method: the robust oriented variant, which helps to improve the adversarial robustness and the detection oriented variant, which can detect adversarial examples with high predictive performance.
- We conduct extensive and comprehensive experiments to demonstrate the improvement of our proposed method over the state-of-the-art defense methods.
- We provide a further understanding of the relationship between the transferability and the overall robustness in ensemble learning context.

Our Proposed Method

In this section, we present our ensemble collaboration strategy, which allows us to collaborate many committee models for improving the ensemble robustness. We start with the definitions and some key properties of secure and insecure sets which later support us in devising promoting and demoting operations for collaborating the committee models to achieve the ensemble robustness. It is worth noting that our ensemble strategy is applicable for ensembling an arbitrary number of committee models; here we focus on presenting the key theories, principles, and operations for the canonical case of ensembling two models for better readability.

Secure and Insecure Sets

Consider a classification problem on a dataset $\mathcal{D}$ with M classes and a pair $(\mathbf{x}, \boldsymbol{y})$ that represents a data example $\mathbf{x}$ and its true label $\boldsymbol{y}$ which is sampled from the dataset $\mathcal{D}$. Given a model f, the crucial aim of defense is to make f robust by giving consistently accurate predictions over a ball, $\mathcal{B}(\mathbf{x}, \epsilon) \coloneqq \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon\}$ around a benign data example $\mathbf{x}$, for every possible $\mathbf{x}$ in the dataset $\mathcal{D}$ and the distortion boundary ϵ. To further clarify and motivate our theory, we define

$$\mathcal{B}_{\text{secure}}(\mathbf{x}, \boldsymbol{y}, f, \epsilon) \coloneqq \{\mathbf{x}' \in \mathcal{B}(\mathbf{x}, \epsilon) : \text{argmax}_i f_i(\mathbf{x}') = \boldsymbol{y}\},$$
$$\mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f, \epsilon) \coloneqq \{\mathbf{x}' \in \mathcal{B}(\mathbf{x}, \epsilon) : \text{argmax}_i f_i(\mathbf{x}') \neq \boldsymbol{y}\}.$$

Intuitively, we define a *secure* set $\mathcal{B}_{\text{secure}}(\mathbf{x}, \boldsymbol{y}, f, \epsilon)$ as the set of elements in the ball $\mathcal{B}(\mathbf{x}, \epsilon)$ for which the classifier f makes the correct prediction. In addition, we define the *insecure* set $\mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f, \epsilon)$ as the set of elements in the ball $\mathcal{B}(\mathbf{x}, \epsilon)$ for which f predicts differently from the true label $\boldsymbol{y}$. By definition, the secure set is the complement of the insecure set, and $\mathcal{B}(\mathbf{x}, \varepsilon) = \mathcal{B}_{\text{secure}}(\mathbf{x}, \boldsymbol{y}, f, \varepsilon) \bigcup \mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f, \varepsilon)$. It is clear that the aim of improving adversarial robustness is to train the classifier f in such the way that $\mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f, \epsilon)$ is either as small as possible (ideally, $\mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f, \epsilon) = \emptyset$, $\forall \mathbf{x} \in \mathcal{D}$) or makes an adversary hard to generate adversarial examples in it. The following simple lemma (see the proof in the supplementary material) shows the connection between those two kinds of sets and the robustness of the ensemble model and facilitates the development of our proposed method.

Lemma 1. *Let us define $f^{en}(\cdot) = \frac{1}{2}f^1(\cdot) + \frac{1}{2}f^2(\cdot)$ for two given models f^1 and f^2. If f^1 and f^2 predict an example $\mathbf{x}$ accurately, we have the following:*

i) $\mathcal{B}_{insecure}(\mathbf{x}, \boldsymbol{y}, f^{en}, \epsilon) \subset \mathcal{B}_{insecure}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon) \cup \mathcal{B}_{insecure}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon)$.

ii) $\mathcal{B}_{secure}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon) \cap \mathcal{B}_{secure}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon) \subset \mathcal{B}_{secure}(\mathbf{x}, \boldsymbol{y}, f^{en}, \epsilon)$.

Dual Collaborative Ensemble

Transferring Flow. Consider the canonical case of an ensemble consisting of two models: $f^{en}(\cdot) = \frac{1}{2}f^1(\cdot) + \frac{1}{2}f^2(\cdot)$, where f^{en} is the ensemble model and $\{f^1, f^2\}$ is the set of ensemble committee (or the committee). Based on the definitions of secure and insecure sets, an arbitrary adversarial example $\mathbf{x}_a$ must lie in one of four subsets as shown in Table 1. Let us further clarify these subsets. In the first subset $S_{11} = \mathcal{B}_{\text{secure}}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon) \bigcap \mathcal{B}_{\text{secure}}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon)$, the example $\mathbf{x}_a$ is predicted correctly by both models, hence also by the ensemble model f^{en} (Lemma 1 (ii)). The subsets S_{10}, S_{01} are the intersection of a secure set of one model and an insecure set of another model, hence an example of two sets is predicted correctly by one model and incorrectly by the other. Lastly, in the subset $S_{00} = \mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon) \bigcap \mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon)$, both models offer predictions other than the true label, but there is also no guarantee that their incorrect predictions are in the same class. There is still a chance that the incorrect prediction in subset S_{10}, S_{01} dominates the correct ones, which leads to the incorrect prediction on average. Therefore, the insecure region of the overall ensemble should be related to the union $S_{10} \cup S_{01} \cup S_{00}$ or the total volume (i.e., $|S_{10}| + |S_{01}| + |S_{00}|$) of the subsets S_{10}, S_{01}, S_{00}.

As the result, to obtain a robust ensemble model, we need to maintain the subset S_{00} as small as possible, which is in turn equivalent to making the insecure regions of the two models as disjoint as much as possible (i.e., concurred with Lemma 1 (i)). For the data points in either S_{10} or S_{01}, we need to increase the chance that the correct predictions dominate the incorrect ones. Our approach is to encourage adversarial examples inside S_{00} to move to the subsets S_{10}, S_{01} during the course of training, and those of S_{10}, S_{01} to move

to the subset S_{11}. We term this movement as the *transferring flow*, which is described in Table 1. In what follows, we present how to implement the transferring flow for our ensemble model.

Promoting Adversarial Robustness (PO). We refer to promoting adversarial robustness as an operation to leverage the information of an example $\mathbf{x}_a^i$ (adversarial example of model f^i) for improving the robustness of a model f^j (i, j can be different). There are several adversarial defense methods that can be applied to promote adversarial robustness, notably (Madry et al. 2018; Zhang et al. 2019; Qin et al. 2019). In this work, to promote the adversarial robustness of a given adversarial example $\mathbf{x}_a^i$ w.r.t the model f^j, we use adversarial training (Madry et al. 2018) by minimizing the cross-entropy loss w.r.t the true label as $\min\ \mathcal{C}\left(f^j(\mathbf{x}_a^i), \boldsymbol{y}\right)$. After undertaking this PO, $\mathbf{x}_a^i$ is expected to move to the secure set $\mathcal{B}_{\text{secure}}\left(\mathbf{x}, \boldsymbol{y}, f^j, \epsilon\right)$. We introduce two types of PO: direct PO (dPO) when $i = j$ and crossing PO (cPO) when $i \neq j$.

Demoting Adversarial Robustness (DO). In contrast to promoting adversarial robustness, we refer to demoting adversarial robustness as an operation to sacrifice the robustness of a model for an example $\mathbf{x}_a^i$ (adversarial example of model f^i). Here, we demote the adversarial robustness of a given adversarial example $\mathbf{x}_a^i$ w.r.t the model f^j by $\max\ \mathcal{H}\left(f^j(\mathbf{x}_a^i)\right)$ where $\mathcal{H}$ is the entropy. Without any further knowledge, the prediction is likely uniformly distributed, hence the example $\mathbf{x}_a^i$ likely falls into the insecure set $\mathcal{B}_{\text{insecure}}\left(\mathbf{x}, \boldsymbol{y}, f^j, \epsilon\right)$ instead of the secure set $\mathcal{B}_{\text{secure}}\left(\mathbf{x}, \boldsymbol{y}, f^j, \epsilon\right)$.

Collaboration of the Promoting and Demoting Operations. We now present how to coordinate PO/DO to enforce the transferring flow for enhancing the adversarial robustness of the ensemble model in the canonical case of a committee of two members $\{f^1, f^2\}$, parameterized by θ_1 and θ_2. Let $\mathbf{x}_a^1$ and $\mathbf{x}_a^2$ be white-box adversarial examples of f^1 and f^2 respectively. With a strong adversary, we can assume that $\mathbf{x}_a^1 \in \mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f^1, \epsilon)$ (i.e., $\mathbf{x}_a^1 \in S_{01} \cup S_{00}$) and $\mathbf{x}_a^2 \in \mathcal{B}_{\text{insecure}}(\mathbf{x}, \boldsymbol{y}, f^2, \epsilon)$ (i.e., $\mathbf{x}_a^2 \in S_{10} \cup S_{00}$). For ease of comprehensibility, we present the treatment for $\mathbf{x}_a^1$ and the same treatment is applied to $\mathbf{x}_a^2$. To strengthen model f^1, we always use $\mathbf{x}_a^1$ to promote the robustness of model f^1 by minimizing the cross-entropy loss $\mathcal{C}\left(f^1(\mathbf{x}_a^1), \boldsymbol{y}\right)$ (i.e., flow $S_{01} \Rightarrow S_{11}$ or $S_{00} \Rightarrow S_{10}$). Meanwhile, we consider two cases of $\mathbf{x}_a^1$ w.r.t model f^2: i) being correctly predicted by f^2 (i.e., $\mathbf{x}_a^1 \in S_{01}$) and ii) being incorrectly predicted by f^2 (i.e., $\mathbf{x}_a^1 \in S_{00}$). For the first case, we use $\mathbf{x}_a^1$ to promote model f^2 to make sure $\mathbf{x}_a^1$ stays in the secure set of model f^2 (i.e., $S_{11} \cup S_{01}$). For the second case, we demote $\mathbf{x}_a^1$ w.r.t f^2 by maximizing the entropy $\mathcal{H}\left(f^2(\mathbf{x}_a^1)\right)$ in order to keep $\mathbf{x}_a^1$ in the insecure set of model f^2 (i.e., $S_{10} \cup S_{00}$).

Therefore, with the collaboration of two models f^1 and f^2 on the same example $\mathbf{x}_a^1$, we deploy either flow $S_{01} \Rightarrow S_{11}$ or $S_{00} \Rightarrow S_{10}$ depending on the scenario of $\mathbf{x}_a^1$. It is worth noting that DO encourages $f^2(\mathbf{x}_a^1)$ to be close to the uniform prediction, hence causing a minimal effect on the ensemble prediction $f^{en}\left(\mathbf{x}_a^1\right)$. As a consequence, $f^{en}\left(\mathbf{x}_a^1\right) = \frac{1}{2}\left(f^1\left(\mathbf{x}_a^1\right) + f^2\left(\mathbf{x}_a^1\right)\right)$ is dominated by $f^1\left(\mathbf{x}_a^1\right)$, which likely offers a correct prediction via the corresponding PO: $\min\ \mathcal{C}\left(f^1(\mathbf{x}_a^1), \boldsymbol{y}\right)$. We summarize the PO/DO to deploy the transferring flow in Table 2.

Scenario	f^1	f^2
$\mathbf{x}_a^1 \in S_{01}$	$\min \mathcal{C}\left(f^1(\mathbf{x}_a^1), \boldsymbol{y}\right)$	$\min \mathcal{C}\left(f^2(\mathbf{x}_a^1), \boldsymbol{y}\right)$
$\mathbf{x}_a^1 \in S_{00}$	$\min \mathcal{C}\left(f^1(\mathbf{x}_a^1), \boldsymbol{y}\right)$	$\max \mathcal{H}\left(f^2(\mathbf{x}_a^1)\right)$
$\mathbf{x}_a^2 \in S_{10}$	$\min \mathcal{C}\left(f^1(\mathbf{x}_a^2), \boldsymbol{y}\right)$	$\min \mathcal{C}\left(f^2(\mathbf{x}_a^2), \boldsymbol{y}\right)$
$\mathbf{x}_a^2 \in S_{00}$	$\max \mathcal{H}\left(f^1(\mathbf{x}_a^2)\right)$	$\min \mathcal{C}\left(f^2(\mathbf{x}_a^2), \boldsymbol{y}\right)$

Table 2: Promoting and demoting operations for the transferring flow

The objective functions for model f^1 and f^2 to deploy the transferring flow are:

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \boldsymbol{y}, \theta_1) &= \mathcal{C}\left(f^1(\mathbf{x}), \boldsymbol{y}\right) + \mathcal{C}\left(f^1(\mathbf{x}_a^1), \boldsymbol{y}\right) \\ &\quad + \lambda_{pm}\mathbb{I}\left(f^1(\mathbf{x}_a^2), \boldsymbol{y}\right)\mathcal{C}(f^1(\mathbf{x}_a^2), \boldsymbol{y}) \\ &\quad - \lambda_{dm}\left(1 - \mathbb{I}\left(f^1(\mathbf{x}_a^2), \boldsymbol{y}\right)\right)\mathcal{H}\left(f^1(\mathbf{x}_a^2)\right), \quad (1) \\ \mathcal{L}(\mathbf{x}, \boldsymbol{y}, \theta_2) &= \mathcal{C}\left(f^2(\mathbf{x}), \boldsymbol{y}\right) + \mathcal{C}\left(f^2(\mathbf{x}_a^2), \boldsymbol{y}\right) \\ &\quad + \lambda_{pm}\mathbb{I}\left(f^2(\mathbf{x}_a^1), \boldsymbol{y}\right)\mathcal{C}(f^2(\mathbf{x}_a^1), \boldsymbol{y}) \\ &\quad - \lambda_{dm}\left(1 - \mathbb{I}\left(f^2(\mathbf{x}_a^1), \boldsymbol{y}\right)\right)\mathcal{H}\left(f^2(\mathbf{x}_a^1)\right). \quad (2)\end{aligned}$$

where λ_{pm} and λ_{dm} are the hyper-parameters for promoting and demoting effects, respectively, and $\mathbb{I}\left(f^1(\mathbf{x}_a^2), \boldsymbol{y}\right)$ is the indicator to indicate whether $\mathbf{x}_a^2$ is predicted correctly (i.e., $\mathbb{I} = 1$, hence $\mathbf{x}_a^2 \in S_{10}$) or incorrectly (i.e., $\mathbb{I} = 0$, hence $\mathbf{x}_a^2 \in S_{00}$) by f^1, which helps to switch on/off the cPO/DO for model f^1.

For the final objective function, we approximate the hard indicator $\mathbb{I}\left(f^1(\mathbf{x}_a^2), \boldsymbol{y}\right)$ by the soft version $f^1_{\boldsymbol{y}}(\mathbf{x}_a^2) = p\left(\boldsymbol{y} \mid \mathbf{x}_a^2, f^1\right)$, which represents the probability the model f^1 assigning $\mathbf{x}_a^2$ to the label $\boldsymbol{y}$. We hence arrive at the following objective functions for both f^1 and f^2, respectively.

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \boldsymbol{y}, \theta_1) &= \mathcal{C}\left(f^1(\mathbf{x}), \boldsymbol{y}\right) + \mathcal{C}\left(f^1(\mathbf{x}_a^1), \boldsymbol{y}\right) \\ &\quad + \lambda_{pm} f^1_y(\mathbf{x}_a^2)\mathcal{C}(f^1(\mathbf{x}_a^2), \boldsymbol{y}) \\ &\quad - \lambda_{dm}\left(1 - f^1_y(\mathbf{x}_a^2)\right)\mathcal{H}\left(f^1(\mathbf{x}_a^2)\right), \quad (3) \\ \mathcal{L}(\mathbf{x}, \boldsymbol{y}, \theta_2) &= \mathcal{C}\left(f^2(\mathbf{x}), \boldsymbol{y}\right) + \mathcal{C}\left(f^2(\mathbf{x}_a^2), \boldsymbol{y}\right) \\ &\quad + \lambda_{pm} f^2_y(\mathbf{x}_a^1)\mathcal{C}(f^2(\mathbf{x}_a^1), \boldsymbol{y}) \\ &\quad - \lambda_{dm}\left(1 - f^2_y(\mathbf{x}_a^1)\right)\mathcal{H}\left(f^2(\mathbf{x}_a^1)\right). \quad (4)\end{aligned}$$

We note that in our implementation, the soft indicators $f^1_y(\mathbf{x}_a^2)$ and $f^2_y(\mathbf{x}_a^1)$ are used as values by performing a stopping gradient to prevent the back-propagation process to go inside them for further updating f^1 and f^2.

Crossing Collaborative Ensemble

We now extend our collaboration strategy to enable us to ensemble many individual members, which we term as a *Crossing Collaborative Ensemble (CCE)*. Specifically, given an ensemble of N members $f^{en}(\cdot) = \frac{1}{N}\sum_{n=1}^{N} f^n(\cdot)$ parameterized by θ_n, the loss function for a model $f^n, n \in [1, N]$ as follow:

$$\mathcal{L}^n(\mathbf{x}, \boldsymbol{y}, \theta_n) = \mathcal{C}\left(f^n(\mathbf{x}), \boldsymbol{y}\right) + \mathcal{C}\left(f^n(\mathbf{x}_a^n), \boldsymbol{y}\right) + \frac{1}{N-1}\sum_{i \neq n}\Bigg(\lambda_{pm} f_y^n(\mathbf{x}_a^i)\mathcal{C}(f^n(\mathbf{x}_a^i), \boldsymbol{y}) - \lambda_{dm}\left(1 - f_y^n(\mathbf{x}_a^i)\right)\mathcal{H}\left(f^n(\mathbf{x}_a^i)\right)\Bigg). \quad (5)$$

It appears from the above loss that we encourage each individual model to (i) minimize the loss of the adversarial example itself for improving its robustness (dPO) and (ii) promoting or demoting its robustness (cPO/DO) with other adversarial examples depending on the soft indicator.

Connections to Traditional Ensemble Learning. Firstly, in our method, N members $\{f^n\}$ are reinforced with the joint of $N+1$ data sources: clean data $\{\mathbf{x}\}$ and N adversarial examples $\{\mathbf{x}_a^n\}_{n=1}^N$. However, depending on different scenarios, they have the same task (PO-PO) or opposite tasks (PO-DO) on the same adversarial set $\{\mathbf{x}_a^n\}$. Our approach can be linked to the bagging technique in the literature, in which each classifier was trained on different sets of data. Secondly, by assigning opposite tasks for ensemble members, our method produces a negative correlation which was described in (Liu and Yao 1999; Kuncheva and Whitaker 2003; Bagnall, Bunescu, and Stewart 2017). It has been claimed that negative relationship among ensemble members can further improve the ensemble accuracy better than the independent correlation.

Experiments

In this section, we first introduce the experimental setting for adversarial defenses and attackers followed by an extensive evaluation to compare our method with state-of-the-art adversarial defenses. We show that our method surpasses these methods for common benchmark datasets. Next, we provide an ablation study to understand the transferability among ensemble members of adversarial examples. Finally, we show that our method not only detects adversarial examples accurately and consistently but also predicts benign examples with a significant improvement.

Experimental Setting

General Setting. We use CIFAR10 and CIFAR100 as the benchmark datasets in our experiment.[1] Both datasets have 50,000 training images and 10,000 test images. The inputs were normalized to $[0, 1]$. We apply random horizontal flips and random shifts with scale 10% for data augmentation as used in (Pang et al. 2019). We use both standard CNN architecture and ResNet architecture (He et al. 2016) in our experiment. The architecture and training setting for each dataset are provided in our supplementary material.

Crafting Adversarial Examples for Defenders. In our experiments, we use PGD $\{k, \epsilon, \eta, l_\infty\}$ as the common adversary to generate adversarial examples for the adversarial training of all defenders where k is the iteration steps, ϵ is the distortion bound and η is the step size. Specifically, the configuration for the CIFAR10 dataset is $k = 10, \epsilon = 8/255, \eta = 2/255$ and that for the CIFAR100 dataset is $k = 10, \epsilon = 0.01, \eta = 0.001$. For the CIFAR10 dataset with ResNet architecture, we use the same setting in (Pang et al. 2019) which is $k = 10, \epsilon \sim U(0.01, 0.05), \eta = \epsilon/10$.

Baseline Methods. Because the model capacity has significant impact on the inference performance, therefore, for a fair comparison, we compare our method with the start-of-the-art ensemble-based method, i.e., ADV-EN (Madry et al. 2018) and ADP (Pang et al. 2019), which have the same number of committee members and also the member's architecture. More specifically, ADV-EN is the variant of PGD adversarial training method (ADV) in the context of ensemble learning, in which the entire ensemble model is treated as one unified model applied with adversarial training. We also compare with the ADV method which is adversarial training on a single model. For ADP, we choose the best setting $ADP_{2,0.5}$ with adversarial version, which was reported in the paper (Pang et al. 2019), and use the official code.[2]

Throughout our experiments, we use two variants of our method: (i) Robustness Mode (i.e., CCE-RM) for which we set $\lambda_{pm} = \lambda_{dm} = 1$ and (ii) Detection Mode (i.e., CCE-DM) for which we disable cPO ($\lambda_{pm} = 0$) and strengthen DO (i.e., $\lambda_{dm} = 5$).

Attack Setting. We use different state-of-the-art attacks to evaluate the defense methods including:

(i) **Gradient based attacks** (with *cleverhans*[3] lib). We use PGD (Madry et al. 2018), the Basic Iterative Method (BIM) (Kurakin, Goodfellow, and Bengio 2017) and the Momentum Iterative Method (MIM) (Dong et al. 2018). They share the same hyper-parameters configuration, i.e., $\{k, \epsilon, \eta\}$, which is described in each individual experiment.

(ii) **B&B attack** (Brendel et al. 2019) (with *foolbox*[4] lib) which is a decision based attack. We argue that the B&B attack setting in the paper of (Tramer et al. 2020) may not be appropriate to evaluate the ADP method. It is because the ADP method used PGD ($\epsilon \sim U(0.01, 0.05), k = 10$) for its adversarial training, while B&B attack used PGD ($\epsilon = 0.15, k = 20$) as an initialized attack which is much stronger than the defense capacity. More specifically, the initialized PGD attack alone can reduce the accuracy to 0.1%. Therefore, B&B attack contributes very little to the final attack performance. To have a fair evaluation, we use two initialized attacks with lower strength: PGD1 ($\epsilon = 8/255, \eta = 2/255, k = 20$) and PGD2 ($\epsilon = 16/255, \eta = 2/255, k = 20$) then apply B&B attack with 100 steps and repeat for three times. It is worth noting that, PGD2 is still much stronger than the defense capacity, however, we use this setting to mimic the evaluation in the paper of (Tramer et al.

[1] Recently, (Tsipras et al. 2020) found the labeling issue in the ImageNet dataset, which highly affects the fairness of robustness evaluation on this dataset.

[2] https://github.com/P2333/Adaptive-Diversity-Promoting

[3] https://github.com/tensorflow/cleverhans

[4] https://foolbox.readthedocs.io/en/stable/

Attack	ADV_1	ADV_2	ADP_2	CCE_2	ADV_3	ADP_3	CCE_3
Nat. acc.	83.9	85.3	**85.3**	84.5	86.1	**86.2**	84.9
PGD	41.4	42.8	44.2	**45.8**	43.8	45.1	**48.6**
BIM	41.5	42.9	44.1	**45.8**	44.0	45.2	**48.8**
MIM	41.9	43.3	44.8	**46.3**	44.5	45.7	**49.1**
B&B (PGD1)	37.0	38.3	37.3	**42.2**	39.3	38.3	**44.2**
B&B (PGD2)*	4.9	2.9	3.9	**6.0**	4.2	4.3	**7.1**
SPSA	50.0	53.5	52.8	**56.2**	53.8	53.9	**56.6**
Auto-Attack	16.1	18.5	17.3	**18.8**	18.4	17.6	**20.8**

Table 3: Robustness evaluation on the CIFAR10 dataset with ResNet architecture. For the gradient based attacks, we use $\epsilon = 8/255, \eta = 2/255, k = 250$. (*) The low robust accuracies (even with standard method ADV) because the attack strength of PGD2 is double of the defense capacity, which makes the adversarial examples to be recognizable. CCE represents for CCE-RM version.

2020).

(iii) **Auto-Attack** (Croce and Hein 2020) (with the official implementation[5]) which is an ensemble based attack. We use $\epsilon = 8/255$ for the CIFAR10 dataset and $\epsilon = 0.01$ for the CIFAR100 dataset, both with standard version which is an ensemble of four different attacks.

(iv) **SPSA attack** (Uesato et al. 2018) (with *cleverhans* lib) which is a gradient-free optimization method. We use $\epsilon = 8/255$ for the CIFAR10 dataset and $\epsilon = 0.01$ for the CIFAR100 dataset, both with 50 steps**.**

The distortion metric we use in our experiments is l_∞ for all measures. We use the full test set for the attacks (i) and 1000 test samples for the attacks (ii-iv).

Robustness Evaluation

We conduct extensive experiments on the CIFAR10 and CIFAR100 datasets to compare our method with the other methods. We consider the ensemble of both two and three committee members (denoted by a subscript number in each method). It can be observed from the experimental results in Table [3, 4, 5] that:

(i) There is a gap of 2%$\sim$3% when comparing ADV-EN_3 with ADV_1 showing that increasing model capacity (by increasing number of ensemble member) can improve the robustness of the model.

(ii) There is a gap of 3%$\sim$4% between ADP_3 and ADV_1, and especially, a gap of 7%$\sim$8% when comparing our CCE-RM_3 with ADV_1, which shows the potential of the ensemble learning to tackle with the adversarial attacks.

(iii) With the same model capacity, our CCE-RM is consistently the best with all attacks and in some attacks, ours surpasses other baselines in a large margin (4%$\sim$5%).

(iv) There is a gap of 3% between CCE-RM_3 and CCE-RM_2, which is larger than the gap of 1% between ADP_3 and ADP_2 or that of ADV-EN_3 and ADV-EN_2, showing that our method collaborates members better and gets more benefit from ensembling more committee members.

The effectiveness of adversarial training method depends on the diversity (or the hardness) of the adversarial examples (Madry et al. 2018). Fort et al. (2019) found that differently initializing members' parameters, even with the same training data, can end up with different local optimal in the solution space. Therefore, the potential of ensemble learning (in the remark ii) can be explained by the fact that the adversarial space of an ensemble model $\mathcal{B}_{\text{insecure}}\left(\mathbf{x}, \boldsymbol{y}, f^{en}, \epsilon\right)$ is more diverse than that of a single model $\mathcal{B}_{\text{insecure}}\left(\mathbf{x}, \boldsymbol{y}, f, \epsilon\right)$.

Our advantages over others (in the remark iii, iv) can be explained by the fact that our proposed method encourages the diversity of its committee members. Specifically it can be elaborated on with the following three key points. Firstly, while other ensemble-based defenses use the adversarial examples of the entire ensemble $\mathbf{x}_a^{en} \sim \mathcal{B}_{\text{insecure}}\left(\mathbf{x}, \boldsymbol{y}, f^{en}, \epsilon\right)$, our method makes use of the broader joint adversarial space $\mathbf{x}_a^i \sim \mathcal{B}_{\text{insecure}}\left(\mathbf{x}, \boldsymbol{y}, f^i, \epsilon\right)$ (Lemma 1 (i)). Secondly, each member has different loss landscape (Fort, Hu, and Lakshminarayanan 2019), in addition with the randomness of an adversary (e.g., random starting points in PGD), each member has its individual adversarial set (partly collapsed as shown in the next experiment). Therefore, similar with the bagging technique, by promoting each member with its adversarial examples independently, we can increase the diversity of the joint adversarial space. Last but not least, inspired from traditional ensemble learning (Liu and Yao 1999), by elegantly collaborating PO and DO, we encourage the negative correlation among ensemble members, therefore, further improve the diversity of the joint adversarial space.

[5]https://github.com/fra31/auto-attack

Attack	ADV_1	ADV_2	ADP_2	CCE_2	ADV_3	ADP_3	CCE_3
Nat. acc.	75.7	76.0	75.9	**76.0**	**76.7**	76.6	75.7
PGD	38.0	39.7	42.2	**44.7**	40.8	43.9	**46.8**
BIM	38.2	39.7	42.2	**44.9**	40.8	43.8	**46.8**
MIM	38.5	40.5	42.4	**45.4**	41.3	44.2	**47.2**
mul-PGD	26.0	27.7	27.8	**31.9**	28.3	32.4	**36.9**
mul-BIM	25.9	27.2	27.2	**31.6**	27.7	29.8	**34.1**
mul-MIM	26.2	28.1	28.3	**32.3**	29.0	30.7	**34.6**
SPSA	40.6	44.3	41.5	**45.2**	45.1	46.1	**47.5**
Auto-Attack	25.1	25.0	24.4	**29.9**	25.5	28.1	**31.9**

Table 4: Robustness evaluation on the CIFAR10 dataset with standard CNN architecture. We use $\epsilon = 8/255, \eta = 2/255, k = 100$ for gradient based attacks. Note that $mul\mathcal{A}$ represents for multiple-targeted attack by adversary $\mathcal{A}$ with $k = 20$. CCE represents for CCE-RM version.

Attack	ADV_1	ADV_2	ADP_2	CCE_2	ADV_3	ADP_3	CCE_3
Nat. acc.	40.8	41.4	48.0	**53.4**	40.8	52.6	**54.4**
PGD	26.8	29.7	30.9	**35.3**	32.8	36.2	**39.5**
BIM	26.9	29.1	31.0	**35.2**	32.8	36.2	**39.4**
MIM	27.0	29.0	30.8	**35.3**	32.9	36.1	**39.6**
mul-PGD	16.4	15.8	20.1	**24.2**	16.6	24.8	**28.4**
mul-BIM	15.9	15.5	19.4	**23.7**	16.3	24.5	**28.1**
mul-MIM	16.7	16.1	20.3	**24.1**	16.8	25.1	**28.6**
SPSA	25.6	25.5	24.1	**31.8**	26.0	32.5	**35.0**
Auto-Attack	15.3	15.1	14.8	**21.9**	15.8	23.0	**25.9**

Table 5: Robustness evaluation on the CIFAR100 dataset with standard CNN architecture. We use $\epsilon = 0.01, \eta = 0.001, k = 100$ for gradient based attacks. Note that $mul\mathcal{A}$ represents for multiple-targeted attack by adversary $\mathcal{A}$ with $k = 20$. CCE represents for CCE-RM version.

Transferability Among Ensemble Members

The transferability is a phenomenon when adversarial examples generated to attack a specific model also mislead other models trained for the same task. In the ensemble learning context, adversarial examples which are transferred well among members will likely fool the entire ensemble. Therefore, reducing the transferability among members is a principled approach to achieve better robustness as claimed in the previous works (Pang et al. 2019; Kariyappa and Qureshi 2019). In this sub-section, we provide a further understanding of the transferability to the overall robustness and show the impact of the transferring flow.

We first summarize the experiments setting. The experiments are conducted on the CIFAR10 dataset with an ensemble of two members under PGD attack with $k = 20, \epsilon = 8/255, \eta = 2/255$. The results are reported in Table 6. CCE-Base is our model which disables the crossing PO and DO by setting $\lambda_{pm} = \lambda_{dm} = 0$. $a^{(i,j)}$ represents for the robust accuracy when adversarial examples $\{x_a^i\}$ attack model f^j. $|S|$ shows the cardinality of a subset S, i.e., the percentage of the images that go into the subset S, which can be one of $\{S_{11}, S_{01}, S_{10}, S_{00}\}$. From the definition of the transferability as mentioned above, to measure the transferability of adversarial examples $\{\mathbf{x}_a^i\}$, we can compute the accuracy difference of model f^i and $f^j, j \neq i$ against the same attack $\{\mathbf{x}_a^i\}$. The smaller gap implies that adversarial examples $\{\mathbf{x}_a^i\}$ are more transferable. The overall transferability of an ensemble method can be evaluated by the sum the accuracy differences over all its members, i.e., $T = a^{(1,2)} - a^{(1,1)} + a^{(2,1)} - a^{(2,2)}$.

We would like to emphasize some following important empirical observations (Table 6):

1) **The impact of the transferring flow.** It can be observed that the cardinality $|S_{11}|$ in CCE-RM (39.9%) is larger than that in CCE-Base (36.1%), while the cardinality $|S_{01}|, |S_{10}|, |S_{00}|$ is smaller than those in CCE-Base which serves as evidence that the adversarial examples are successfully transferred from subsets S_{10}, S_{01}, S_{00} to subset S_{11} as we expect. This helps improve the overall robustness of the ensemble model from 43.3% for CCE-Base to 45.5% for CCE-RM.

2) **The transferable space is just a subset of the adversarial space.** By definition, the subset S_{00} consists of adversarial examples which fools both models f^1, f^2, therefore, S_{00} represents for the transferable space of the ensemble model f^{en}. In fact, the cardinality of $|S_{00}|$ is smaller than the insecure region of the ensemble model f^{en} (i.e., the total classification error $100\% - a^{(en,en)}$) in all methods showing that the transferable space cannot represent for the insecure region of the ensemble model f^{en}, and the former is just the subset of the latter.

3) **Reducing transferability among ensemble members is not enough to improve adversarial robustness.** In fact, the transferability metric T for CCE-RM is 33.7% which is much smaller than those for ADP and ADV-EN (59.3% and 65.5%, respectively). The smaller value of T shows that the adversarial examples $\{\mathbf{x}_a^1\}, \{\mathbf{x}_a^2\}$ in our method are more transferable than those in ADV-EN and ADP. However, the fact that the overall robustness of our method is significantly better evidently shows that *transferability is not the only factor for improving the robustness.* This is because the robustness of each individual member under a direct attack (i.e., $a^{(1,1)}$ or $a^{(2,2)}$) is much lower than our method. In addition, the cardinality $|S_{11}|$ in our method is 39.9% which is much bigger than those in ADV-EN (24.0%) and ADP (25.7%).

We provide two additional metrics which are (i) $nT = 100\% - a^{(en,en)} - |S_{00}|$ to measure the cardinality of *adversarial examples set which successful attack model f^{en} but non transferable among f^1, f^2* and (ii) $a_{single} = a^{(en,en)} - |S_{11}|$ to measure the cardinality of *adversarial examples set which are correctly predicted by only one model either f^1 or f^2 but still being correctly predicted by model f^{en}*. The comparison on the metric nT in Table 6 shows that most of successful adversarial examples in our method are predicted incorrectly by both members. While the comparison on the metric a_{single} shows that most of unsuccessful adversarial examples in our method are predicted correctly by both members. The two comparisons demonstrate that our method have better robustness than other methods because (i) the adversarial examples have to fool both ensemble members for a successful attack and (ii) our ensemble model can predict correctly by both members which explains the higher performance.

The remarks (2, 3) further imply that:

An ensemble model cannot be secure against white-box attacks unless its members are robust against direct attacks (even they are secure against transferred attacks).

This hypothesis provides more understanding of the correlation between the transferability and the overall robustness of an ensemble model.

Improving Natural Accuracy and Adversarial Detectability

The parameter $\lambda_{pm}(\lambda_{dm})$ controls the level of the agreement (disagreement) of models $\{f^i\}, i \in [1, N]$ and model $f^j, j \neq i$ on the same adversarial example x_a^j. By disabling the crossing PO ($\lambda_{pm} = 0$) and strengthening DO (i.e., $\lambda_{dm} = 5$), our method encourages the disagreement among members on the same data example, therefore, increases the negative correlation among them. This setting of CCE-DM leads to two important properties, which are empirically proved by the experiments below.

Improving Natural Accuracy. We compare natural accuracies of two variants: CCE-RM and CCE-DM against the baselines. Table 7 shows that CCE-DM significantly improves natural accuracy of the ensemble model by a large margin. In traditional ensemble learning, the key ingredient to improve natural performance is making ensemble members more diverse (Kuncheva and Whitaker 2003). By disabling the crossing PO and strengthening DO, CCE-DM variant enforces the diversity more strictly, which explains the improvement of the natural performance. This result demonstrates the promising usage of adversarial examples to improve the traditional ensemble learning.

Model	$a^{(en,en)}$	$a^{(1,1)}$	$a^{(2,2)}$	$\|S_{11}\|$	$\|S_{01}\|$	$\|S_{10}\|$	$\|S_{00}\|$	T	nT	a_{single}
ADV-EN	40.7	31.1	33.2	24.0	17.0	13.0	46.0	65.5	13.3	16.7
ADP	42.9	31.0	33.1	25.7	13.1	11.7	49.5	59.3	7.6	17.2
CCE-RM	45.5	41.7	41.4	39.9	5.2	5.5	49.5	33.7	5.0	5.6
CCE-Base	43.3	40.3	40.5	36.1	6.5	7.2	50.3	36.1	6.4	7.2

Table 6: Evaluation on the transferability among ensemble members on the CIFAR10 dataset. $\{T, nT, a_{single}\}$ are the metrics of interest.

Model	ADV-EN	ADP	CCE-RM	CCE-DM
CNN_2	76.0	75.9	76.0	86.0
CNN_3	76.7	76.6	75.7	87.2
$ResNet_2$	85.3	85.3	84.5	91.0
$ResNet_3$	86.1	86.2	84.9	91.6

Table 7: Comparison of the natural performance on the CIFAR10 dataset (the subscript number denotes the number members).

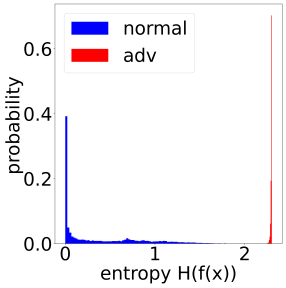
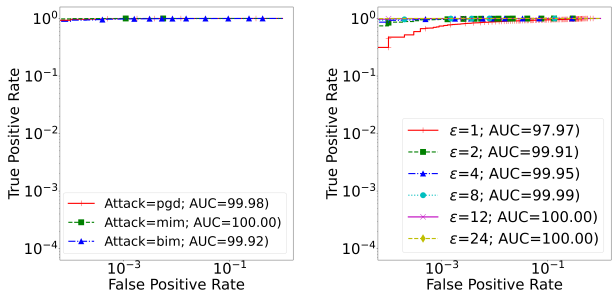


Figure 1: Histogram of prediction entropy in CCE-RM

Adversarial Detectability. CCE-DM can distinguish between benign and adversarial examples more easily. It is because the committee members produce a uniform prediction for adversarial examples, while yielding a very high confident prediction for benign examples. The histogram for all images in the test set and their adversarial examples in Figure 1 demonstrate the consistency of this observation over the data distribution.

These results further inspire us to develop a simple yet effective method to detect adversarial examples based on the entropy of the model prediction. Following the evaluation in (Pang et al. 2018, 2019), we try with different thresholds to distinguish the benign and adversarial examples and report the AUC score of each adversarial attack. It is worth noting that, we do not intend to compete with other adversarial detectors but just to show the advantage and flexibility of our CCE. The experiment is on the CIFAR10 dataset with an ensemble of two members. We conduct two evaluations to justify our understanding. First, we study our detection method against three different attacks: PGD, BIM and MIM with the same hyper-parameter setting $k = 20, \epsilon = 8/255, \eta = 1/255$. The result in Figure 2a shows that our method can accurately and consistently detect all three kind of attacks. Secondly, we study our detection method on different attack strengths. We use



(a) multiple types of attack (b) multiple attack strengths

Figure 2: ROC of CCE-RM under different attack scenarios

the PGD attack $k = 20, \eta = 1/255$ and vary the distortion bound ϵ from $1/255$ to $24/255$. The result in Figure 2b shows that our method can perform well on a wide range of attack strengths. The adversary is obviously less distinguishable when decreasing its strength. However, our method still obtains a very high AUC score $(93.4/100)$ even under a very weak attack $(\epsilon = 1/255)$, in which adversarial images look nearly identical to the original ones.

Conclusion

In this paper, we explore the use of ensemble-based learning to improve adversarial robustness. In particular, we propose a cross-collaborative strategy by means of enforcing the transferring flow of adversarial examples, thereby implicitly increasing the diversity of adversarial space and improving the robustness of the ensemble. Moreover, our proposed method can be performed in both detection and robustness modes. We conduct extensive and comprehensive experiments to show the improvement of our proposed method on state-of-the-art baselines. We also provide the detailed understanding of the relationship between the transferability and the overall robustness in the ensemble learning context.

Acknowledgements

This work was partially supported by the Australian Defence Science and Technology (DST) Group under the Next Generation Technology Fund (NTGF) scheme.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *International Conference on Machine Learning*, 274–283.

Bagnall, A.; Bunescu, R.; and Stewart, G. 2017. Training ensembles to detect adversarial examples. *arXiv preprint arXiv:1712.04006* .

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Srndic, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402. Springer.

Brendel, W.; Rauber, J.; Kümmerer, M.; Ustyuzhaninov, I.; and Bethge, M. 2019. Accurate, reliable and fast robustness evaluation. In *Advances in Neural Information Processing Systems*, 12861–12871.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690* .

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Fort, S.; Hu, H.; and Lakshminarayanan, B. 2019. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757* .

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL http://arxiv.org/abs/1412.6572.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, W.; Wei, J.; Chen, X.; Carlini, N.; and Song, D. 2017. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th {USENIX} Workshop on Offensive Technologies ({WOOT} 17)*.

Kariyappa, S.; and Qureshi, M. K. 2019. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981* .

Kuncheva, L. I.; and Whitaker, C. J. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning* 51(2): 181–207.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. URL https://openreview.net/forum?id=HJGU3Rodl.

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C.-J. 2018. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 369–385.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* .

Liu, Y.; and Yao, X. 1999. Ensemble learning via negative correlation. *Neural networks* 12(10): 1399–1404.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Pang, T.; Du, C.; Dong, Y.; and Zhu, J. 2018. Towards robust detection of adversarial examples. In *Advances in Neural Information Processing Systems*, 4579–4589.

Pang, T.; Xu, K.; Du, C.; Chen, N.; and Zhu, J. 2019. Improving Adversarial Robustness via Promoting Ensemble Diversity. In *International Conference on Machine Learning*, 4970–4979.

Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* .

Papernot, N.; McDaniel, P. D.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The Limitations of Deep Learning in Adversarial Settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, 372–387. IEEE. doi:10.1109/EuroSP.2016.36. URL https://doi.org/10.1109/EuroSP.2016.36.

Qin, C.; Martens, J.; Gowal, S.; Krishnan, D.; Dvijotham, K.; Fawzi, A.; De, S.; Stanforth, R.; and Kohli, P. 2019. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, 13824–13833.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! In *Advances in Neural Information Processing Systems*, 3353–3364.

Strauss, T.; Hanselmann, M.; Junginger, A.; and Ulmer, H. 2017. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1709.03423* .

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. URL http://arxiv.org/abs/1312.6199.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On Adaptive Attacks to Adversarial Example Defenses. *Advances in Neural Information Processing Systems* 33.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. D. 2018. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018*.

Tramèr, F.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453* .

Tsipras, D.; Santurkar, S.; Engstrom, L.; Ilyas, A.; and Madry, A. 2020. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks. *arXiv preprint arXiv:2005.11295* .

Uesato, J.; O Donoghue, B.; Kohli, P.; and Oord, A. 2018. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. In *International Conference on Machine Learning*, 5025–5034.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. *arXiv preprint arXiv:1901.08573* .