Improving Adversarial Robustness via Probabilistically Compact Loss with Logit Constraints

Xin Li*, Xiangrui Li*, Deng Pan*, Dongxiao Zhu†

Department of Computer Science, Wayne State University, Detroit, MI 48202
{xiangruili, xinlee, pan.deng, dzhu}@wayne.edu

Abstract

Convolutional neural networks (CNNs) have achieved state-of-the-art performance on various tasks in computer vision. However, recent studies demonstrate that these models are vulnerable to carefully crafted adversarial samples and suffer from a significant performance drop when predicting them. Many methods have been proposed to improve adversarial robustness (e.g., adversarial training and new loss functions to learn adversarially robust feature representations). Here we offer a unique insight into the predictive behavior of CNNs that they tend to misclassify adversarial samples into the most probable false classes. This inspires us to propose a new Probabilistically Compact (PC) loss with logit constraints which can be used as a drop-in replacement for cross-entropy (CE) loss to improve CNN's adversarial robustness. Specifically, PC loss enlarges the probability gaps between true class and false classes meanwhile the logit constraints prevent the gaps from being melted by a small perturbation. We extensively compare our method with the state-of-the-art using large scale datasets under both white-box and black-box attacks to demonstrate its effectiveness. The source codes are available from the following url: https://github.com/xinli0928/PC-LC.

Introduction

Convolutional neural networks (CNNs) have achieved significant progress for various challenging tasks in computer vision, including image classification (Li et al. 2020), semantic segmentation (He et al. 2017), and image generation (Goodfellow et al. 2014). Despite their success, CNNs are highly vulnerable to adversarial samples (Szegedy et al. 2013). With imperceptibly small perturbation added to a clean image, adversarial samples can drastically change models' prediction, resulting in a significant drop in CNN's predictive performance. This phenomenon poses a serious threat to security-critical applications of deep learning, such as autonomous driving (Al-Qizwini et al. 2017), surveillance system (Sreenu and Durai 2019), and medical imaging system (Li and Zhu 2020). Furthermore, studies have shown that adversarial robustness is also a key property to obtain human interpretation in computer vision and other application fields (Noack et al. 2019; Pan et al. 2020). Therefore,

---

*These authors contributed equally.

†Corresponding author.

improving models' adversarial robustness is critical to build trustworthy Artificial Intelligence systems to prevent unforeseen hazardous situations.

To improve CNN's adversarial robustness, many methods have been proposed. One strategy is to modify the inputs during inference time via noise removal (Meng and Chen 2017), super-resolution (Mustafa et al. 2019b) and JPEG compression (Das et al. 2017) to diminish the impact of perturbation, but can be easily evaded by strong attacks (Athalye, Carlini, and Wagner 2018). Another type of strategy (Tramèr et al. 2017; Kurakin, Goodfellow, and Bengio 2016; Sinha, Namkoong, and Duchi 2017; Zhang et al. 2019; Shafahi et al. 2019) is based on adversarial training (Goodfellow, Shlens, and Szegedy 2014) that can effectively increase the model's robustness by utilizing crafted adversarial examples as data augmentation. However, it is computationally expensive and compromise model classification performance on clean images (Tsipras et al. 2018). Other than modifying data, some techniques directly enhance model robustness by altering network architectures (Taghanaki et al. 2019; Mustafa et al. 2019a), or constructing ensembles of networks (Tramèr et al. 2017; Pang et al. 2019b). However, they require additional processes and are not flexible to be adopted to other models.

While previous works have successfully improved CNN robustness against adversarial attacks, the connection of CNN predictions between adversarial and clean samples is not known. In this paper, we investigate this connection in the typical setting when CNNs are trained with the cross-entropy (CE) loss. When an adversarial sample successfully fools the trained CNN with small perturbations, it tends to be misclassified into the first several most probable false classes when predicting the original clean sample, i.e., the classes with larger predicted probabilities. This consistent pattern of CNN's predictive behaviors is intuitive and potentially implies a deeper connection between the CNN feature learning and its adversarial robustness.

The tendency of misclassifying adversarial samples enlightens us that the adversarial robustness of CNN can be benefited from the training that focuses on the differentiation between sample's true class and its first several most probable false classes. Hence, in this paper, we propose a novel training objective, termed as Probabilistically Compact (PC) loss with logit constraints, which can improve adversarial
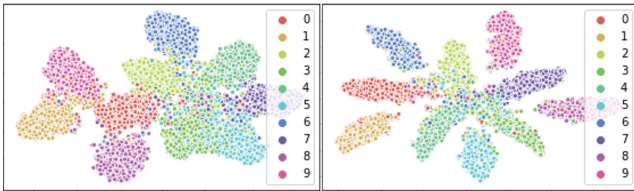
Figure 1: T-SNE visualization of the penultimate layer of ResNet-56 trained with CE loss (left) and PC loss (right) on CIFAR-10.

robustness and achieve comparable classification accuracy without extra training procedure and computational burden.

Unlike CE loss which focuses only on maximizing the output probability of the true class, PC loss aims at maximizing probability gaps between the true class and the most probable false classes. Meanwhile, the logit constraints suppress logit which ensures that the gaps is not only large, but also difficult to be crossed. Consequently, this formulation helps to widen the gaps between different classes in feature space, and ensures that it is difficult for an adversary to fool the trained model with small perturbations. We demonstrate the synergistic effect of the gaps at both probability and feature levels using benchmark datasets. For example, the average probability gaps between the true class and the most probable false class of ResNet-56 on CIFAR-100 test data are 0.527 (CE loss) vs. 0.558 (PC loss), respectively. And for Tiny ImageNet test data the gaps become 0.131 (CE loss) vs. 0.231 (PC loss). These results demonstrate that our PC loss can directly enlarge the probability gap of prediction and the effect is more pronounced for more challenging dataset (Tiny ImageNet). As shown in Figure 1, ResNet-56 trained with our PC loss has clear margin boundaries and samples of each classes are evenly distributed around the center with a minimal overlap on CIFAR-10 test data.

Our main contributions are summarized as follows: (1) We offer an unique insight into the predictive behavior of CNN on adversarial samples that the former tends to misclassify the latter into the first several most probable classes. (2) We formulate the problem by proposing a new loss function, i.e., PC loss with logit constraints to improve CNN's adversarial robustness, where these two components are systematically integrated and simultaneously optimized during training process. (3) Our PC loss can be used as a drop-in replacement of the CE loss to supervise CNN training without extra procedure nor additional computational burden for improving adversarial robustness. Experimental results show that when trained with our method, CNNs can achieve significantly improved robustness against adversarial samples without compromising performance on predicting clean samples.

Related Work

For generating adversarial samples, various computational methods have been developed. Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) and its variant Basic Iterative Method (BIM) efficiently generates adversarial samples by perturbing the pixels according to the gradient of the loss function (Kurakin, Goodfellow, and Bengio 2016). Projected Gradient Descent (PGD) (Madry et al. 2017) introduces a random starting point at each iteration in FGSM within a specified l_∞ norm-ball to enhance the attack effect. Momentum iterative method (MIM) (Dong et al. 2018) uses momentum to help iterative gradient-based methods to avoid sticking into local maximum thus further boosts their attacking performance. As an optimization-based attack, the Carlini and Wagner (C&W) (Carlini and Wagner 2017) approach uses binary search mechanism to find the minimal perturbation for a successful attack. SPSA (Uesato et al. 2018) is a gradient-free method which approximates gradient to generate attacks and defeats many defenses. It outperforms gradient-based attacks when the loss surface is hard to optimize.

To counter adversarial attack and enhance model robustness, various defensive techniques have been proposed. Among these proposed defenses, one line of those approaches (Kurakin, Goodfellow, and Bengio 2016; Sinha, Namkoong, and Duchi 2017; Zhang et al. 2019; Shafahi et al. 2019) are based on adversarial training (Goodfellow, Shlens, and Szegedy 2014) and can achieve effective robustness against different adversarial attacks, where the training dataset is augmented with adversarial samples. However, these methods have trade-off between accuracy on clean images and adversarial robustness (Tsipras et al. 2018) and are computationally expensive in adversarial samples generation (Zhang et al. 2019). To reduce the computational burden, Shafahi et al. (Shafahi et al. 2019) propose a training algorithm, which improves the efficiency of adversarial training by updating both model parameters and image perturbation in one backward pass.

Another line of defending strategy against adversaries, other than augmenting the training dataset, is to learn feature representations with adversarial robustness by using model ensembles or altering network architectures (Taghanaki et al. 2019; Mustafa et al. 2019a; Tramèr et al. 2017; Pang et al. 2019b). For example, (Taghanaki et al. 2019) augment CNNs with the radial basis function kernel to further transform features via kernel trick to improve the class separability in feature space and reduce the effect of perturbation. (Mustafa et al. 2019a) propose a prototype objective function, together with multi-level deep supervision. Their method ensures the feature space separation between classes and shows significant improvement of robustness. (Pang et al. 2019b) obtain a strong ensemble defense by introducing a new regularizer to encourage diversity among models within the ensemble system, which encourages the feature representation from the same class to be close. Although these approaches avoid high computationally cost of adversarial training, they have to modify the network or require extra training process, which limits their flexibility to be adapted to different tasks.

More efficient approaches are designing new loss functions to improve model adversarial robustness. By explicitly imposing regularization on latent features, CNNs are encouraged to learn feature representations with more inter-class separability and intra-class compactness (Pang et al.
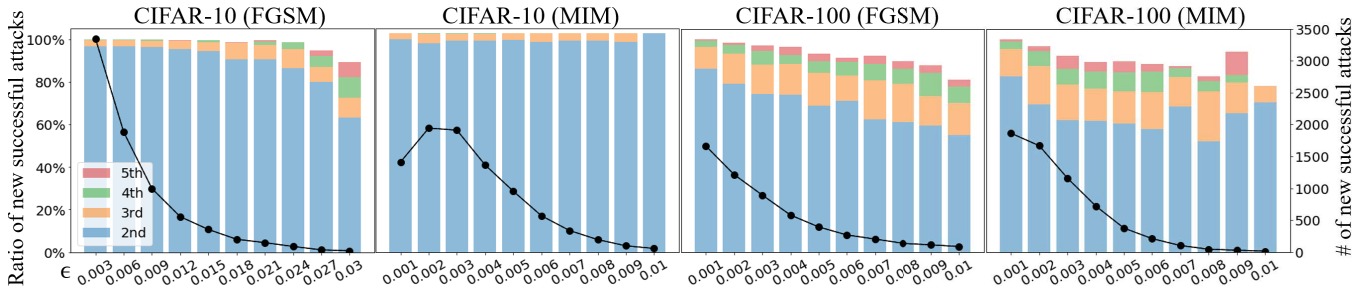
Figure 2: Empirical investigation on the predictive behavior of CNN on adversarial samples from CIFAR-10 and CIFAR-100. The line (black, right y-axis) represents the number of increased successful attacks when ϵ is increased from its previous grid value. Each bar (left y-axis) represents the percentage of misclassification for the increased successful attacks, measuring number of adversarial samples are misclassified into the 2nd, 3rd, 4th and 5th most probable classes. FGSM and MIM are attack methods.

2019a; Mustafa et al. 2019a). For example, (Pang et al. 2019a) propose a Max-Mahalanobis center (MMC) loss to learn discriminative features. They first calculate Max-Mahalanobis (Pang, Du, and Zhu 2018) centers for each class and then encourage the features to gather around the centers using Center Loss (Wen et al. 2016). However, the assumption of geometrical compactness for latent features (in terms of Euclidean distance or L_2-norm) may not hold due to inherent intra-class variations in the data and usually requires suitable assumptions on distribution of the latent features. Differently, our PC loss with logit constraints is motivated by the predictive behavior of CNN on adversarial samples from the probability perspective, which avoids this issue by learning probabilistically compact features without geometric assumptions. The work that is closest to ours is (Chen et al. 2019b) that encourages the predicted probabilities of false classes to be equally distributed, whereas our PC loss directly enlarges the gap of probabilities between true class and the first several most probable false classes.

Proposed Method

Notation Let $D = (\boldsymbol{x}_i, y_i)_{i=1}^N$ be the set of training samples of size N, where $\boldsymbol{x}_i \in \mathbf{R}^p$ is the p-dimensional feature vector and $y_i = k(k = 1, \cdots, K)$ is the true class label, and $S_k = \{(\boldsymbol{x}_i, y_i) : y_i = k\}$ the subset of D for the k-th class. The bold $\boldsymbol{y}_i = (y_i^1, \cdots, y_i^K)$ is used to represent the one-hot encoding for y_i: $y_i^k = 1$ if $y_i = k$, 0 otherwise.

Cross-entropy (CE) loss Assume that CNN's output layer, after convolutional layers, is a fully connected layer of K neurons with bias terms, then the predicted probability for sample $\boldsymbol{x}$ being classified into k-th class is calculated using the softmax activation (the k-th logit $a_k = \boldsymbol{W}_k \boldsymbol{h}_{\boldsymbol{x}} + b_k$):

$$f_k(\boldsymbol{x}) = p(y = k | \boldsymbol{x}) = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)} \quad (k = 1, \cdots, K), \tag{1}$$

where $\boldsymbol{h}_{\boldsymbol{x}}$ is the feature representation of $\boldsymbol{x}$, $\boldsymbol{W}_k$ and b_k are parameters of the k-th neuron in the output layer. Then CE loss, which is equivalent to the maximum likelihood ap-

proach, is given as follows:

$$L(\boldsymbol{\theta}) = -\sum_{k=1}^K \sum_{i_k \in S_k} \log f_k(\boldsymbol{\theta}; \boldsymbol{x}_{i_k}), \tag{2}$$

where $\boldsymbol{\theta}$ is the vector of trainable model parameters.

Motivation: Predictive Behavior of CNN on Adversarial Samples

Previous studies have shown that when trained to optimum, *i.e.* $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, CNNs can misclassify adversarial samples that are only slightly different from the original clean samples. This vulnerability has recently inspired many methods for generating adversarial samples (attack), defending adversarial attacks and detecting adversarial samples. Here, we take a different perspective on the attacks and empirically investigate if there is a systematic tendency on how CNNs misclassify adversarial samples.

Specifically, for a testing (clean) sample $(\boldsymbol{x}, y)$, the (untargeted) attack seeks a small perturbation ϵ that leads to the misclassification of $\boldsymbol{x}$ when the perturbation is added to $\boldsymbol{x}$:

$$\min_{\boldsymbol{\epsilon}} ||\boldsymbol{\epsilon}||_p, \text{s.t. } y' = \arg\max_k f_k(\boldsymbol{x} + \boldsymbol{\epsilon}) \text{ and } y \neq y', \quad (3)$$

where $|| \cdot ||_p$ is the norm such as L_1, L_2 and L_∞. When Eq. (3) is optimized and the attack succeeds, a natural question to ask is *"are there any connections between y' and y for the trained CNN misclassifying $\boldsymbol{x} + \boldsymbol{\epsilon}$ into class y'?"* Note that $y = \arg\max_k f_k(\boldsymbol{x})$.

Intuitively, we could expect that y' is likely to be the most probable class except the true class label y, *i.e*, the class corresponding to the 2nd largest value of CNN predicted probabilities. This conjectures that solving Eq. (3) is equivalent to solve

$$\min_{\boldsymbol{\epsilon}} ||\boldsymbol{\epsilon}||_p, \text{s.t. } \arg\max_k f_k(\boldsymbol{x} + \boldsymbol{\epsilon}) = \arg_{\#2} \max_k f_k(\boldsymbol{x}), \quad (4)$$

where $\arg_{\#2} \max$ represents the operation of taking the 2nd largest value[1].

---

[1] We may relax it to the first several most probable classes such as 3rd and 4th.

Here we provide an analysis as our motivation behind this conjecture. Assuming that the CNN is Lipschitz continuous (Fazlyab et al. 2019), then we have the inequality:

$$||\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{f}(\boldsymbol{x})||_p \leq l||(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{x}||_p = l||\boldsymbol{\epsilon}||_p, \quad (5)$$

where l is the Lipschitz constant and $\boldsymbol{f}(\cdot) = (f_1(\cdot), \cdots, f_K(\cdot))$. The Lipschitz continuity implies that the change of the ouput is bounded by the change of the input, which is the small perturbation in adversarial attacks. To misclassify $\boldsymbol{x}+\boldsymbol{\epsilon}$, the possible minimal value of the LHS in Eq. (5) is to seek an $\boldsymbol{\epsilon}$ such that $f_j(\boldsymbol{x}+\boldsymbol{\epsilon}) \geq f_y(\boldsymbol{x}+\boldsymbol{\epsilon})$, where y is true class and j is 2nd most probable class. To see this, consider the following two cases:

- Case 1. $f_y(\boldsymbol{x}+\boldsymbol{\epsilon}) \geq f_y(\boldsymbol{x})$. To misclassify $\boldsymbol{x}+\boldsymbol{\epsilon}$, the possible minimal value $||\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{f}(\boldsymbol{x})||_p$ is to reduce $f_{k'}(\boldsymbol{x})$ $(k' \neq y, j)$ to compensate $f_j(\boldsymbol{x}+\boldsymbol{\epsilon})$ so that $f_j(\boldsymbol{x}+\boldsymbol{\epsilon}) \geq f_y(\boldsymbol{x}+\boldsymbol{\epsilon})$.

- Case 2. $f_y(\boldsymbol{x}+\boldsymbol{\epsilon}) < f_y(\boldsymbol{x})$. The possible minimal value is that $f_y(\boldsymbol{x}+\boldsymbol{\epsilon}) = f_y(\boldsymbol{x}) - (\frac{f_y(\boldsymbol{x})-f_j(\boldsymbol{x})}{2})$, $f_j(\boldsymbol{x}+\boldsymbol{\epsilon}) = f_j(\boldsymbol{x}) + (\frac{f_y(\boldsymbol{x})-f_j(\boldsymbol{x})}{2})$ and all other $f_{k'}(\boldsymbol{x})$ $(k' \neq y, j)$ remain unchanged.

Those two cases may not be achievable in practice, but provide a lower bound on $||\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{f}(\boldsymbol{x})||_p$. The same analysis can be further relaxed to the 3rd and 4th most probable classes. Observing Eq. (5), solving Eq. (3) provides an upper-bound for the LHS of Eq. (5). With Lipschitz continuity, we hence conjecture that CNN tends to misclassify adversarial samples into classes that have large predicted probabilities when predicting the original clean samples.

To verify our conjecture, we perform an empirical study on CIFAR-10 and CIFAR-100 datasets. FGSM and MIM are used as the adversarial attack algorithms and generate adversarial samples for the standard testing data of CIFAR-10 and CIFAR-100. We do not solve Eq. (3) for each testing samples as it is computationally expensive for the test data of size 10,000. Instead, we take a fine grid of perturbation values and summarize the misclassification of newly successful attacks when the perturbation ϵ is increased from ϵ_m to $\epsilon_{m+1} = \epsilon_m + \Delta$ (Δ is the value of increment). Figure 2 displays the summary of the misclassification results. From the figure, we can see that for CIFAR-10, every time the perturbation is increased, the newly successful attacks are mostly misclassified into the 2nd most probable class for the clean samples. For CIFAR-100, the misclassification follows a similar trend considering it has 100 classes. We also notice that as the perturbation gets larger in FGSM, more newly successful attacks are classified into the 3rd, 4th and 5th most probable classes of predicting clean samples. A possible reason is that the large perturbation results in overshoot in the misclassification as the difference between 2nd most probable class and 3rd most probable class is small when a clean sample needs large perturbation to be adversarial. Different from FGSM, as perturbation increases, MIM always maintains a high percentage of classifying newly successful adversarial samples into the 2nd most probable class of predicting clean samples for CIFAR-10, due to its iterative procedure in generating adversarial attacks. Overall, Figure 2

empirically agrees with our analysis that motivates our proposed PC loss.

Probabilistically Compact Loss

The predictive behavior of CNN on adversarial samples in the last section inspires us that to improve model robustness to adversaries, CNN needs to focus on the differentiation between the true class and the first several most probable classes. In terms of predicted probability, CNN robustness is benefited from the large gap between true class $f_y(\boldsymbol{x})$ and false class $f_{y'}(\boldsymbol{x})$ $(y' \neq y)$. Indeed, (Neyshabur et al. 2017) shows that the gap $f_y(\boldsymbol{x}) - \max_{y'} f_{y'}(\boldsymbol{x})$ can be used to measure the generalizability of deep neural networks.

With the aforementioned motivation, we propose the PC loss to improve CNN's adversarial robustness as follows:

$$L_{pc}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{y' \neq y_i, i \in D} \max(0, f_{y'}(\boldsymbol{x}_i) + \xi - f_{y_i}(\boldsymbol{x}_i)), \quad (6)$$

where N is the number of training samples, $\xi > 0$ is the probability margin treated as hyperparameter. Here, we include all non-target classes in the formulation and penalize any classes for each training sample that violate the margin requirement for two considerations: (1) if one of the most probable classes satisfies the margin requirement, all less probable classes will automatically satisfy this requirement and hence have no effect in PC loss; (2) since the first several most probable classes are unknown and can change during the training process, it is necessary to maintain the margin requirement for all classes.

Compared with previous works (Mustafa et al. 2019a; Pang et al. 2019a; Taghanaki et al. 2019) that improve adversarial robustness via explicitly learning features with large intra-class compactness, PC loss avoids assumptions on the feature space. Instead, PC loss only encourages the feature learning that leads to probabilistic intra-class compactness by imposing a probability margin ξ.

In training CNN with PC loss, the latter is differentiable and hence can be optimized with stochastic gradient descent. The gradient of PC loss can be calculated as (w.r.t. logit a_y)

$$\frac{\partial(f_{y'} + \xi - f_y)}{\partial a_y} = -f_y(1 - f_y + f_{y'}), \quad (7)$$

where the gradient is computed for the softmax function. For other logits ($a_{y'}$), the gradients can be similarly computed.

The Logit Constraints

In last section we introduce PC loss to enhance adversarial robustness by enlarging the probability gaps. In this section, we further propose logit constraints as a complement of PC loss, which suppress logit to ensure that the gaps are not only large, but also difficult to be crossed. Here we explain the necessity of using both parts together. We use $|\cdot|$ to denote $||\cdot||_2$ for simplicity. The conclusion can be extended to other L_p norms. Assume there is a clean image $\boldsymbol{x}$, and a corrupted image $\boldsymbol{x}+\boldsymbol{\epsilon}$ attacked by some adversarial algorithm, where $|\boldsymbol{\epsilon}| < \tau$, with $\tau > 0$ be a small constant.

We take the log probability for simplicity, i.e., PC loss is equivalently enlarging $\log f_y(\boldsymbol{x}) - \log f_j(\boldsymbol{x})$ between true

class y and most probable false class j. Given a perturbation ϵ, the corresponding log probabilities can be estimated via first order approximation

$$\log f_y(\boldsymbol{x} + \boldsymbol{\epsilon}) = \log f_y(\boldsymbol{x}) + \boldsymbol{\epsilon} \cdot \nabla_{\boldsymbol{x}} \log f_y(\boldsymbol{x}), \quad (8)$$

$\log f_j(\boldsymbol{x} + \boldsymbol{\epsilon})$ can also be approximated in the same manner. To prevent $\log f_y(\boldsymbol{x} + \boldsymbol{\epsilon}) - \log f_j(\boldsymbol{x} + \boldsymbol{\epsilon}) < 0$ (i.e., false prediction with perturbation $\boldsymbol{\epsilon}$), we should solve $\min_{\boldsymbol{\theta}} \boldsymbol{\epsilon} \cdot (\nabla_{\boldsymbol{x}} \log f_j(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log f_y(\boldsymbol{x}))$.

Lets denote vector $\boldsymbol{b} = \nabla_{\boldsymbol{x}} \log f_j(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log f_y(\boldsymbol{x})$. As the attackers can always choose the worst $\hat{\boldsymbol{\epsilon}}$ that maximizes $\boldsymbol{\epsilon} \cdot \boldsymbol{b}$ by letting $\hat{\boldsymbol{\epsilon}}$ in the same direction as $\boldsymbol{b}$, i.e. $\boldsymbol{\epsilon} \cdot \boldsymbol{b} \leq |\hat{\boldsymbol{\epsilon}}| \cdot |\boldsymbol{b}|$. Our goal becomes to minimize the upper bound $|\hat{\boldsymbol{\epsilon}}| \cdot |\boldsymbol{b}|$

$$\min_{\boldsymbol{\theta}} |\hat{\boldsymbol{\epsilon}}||\boldsymbol{b}| \implies \min_{\boldsymbol{\theta}} |\boldsymbol{b}|$$
$$\implies \min_{\boldsymbol{\theta}} |\nabla_{\boldsymbol{x}} \log f_j(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log f_y(\boldsymbol{x})|. \quad (9)$$

Hence to prevent the prediction changes after perturbation ϵ, we should minimize $|\nabla_{\boldsymbol{x}}(\log f_y(\boldsymbol{x}) - \log f_j(\boldsymbol{x}))|$. Let a_k denote the logit for the kth class softmax output, observe that

$$\nabla_{\boldsymbol{x}} \log f_y - \nabla_{\boldsymbol{x}} \log f_j = \frac{\nabla_{\boldsymbol{x}} f_y}{f_y} - \frac{\nabla_{\boldsymbol{x}} f_j}{f_j} \quad (10)$$
$$= \nabla_{\boldsymbol{x}} a_y - \nabla_{\boldsymbol{x}} a_j,$$

because $\nabla_{\boldsymbol{x}} f_y = -\sum_k f_k f_y \nabla_{\boldsymbol{x}} a_k + f_y \nabla_{\boldsymbol{x}} a_y$, and the same holds for $\nabla_{\boldsymbol{x}} f_j$. We can equivalently change our objective to $\min_{\boldsymbol{\theta}} |\nabla_{\boldsymbol{x}}(a_y - a_j)|$. We can estimate $|\nabla_{\boldsymbol{x}}(a_y(\boldsymbol{x}) - a_j(\boldsymbol{x}))|$ using

$$|\nabla_{\boldsymbol{x}}(a_y - a_j)| \approx |(a_y(\boldsymbol{x}) - a_j(\boldsymbol{x}))$$
$$- (a_y(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}))|/|\hat{\boldsymbol{\epsilon}}|, \quad (11)$$

where we denote $|\boldsymbol{\epsilon}| < |\hat{\boldsymbol{\epsilon}}| = \tau$ that upper bounds $|\boldsymbol{\epsilon}|$. Note that an adversarial attack tends to minimize $a_y(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}})$ so that $a_y(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}) < a_y(\boldsymbol{x}) - a_j(\boldsymbol{x})$. And a robust model should instead prevent $a_y(\boldsymbol{x} + \boldsymbol{\epsilon}) - a_j(\boldsymbol{x} + \boldsymbol{\epsilon}) < 0$ to ensure a correct prediction when attacked, so we have the following inequality for a robust model under attack

$$0 < a_y(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\boldsymbol{x} + \hat{\boldsymbol{\epsilon}}) < a_y(\boldsymbol{x}) - a_j(\boldsymbol{x}). \quad (12)$$

Then Eq. (11) can be upper bounded by

$$|\nabla_{\boldsymbol{x}}(a_y - a_j)| < |a_y(\boldsymbol{x}) - a_j(\boldsymbol{x})|/|\hat{\boldsymbol{\epsilon}}|. \quad (13)$$

Substitute this inequality back to Eq. (10), we get a logit constraint condition to ensure model robustness

$$|\nabla_{\boldsymbol{x}}(\log f_y - \log f_j)| < |a_y(\boldsymbol{x}) - a_j(\boldsymbol{x})|/|\hat{\boldsymbol{\epsilon}}| < C, \quad (14)$$

where C is an arbitrary positive constant thresholding robustness, hence we can optimize PC loss subject to the above condition

$$\min_{\boldsymbol{\theta}} L_{pc}(\boldsymbol{\theta}), \text{s.t.} |a_y(\boldsymbol{x}; \boldsymbol{\theta}) - a_j(\boldsymbol{x}; \boldsymbol{\theta})| < C' \text{ for } \forall \boldsymbol{x}, \quad (15)$$

where $C' = |\hat{\boldsymbol{\epsilon}}|C$. It is equivalent to write the above minimization problem with a multiplier λ

$$\min_{\boldsymbol{\theta}, \boldsymbol{\lambda}} \left(L_{pc}(\boldsymbol{\theta}) + \frac{\lambda}{N} \sum_{\boldsymbol{x} \in D} (d_{yj} - C') \right), \quad (16)$$

where N is the number of samples, and D is the set of training samples. λ is treated as a hyper-parameter in training, $d_{yj} = \max(0, a_y(\boldsymbol{x}; \boldsymbol{\theta}) - a_j(\boldsymbol{x}; \boldsymbol{\theta}))$. As shown in Eq. (16), PC loss and logit constraints are systematically integrated and simultaneously optimized during training process to enhance model adversarial robustness.

Experiments

In this section, we evaluate our proposed PC loss with logit constraints along with analysis that our method does not rely on the 'gradient masking' that provides a false sense of security (Athalye, Carlini, and Wagner 2018).

Datasets and models: We analyze seven benchmark datasets: MNIST, KMNIST, Fashion-MNIST (FMNIST), CIFAR-10, CIFAR-100, Street-View House Numbers (SVHN), and Tiny Imagenet. We scale all pixel values to $[0, 1]$ following the preprocessing procedure in (Mustafa et al. 2019a; Pang et al. 2019b). For gray-scale image datasets (K/F/MNIST), we use a LeNet-5 model (LeCun et al. 1998), and for color image datasets (CIFAR-10, CIFAR-100, SVHN, Tiny Imagenet), we use a VGG-13 model (Simonyan and Zisserman 2014). All these models are trained using Adam optimizer with a initial learning rate of 0.01 and a batch size of 256. For our method, we first warm up the training process for T epochs ($T = 50$ for K/F/MNIST and $T = 150$ for other datasets) using CE loss, and then train the model using our method shown in Eqs. (6) and (16) ($\xi = 0.995, \lambda = 0.05$) for another T epochs whereas we directly train the baseline using CE loss for $2T$ epochs.

Attack types In the adversarial setting, there are two main threat models: white-box attacks where the adversary possesses complete knowledge of target model, including its architecture, training method and learned parameters, and black-box attacks where the adversary does not have access to the information about trained classifier but is aware of the classification task. We evaluate the robustness of our proposed method against both white-box and black-box attacks.

Results

Performance on white-box attacks Following the attack settings in (Chen et al. 2019a), we crafted adversarial examples in a non-targeted way with respect to allowed perturbation ϵ for gradient-based attacks, i.e., FGSM, BIM, PGD and MIM. The number of iterations is set to 10 for BIM and 40 for MIM and PGD while perturbation of each step is 0.01. For parameters of optimization-based attack C&W, the maximum iteration steps are set to 100, with a learning rate of 0.001, and the confidence is set to 0.

The results (Table 1) demonstrate that our proposed PC loss with logit constraints outperforms the CE loss under white-box attacks while maintaining the comparable level of performance on the clean image classification. The improvement is even more significant on stronger attacks.

Besides comparing to the standard CE loss, we also compare our defense approach with a closely related Guided Complement Entropy (GCE) approach (Chen et al. 2019a). To ensure a fair comparison we use the exactly same models (LeNet-5 for MNIST and ResNet56 for CIFAR-10) and parameters (max iterations of C&W is 1000) as in the GCE paper. In Table 2, it is evident that our method outperforms GCE in the vast majority of settings.

Performance on black-box attacks The performance under black-box setting is critical to substantiate adversarial robustness since it is closer to the real-world scenario where

Attacks	Param.	MNIST		KMNIST		FMNIST		Param.	CIFAR-10		SVHN	
		CE	Ours	CE	Ours	CE	Ours		CE	Ours	CE	Ours
Clean	-	99.2	99.2	95.5	95.4	90.1	90.2	-	91.6	91.2	94.9	94.7
FGSM	0.1	71.5	**80.5**	26.2	**62.8**	17.5	**58.0**	0.04	4.3	**53.1**	8.7	**39.5**
	0.2	51.6	**76.3**	1.8	**39.7**	9.5	**43.3**	0.12	11.7	**30.3**	4.5	**24.2**
	0.3	31.8	**65.0**	1.2	**34.1**	7.6	**31.8**	0.2	11.5	**18.7**	2.6	**17.1**
BIM	0.1	52.8	**72.0**	55.8	**82.5**	0.0	**18.7**	0.04	0.0	**29.0**	1.3	**26.2**
	0.2	4.5	**48.6**	28.7	**73.3**	0.3	**8.4**	0.12	0.0	**21.0**	0.0	**18.2**
	0.3	1.5	**39.5**	16.8	**60.5**	0.0	**6.4**	0.2	0.0	**20.0**	0.0	**17.6**
PGD	0.1	49.0	**72.3**	31.3	**62.4**	0.0	**15.7**	0.04	0.0	**27.6**	0.0	**27.6**
	0.2	3.3	**50.2**	3.9	**39.9**	0.0	**7.0**	0.12	0.0	**14.6**	0.0	**22.0**
	0.3	0.8	**39.7**	2.0	**33.2**	0.0	**4.6**	0.2	0.0	**7.5**	0.0	**21.0**
MIM	0.1	49.8	**73.8**	26.0	**65.4**	0.0	**14.8**	0.04	0.0	**34.3**	0.0	**29.2**
	0.2	5.0	**54.0**	4.2	**45.4**	0.0	**6.3**	0.12	0.0	**32.7**	0.0	**27.6**
	0.3	1.5	**43.3**	2.0	**36.9**	0.0	**4.5**	0.2	0.0	**32.4**	0.0	**26.0**
CW	0.0	42.2	**78.0**	19.5	**57.3**	0.2	**21.8**	0.0	0.0	**30.2**	0.0	**36.2**

Table 1: Accuracy (%) on K/F/MNIST, CIFAR-10 and SVHN under white-box setting. For CW, the parameter is the confidence.

Attacks	Param.	MNIST			Param.	CIFAR-10		
		CE	GCE*	Ours		CE	GCE*	Ours
FGSM	0.1	71.5	**87.7**	80.5	0.04	12.7	41.2	**58.4**
	0.2	51.6	62.7	**76.3**	0.12	10.3	14.8	**17.3**
	0.3	31.8	47.2	**65.0**	0.2	7.0	11.8	**12.0**
BIM	0.1	52.8	61.9	**72.0**	0.04	0.0	**19.6**	16.6
	0.2	4.5	34.5	**48.6**	0.12	0.0	3.0	**3.4**
	0.3	1.5	33.5	**39.5**	0.2	0.0	2.0	**2.6**
PGD	0.1	49.0	51.9	**72.3**	0.04	0.0	5.9	**10.2**
	0.2	3.3	9.6	**50.2**	0.12	0.0	1.9	**3.5**
	0.3	0.8	2.2	**39.7**	0.2	0.0	1.6	**2.7**
MIM	0.1	49.8	61.2	**73.8**	0.04	0.0	15.4	**16.0**
	0.2	5.0	39.8	**54.0**	0.12	0.0	**13.1**	11.6
	0.3	1.5	38.8	**43.3**	0.2	0.0	**12.7**	11.2
C&W	0.0	0.0	25.6	**30.1**	0.0	0.0	0.8	**3.3**

Table 2: Accuracy (%) between GCE and our method on MNIST and CIFAR10 under white-box setting. *Results are directly from (Chen et al. 2019a).

an adversary has no access to the trained classifier. During inference time, black-box adversary uses a substitute model trained on the same dataset to generate adversarial samples to attack the target model. In our cases, we use a 3-layer CNN as the substitute model for LeNet-5 and ResNet-56 for VGG-13 to generate black-box attacks. Similar to (Pang et al. 2019b), we adopt PGD and MIM, the two most commonly used attack methods under the black-box setting. We then further evaluate our defense method using a gradient-free attack approach, i.e., SPSA, as in (Carlini et al. 2019), which performs numerical approximation on the gradients using test data. The learning rate of SPSA is set to 0.01, and the step size is $\delta = 0.01$ (Uesato et al. 2018). As shown in Table 3, the model trained with our PC loss improves robustness against the black-box attacks.

Larger-scale experiments on CIFAR-100 and Tiny ImageNet We also evaluate our method on larger and more complex CIFAR-100 and Tiny ImageNet datasets under both white-box attacks (PGD, MIM) and black-box attack (SPSA). Similar to (Pang et al. 2019b), we reduce the perturbation budget to the range of [0.005, 0.015] and attack iterations to 10 due to the increased data complexity and scale.

As shown in Table 4, our method significantly improves the model's adversarial robustness compared to the CE loss and GCE while maintaining the comparable level of performance on the clean image classification. Recall our observation that the most probable false classes are more vulnerable to attacks. GCE flattens the probabilities on false classes and thus enlarges the gap between true class and the most probable false class to increase model's robustness. However, when dataset become complex with more classes, this gap is smaller due to generally lower output probability for the true class, resulting a limited robustness improvement. On the other hand, our method directly maximizes the probability gap and thus is more suitable for large scale datasets.

Combining with adversarial training To demonstrate our method's compatibility and synergy with other adversarial defense techniques, we investigate the performance of our method in combination with adversarial training. Our goal is not to beat adversarial training, instead we attempt to show our method can be combined with it to further improve adversarial robustness. During training, we augment the dataset with adversarial samples generated using FGSM with perturbation range of $[0.1, 0.3]$ for gray-scale image datasets (K/F/MNIST), and 5-step PGD with perturbation range of $[0.0, 0.1]$ for color image dataset (CIFAR-10). The ratio of adversarial examples and clean images in each training mini-batch remains $1 : 1$. For gray-scale image datasets, table 5 shows that integrating our method with adversarial training further improves the model's adversarial robustness under both white-box (PGD, BIM, MIM, CW) and black-box attack settings (SPSA). Furthermore, our PC loss with adversarial training outperforms GCE with adversarial training, which demonstrates our method has better compatibility with other defense techniques. It is worth mentioning that our method alone outperforms the fast version adversarial training on gray-scale datasets, which generates adversarial training examples by one-step FGSM attack. And for color image dataset (CIFAR-10), we augment the dataset with adversarial examples crafted by more advanced PGD attacks. The result shows the same trend, and the performance gain is more pronounced on this more challenging dataset.

Attacks	Param.	MNIST		KMNIST		FMNIST		Param.	CIFAR-10		SVHN	
		CE	Ours	CE	Ours	CE	Ours		CE	Ours	CE	Ours
PGD	0.1	96.6	**98.1**	90.8	**92.5**	65.5	**74.0**	0.04	19.5	**42.2**	43.6	**48.8**
	0.2	84.3	**92.8**	76.7	**85.0**	50.4	**57.1**	0.12	13.2	**38.0**	19.5	**28.1**
	0.3	61.1	**85.6**	59.1	**77.7**	47.8	**53.5**	0.2	16.7	**35.6**	13.2	**24.5**
MIM	0.1	96.4	**98.1**	90.4	**92.2**	62.4	**72.3**	0.04	17.2	**42.0**	40.1	**44.2**
	0.2	84.2	**94.7**	74.9	**83.0**	43.3	**54.2**	0.12	1.3	**16.3**	13.3	**21.7**
	0.3	56.7	**81.2**	48.4	**63.8**	30.5	**36.0**	0.2	0.3	**11.2**	10.0	**16.2**
SPSA	0.3	72.9	**95.7**	50.2	**78.0**	4.3	**39.8**	0.3	0.0	**45.3**	4.0	**58.0**

Table 3: Accuracy (%) on K/F/MNIST, CIFAR-10 and SVHN under black-box setting.
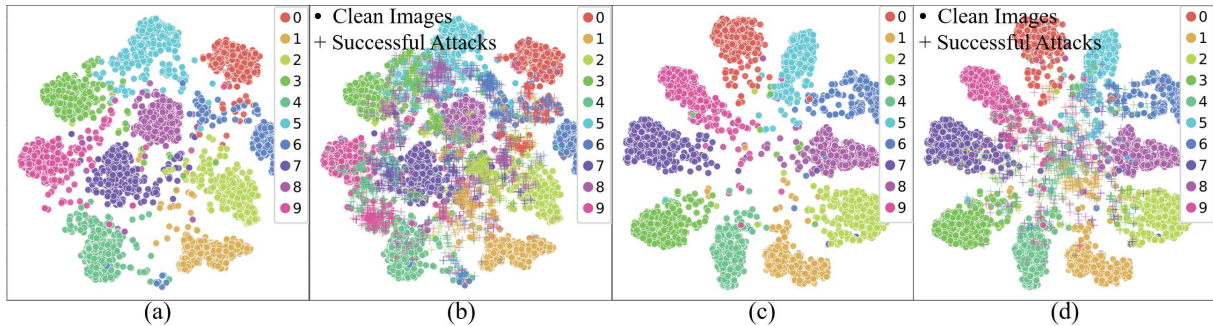

Figure 3: T-SNE visualization of the penultimate layer of the model trained by CE loss (a,b) and our PC loss (c,d) on MNIST dataset. (a,c) display only clean images whereas (b,d) also include successful attacks generated with FGSM ($\epsilon = 0.3$).

Attacks	Param.	CIFAR-100			Tiny ImageNet		
		CE	GCE	Ours	CE	GCE	Ours
Clean	-	40.2	64.5	**67.7**	**38.2**	32.8	37.7
PGD	0.005	11.4	24.4	**56.7**	11.3	8.9	**24.8**
	0.010	2.0	14.8	**54.6**	2.8	2.6	**19.2**
	0.015	0.4	9.1	**52.6**	0.8	1.0	**15.8**
MIM	0.005	8.7	21.9	**55.8**	7.9	7.3	**23.5**
	0.010	1.4	12.3	**52.8**	1.9	2.1	**18.0**
	0.015	0.3	7.4	**48.9**	0.6	1.4	**14.6**
SPSA	0.015	3.9	11.5	**22.1**	6.3	7.3	**16.2**

Table 4: Accuracy (%) on CIFAR-100 and Tiny ImageNet between CE loss, GCE and our new PC loss.

Feature Space Visualization In order to visually dissect the advantages of PC loss over the CE loss, we also inspect the feature space of trained models using t-SNE on MNIST datasets. As shown in Figure 3a, the model trained with CE loss has a large portion of clean images lay across the boundaries between different classes thus easily to be manipulated to become adversarial samples. On the contrary, for the model trained with our PC loss with logit constraints, in Figure 3c, the samples of each class have clear boundaries and are evenly distributed around the center with a minimal overlap. Note that the samples locate near the center are 'hard samples' for a classifier even without attacks.

Looking into the successful attacks (labeled with '+') in Figure 3, we find the predictive behavior of CNN on adversarial samples is consistent to our hypothesis. In Figure 3b, for a model (LeNet-5) trained with CE loss, adversarial sam-

ples are mostly located to the nearest classes corresponding to the most probable false classes. For example, many adversarial attacks generated based on class 5 are located within the class 8 of clean images and *vice versa*. While in Figure 3d, for a model trained with PC loss, due to the large margin between classes, the adversarial samples are harder to cross the boundaries with the only exception that the adversarial samples are distributed near the center of the feature space where hard samples are usually located.

Identifying Gradient Masking

Previous defense strategies (Buckman et al. 2018; Xie et al. 2017) rely on the effect of gradient masking, which was considered as a false sense of security (Athalye, Carlini, and Wagner 2018). Briefly, these defenses deteriorate the gradient information to make gradient-based attack methods hard to generate effective adversarial examples. However, these defenses can be easily defeated by black-box or gradient-free attackers. We show that our method does not rely on gradient masking on the basis of characteristics defined in (Athalye, Carlini, and Wagner 2018; Carlini et al. 2019). (1) Iterative attacks have better performance than one-step attack: Our results in Table 1 indicate that the iteration-based attacks (BIM, MIM, PGD) are more successful in generating adversarial attacks than single step method (FGSM). (2) Robustness against Black-box attacks is higher than white-box attacks: When model's gradients information is manipulated by the defender, the attacker can recover the gradient with black-box attacks and perform more successful attacks than using white-box attacks (Papernot et al. 2017). However, the results in Tables 1 & 3 demonstrate that our method

Attacks	Param.	MNIST				KMNIST				FMNIST				Param.	CIFAR-10	
		CE+AT	GCE+AT	Ours+AT	Ours	CE+AT	GCE+AT	Ours+AT	Ours	CE+AT	CGE+AT	Ours+AT	Ours		CE+AT	Ours+AT
BIM	0.3	27.0	28.3	**86.1**	39.5	58.4	8.1	**65.9**	60.5	0.1	4.8	**18.6**	4.9	0.04	17.3	**38.7**
PGD	0.3	3.2	26.8	**72.3**	39.7	37.2	0.5	**48.7**	33.2	0.0	2.7	**13.4**	2.8	0.04	11.4	**33.7**
MIM	0.3	10.8	27.7	**78.8**	43.3	22.3	15.7	**51.9**	36.9	0.0	1.4	**9.1**	2.2	0.04	9.0	**33.3**
C&W	0.0	75.5	69.4	**96.4**	78.0	47.9	48.2	**67.1**	57.3	4.0	21.5	**29.8**	21.8	0.0	0.0	**33.7**
SPSA	0.3	77.3	56.6	**97.1**	95.7	72.0	69.9	**79.0**	78.0	14.2	30.0	**41.6**	39.8	0.3	6.5	**39.6**

Table 5: Accuracy (%) on K/F/MNIST and CIFAR-10 with adversarial training under both white- and black-box attacks.

is more effective against black-box attacks and thus does not obfuscate gradients. (3) Increasing perturbation budget will increase attack success: As shown in the Table 1, increase of perturbations monotonically enhances the attacks. With a large budget ($\epsilon = 0.3$), the success rate is close to 100%.

Conclusion

We propose a novel PC loss with logit constraints inspired by the predictive behavior of CNN on adversarial samples. A CNN trained with our PC loss can achieve impressive robustness against adversarial samples without compromising performance on clean images nor requires additional procedures/computing, making it scalable to large-scale datasets. In addition, our PC loss is flexible and compatible with other defense methods, e.g., as a drop-in replacement of CE loss to supervise adversarial training. In future work, we plan to extensively investigate the connection of predictions between adversarial and clean samples in more general settings.

Ethical Impact

During the past few years, CNNs have been successfully implemented in various computer vision tasks such as autonomous driving, surveillance, and medical imaging diagnosis. Some of these systems have been already deployed and are integrated in our daily lives. Nevertheless recent studies have revealed the vulnerabilities of CNNs against adversarial attacks. Without adversarially robust defense strategies, the users of these systems can be exposed to excessive hazardous situations such as car accidents, burglaries, and inaccurate diagnosis. Therefore improving the robustness of CNNs is critical to build trustworthy AI systems to benefit humanity and society.

In this paper, we propose a novel loss function that can significantly improve CNN robustness against adversarial attacks and thus help to mitigate the risks at deployment of AI systems. The high efficiency of our approach makes it not only adaptive to most classification tasks with various model architectures in different fields without any extra procedure, but also compatible with other defense techniques to further improve the CNN's robustness. As such, our method has a strong potential to become a central pillar of the next-generation trustworthy AI system.

Acknowledgements

References

Al-Qizwini, M.; Barjasteh, I.; Al-Qassab, H.; and Radha, H. 2017. Deep learning algorithm for autonomous driving using GoogLeNet. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, 89–96. IEEE.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420* .

Buckman, J.; Roy, A.; Raffel, C.; and Goodfellow, I. 2018. Thermometer Encoding: One Hot Way To Resist Adversarial Examples. In *The International Conference on Learning Representations (ICLR)*.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705* .

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.

Chen, H.-Y.; Liang, J.-H.; Chang, S.-C.; Pan, J.-Y.; Chen, Y.-T.; Wei, W.; and Juan, D.-C. 2019a. Improving adversarial robustness via guided complement entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, 4881–4889.

Chen, H.-Y.; Wang, P.-H.; Liu, C.-H.; Chang, S.-C.; Pan, J.-Y.; Chen, Y.-T.; Wei, W.; and Juan, D.-C. 2019b. Complement objective training. *arXiv preprint arXiv:1903.01182* .

Das, N.; Shanbhogue, M.; Chen, S.-T.; Hohman, F.; Chen, L.; Kounavis, M. E.; and Chau, D. H. 2017. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900* .

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.

Fazlyab, M.; Robey, A.; Hassani, H.; Morari, M.; and Pappas, G. 2019. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 11423–11434.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* .

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Li, X.; Li, X.; Pan, D.; and Zhu, D. 2020. On the Learning Property of Logistic and Softmax Losses for Deep Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4739–4746.

Li, X.; and Zhu, D. 2020. Robust Detection of Adversarial Attacks on Medical Images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1154–1158. IEEE.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* .

Meng, D.; and Chen, H. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 135–147.

Mustafa, A.; Khan, S.; Hayat, M.; Goecke, R.; Shen, J.; and Shao, L. 2019a. Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3385–3394.

Mustafa, A.; Khan, S. H.; Hayat, M.; Shen, J.; and Shao, L. 2019b. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing* 29: 1711–1724.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 5947–5956.

Noack, A.; Ahern, I.; Dou, D.; and Li, B. 2019. Does Interpretability of Neural Networks Imply Adversarial Robustness? *CoRR* abs/1912.03430. URL http://arxiv.org/abs/1912.03430.

Pan, D.; Li, X.; Li, X.; and Zhu, D. 2020. Explainable Recommendation via Interpretable Feature Mapping and Evaluation of Explainability. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 2690–2696. ijcai.org. doi:10.24963/ijcai.2020/373. URL https://doi.org/10.24963/ijcai.2020/373.

Pang, T.; Du, C.; and Zhu, J. 2018. Max-mahalanobis linear discriminant analysis networks. *arXiv preprint arXiv:1802.09308* .

Pang, T.; Xu, K.; Dong, Y.; Du, C.; Chen, N.; and Zhu, J. 2019a. Rethinking Softmax Cross-Entropy Loss for Adversarial Robustness. *arXiv preprint arXiv:1905.10626* .

Pang, T.; Xu, K.; Du, C.; Chen, N.; and Zhu, J. 2019b. Improving adversarial robustness via promoting ensemble diversity. *arXiv preprint arXiv:1901.08846* .

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519.

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.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Sinha, A.; Namkoong, H.; and Duchi, J. 2017. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571* 2.

Sreenu, G.; and Durai, M. S. 2019. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data* 6(1): 48.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

Taghanaki, S. A.; Abhishek, K.; Azizi, S.; and Hamarneh, G. 2019. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11340–11349.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* .

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* .

Uesato, J.; O'Donoghue, B.; Oord, A. v. d.; and Kohli, P. 2018. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666* .

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, 499–515. Springer.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2017. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991* .

Zhang, D.; Zhang, T.; Lu, Y.; Zhu, Z.; and Dong, B. 2019. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, 227–238.