

Towards Noise-Robust Neural Networks via Progressive Adversarial Training

Hang Yu, Aishan Liu, Xianglong Liu*, Jichen Yang, Chongzhi Zhang

State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
{hyu0829,liuaishan,sxxayjc}@buaa.edu.cn, {xlliu,chongzhizhang}@nlsde.buaa.edu.cn

Abstract

Adversarial examples, intentionally designed inputs tending to mislead deep neural networks, have attracted great attention in the past few years. Although a series of defense strategies have been developed and achieved encouraging model robustness, most of them are still vulnerable to the more commonly witnessed corruptions, e.g., Gaussian noise, blur, etc., in the real world. In this paper, we theoretically and empirically discover the fact that there exists an inherent connection between adversarial robustness and corruption robustness. Based on the fundamental discovery, this paper further proposes a more powerful training method named Progressive Adversarial Training (PAT) that adds diversified adversarial noises progressively during training, and thus obtains robust model against both adversarial examples and corruptions through higher training data complexity. Meanwhile, we also theoretically find that PAT can promise better generalization ability. Experimental evaluation on MNIST, CIFAR-10 and SVHN show that PAT is able to enhance the robustness and generalization of the state-of-the-art network structures, performing comprehensively well compared to various augmentation methods. Moreover, we also propose Mixed Test to evaluate model generalization ability more fairly.

Introduction

State-of-the-art deep learning models have shown significant successes in many tasks, including computer vision (Krizhevsky, Sutskever, and Hinton 2012), natural language processing (Bahdanau, Cho, and Bengio 2014) and speech (Hinton et al. 2012). Their performances are usually promised by training using sufficient clean data. However, in real-world environment, it is usually impractical to acquire entirely clean data without any noises, such as adversarial examples and corruptions (Hendrycks and Dietterich 2019), which have been proven to pose potential threats to deep learning systems (Liu et al. 2019; Kurakin, Goodfellow, and Bengio 2016), especially those deployments in safety and security-critical environments. Therefore, it is crucial to well understand the noise robustness of deep neural networks.

In the past few years, great efforts have been devoted to exploring model robustness to the adversarial noises (or adversarial examples), maliciously constructed imperceptible perturbations that fool deep learning models, from the

views of *attack* (Goodfellow, Shlens, and Szegedy 2014a; Athalye, Carlini, and Wagner 2018) and *defense* (Xie et al. 2018; Kurakin, Goodfellow, and Bengio 2017; Madry et al. 2017). Most of the existing defense methods attempt to build adversarially robust models via supplying adversaries with non-computable gradients. Though they successfully make the deep models predict stably when encountering the adversarial examples, they were still easily defeated by circumventing obfuscated gradients (Athalye, Carlini, and Wagner 2018). Recently, *adversarial training*, appeared as a strong defense algorithm by adversarially data augmentation, has shown the strong capability of offering robust models against adversarial examples.

Besides the progress in robustness to adversarial examples, recent studies have also paid attention to improving model robustness to another common noises named corruptions (Zheng et al. 2016; Sun et al. 2018). In the real-world deep learning systems, corruptions like Gaussian blur and snow are more likely to witness than adversarial examples. Dodge and Karam found that deep learning models behave distinctly subhuman to input images with Gaussian noises. Likewise, deep learning models show weak performance on various corruptions including blur, pixelation and other types of noises (Hendrycks and Dietterich 2019).

Though the existence of noises especially in real-world environment has drawn intense concerns about the robustness of deep models, the literature mainly focused on either corruptions (Zheng et al. 2016; Sun et al. 2018) or adversarial noises (Kurakin, Goodfellow, and Bengio 2017; Madry et al. 2017). A very few studies have investigated the problem from the view of generalized noise robustness. For example, Fawzi, Moosavi-Dezfooli, and Frossard studied the robustness of classifiers from adversarial examples to random noises and tried to build a robust model from the view of curvature constraints. Fawzi, Fawzi, and Fawzi drew the relationship between in-distribution robustness and unconstrained robustness. More recently, Ford et al. found that adversarial robustness is closely related to robustness under certain kinds of distributional shifts, i.e., additive Gaussian noise. Despite the promising progress, there still remain open questions that need to be answered for deep understanding model robustness: Is there any relationship between the different noises such as adversarial noises and corruptions? Could we build strong models against them?

To answer these questions, this paper first dedicates theo-

*Corresponding author.

retical and empirical analyses to demonstrate that there exists an inherent correlation between adversarial robustness and corruption robustness, which potentially explains why existing adversarial defensive methods may somewhat improve model robustness against corruptions. Based on our theoretical finding, we further devise a powerful training method named *Progressive Adversarial Training* (PAT) to improve both adversarial and corruption robustness. Different from the conventional methods that achieve model robustness by employing more training data (Schmidt et al. 2018; Sun, Zhu, and Lin 2019), in PAT the diversified adversarial noises are aggregated, augmented and injected progressively, which are proven to be beneficial on improving both the adversarial and corruption robustness. We also theoretically prove that PAT can also promise models with strong generalization ability. Extensive experiments on MNIST, CIFAR-10 and SVHN indicate that PAT shows comprehensively excellent results on both adversarial and corruption noise compared to various augmentation methods. Moreover, in order to conduct unbiased evaluation, we also propose *Mixed Test* to evaluate model generalization ability more fairly.

Related Work

Adversarial robustness. In order to improve model robustness against adversarial examples, many adversarial training based methods have been proposed and proved to be comparatively competitive through experiments among others. Goodfellow, Shlens, and Szegedy firstly tried to improve model robustness against adversarial attacks through adversarial training which intuitively adds a regularization term and directly feeds adversarial examples during training. Madry et al. adversarially trained moderately the most robust model with the **PGD** augmented adversarial examples, whereas it consumes too much time and fails to generalize well on clean examples. Meanwhile, Sinha, Namkoong, and Duchi proposed an adversarial training algorithm with a surrogate loss which proved by theoretical analysis, but it is only confined to the small dataset, e.g., MNIST. Besides, adversarial noises are also calculated and added to each hidden layer during training in order to tackle the overfitting problem in (Sankaranarayanan et al. 2018). However, adversarial training still fails to handle more aggressive iterative attack and remains questionable to corruption robustness.

Corruption robustness. When it goes to the robust model towards corruptions, rare defense methods have been proposed. To improve model robustness against JPEG compression, Zheng et al. proposed stability training to stabilize model behavior against small input distortions. By investigating feature distribution shift within convolutional neural networks, Sun et al. employed feature quantization techniques to deal with distorted images including motion blur, salt and pepper. More recently, Metz et al. drew insights from meta-learning which uses a learned optimizer to build a robust model against input noises, e.g., translations.

Noise Robustness

In this section, we introduce the formal definition of noise robustness, taking the widely studied and used convolutional neural networks in image classification as the basic deep models.

Basic Deep Models

Given a training set $\mathcal{S} = \{s_i\}_1^n$ and a test set $\mathcal{Z} = \{z_i\}_1^m$ with a feature vector $x \in \mathcal{X}$ and a label $y \in \mathcal{Y}$, the deep supervised learning model tries to learn a mapping or classification function $f: \mathcal{X} \rightarrow \mathcal{Y}$ or $\mathbb{R}^D \rightarrow \{1, \dots, K\}$ as the mapping from input examples to output labels, where D is the dimension of input examples, $\{1, \dots, K\}$ is the corresponding labels of output. More precisely, $f_S(\theta; x, y)$ represents the prediction results of input (x, y) after the model is trained on training set \mathcal{S} . We use the log-loss in the image classification problem as follows:

$$\ell(\theta; x, y) = - \sum_{i=1}^m y_i \log \left(\frac{e^{f(\theta; x, y_i)}}{\sum_{j=1}^m e^{f(\theta; x, y_j)}} \right),$$

where $f(\theta; x, y_i)$ denotes the value of position y_i in prediction result. For the single-label classification problem, let $y_k = 1$, then the above loss can be expressed as below:

$$\ell(\theta; x, y) = -f(\theta; x, y_k) + \log \left(\sum_{j=1}^m e^{f(\theta; x, y_j)} \right).$$

Model Robustness

Now we introduce the definitions and formulations of the model robustness. Assuming that the test set \mathcal{Z} can be divided into K disjoint subsets, then the number of these partitions is denoted as the covering number of the set \mathcal{Z} .

ϵ -cover. Given a specified metric ρ , for subsets \mathcal{T} and $\hat{\mathcal{T}}$ of \mathcal{Z} , set $\hat{\mathcal{T}}$ is called an ϵ -cover of set \mathcal{T} if $\hat{\mathcal{T}}$ satisfies: for any $t \in \mathcal{T}$, there exists $\hat{t} \in \hat{\mathcal{T}}$ such that $\rho(t, \hat{t}) \leq \epsilon$.

ϵ -covering number. Refer to the concept in (Wellner and others 2013), the ϵ -covering number of set \mathcal{T} is defined as $N(\epsilon, \mathcal{T}, \rho)$, which indicates the smallest covering number among all of the ϵ -covers of set \mathcal{T} .

Let the partition \mathcal{C}_i and $\hat{\mathcal{C}}_i$ respectively denotes the set of data samples that objectively should be classified and the set of data samples that actually is classified in i -th class. Therefore, the points of i -th class in sample set can be regarded as the subset of their intersection, i.e., $\subset (\mathcal{C}_i \cap \hat{\mathcal{C}}_i)$. More precisely, we use \mathcal{C}_i^T and \mathcal{C}_i^E to represent $\mathcal{C}_i \cap \hat{\mathcal{C}}_i$ and $\mathcal{C}_i \setminus (\mathcal{C}_i \cap \hat{\mathcal{C}}_i)$, respectively. Thus, \mathcal{C}_i^E can be regarded as the set of samples that are supposed to be classified as i -th class but actually not, which is close to the error set of i -th class.

Adversarial robustness. The definition of adversarial robustness set \mathcal{R}_{adv} can be written as:

$$\mathcal{R}_{adv} = \{x \in \mathbb{R}^n | x \in (\mathcal{C}_i^T)^\epsilon \wedge x \notin \mathcal{C}_i^T \wedge x \in \mathcal{C}_i\}. \quad (1)$$

Corruption robustness. We choose Gaussian noise as the representation in the analysis of corruptions and we have $\mathbb{P}(\mathcal{A}) = \int_{\mathcal{A}} \varphi(x) dx$ as the integrals of Gaussian probability density function. Then, ϵ -extension of set \mathcal{A} can be defined

as $\mathcal{A}^\varepsilon = \{x \in \mathbb{R}^n | d(x, \mathcal{A}) < \varepsilon\}$, which denotes that the original set is “extended” by samples within a distance constraint. Then, the corruption robustness set \mathcal{R}_{cor} is shown as below:

$$\mathcal{R}_{cor} = \{x \in \mathbb{R}^n | x \in \mathcal{C}_i \wedge x \notin \hat{\mathcal{C}}_i\}. \quad (2)$$

Therefore, $\mathbb{P}(\mathcal{R}_{adv})$ denotes adversarial vulnerability and on the contrary $1 - \mathbb{P}(\mathcal{R}_{adv})$ represents adversarial robustness. Likewise, $\mathbb{P}(\mathcal{R}_{cor})$ and $1 - \mathbb{P}(\mathcal{R}_{cor})$ stand for corruption vulnerability and robustness. More details can be found in Theorem 1 and the supplementary material.

Theoretical Connections Between Adversarial and Corruption Robustness

From above, the ε -extension of sets are defined as $(\mathcal{C}_i^T)^\varepsilon$ and $(\mathcal{C}_i^E)^\varepsilon$, where \mathcal{C}_i^T denotes a set of points that *should* be and *actually* be classified as i -th class, and \mathcal{C}_i^E denotes a set of points that *should* be but *actually* not be classified as i -th class. In the following, we use $\mathcal{A}^\varepsilon \setminus \mathcal{A}$ to represent the surrounding area of ε band with respect to ε -extension of \mathcal{A} . More illustration can be found in the supplementary material. Then we have the following theorem holds:

Theorem 1 *For class i in single-label classification problem, we define learning function $f: \mathbb{R}^D \rightarrow \{1, \dots, K\}$. Then we have*

$$\mathbb{P}(\mathcal{C}_i^\varepsilon \setminus \mathcal{C}_i) - \mathbb{P}((\mathcal{C}_i^T)^\varepsilon \setminus \mathcal{C}_i^T) + 2\mathbb{P}(\mathcal{R}_{adv}) \geq \Phi(\Phi^{-1}(\mathbb{P}(\mathcal{R}_{cor})) + \varepsilon) - \mathbb{P}(\mathcal{R}_{cor}),$$

where \mathcal{C}_i is fixed and $\mathcal{C}_i^T \setminus \mathcal{C}_i$ is increased by augmentation. $\Phi(\cdot)$ represents the cumulative distribution function of normal distribution $N(0, 1)$ as shown below:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

For probability $\mathbb{P}(\mathcal{R}_{cor}) \in (0, \Phi(-\frac{\varepsilon}{2}))$, the right hand side of inequality $\Phi(\Phi^{-1}(\mathbb{P}(\mathcal{R}_{cor})) + \varepsilon) - \mathbb{P}(\mathcal{R}_{cor})$ is a monotonically increasing function of $\mathbb{P}(\mathcal{R}_{cor})$.

For the proof and explanation, please refer to the supplementary material. According to this theorem, we draw the observation that if we enhance the adversarial robustness by reducing $\mathbb{P}(\mathcal{R}_{adv})$, the intersection of $\hat{\mathcal{C}}_i$ and \mathcal{C}_i will increase with \mathcal{C}_i staying the same, which results in the rise of $(\mathcal{C}_i^T)^\varepsilon \setminus \mathcal{C}_i^T$, leading to the smaller left-side inequality, i.e. the tighter upper bound of right-side inequality. Thus, when $\mathbb{P}(\mathcal{R}_{cor})$ is in this proved interval, the stronger corruption robustness can be provable as well. Thus, we draw a key conclusion that there exists a positive correlation between adversarial and corruption robustness under some constraints.

Based on the theoretical finding, we can further explain why the existing solutions can hardly obtain the generalized noise robustness in practice, which usually focus on either adversarial examples or the corruptions. In the literature, in order to improve corruption robustness, it is an intuitive idea to add random Gaussian noise through the training process as Gaussian data augmentation (GDA) (Ford et al. 2019). Though proved to be effective against corruptions (Ford et al. 2019), it has little contribution to the actual adversarial

robustness, since improving corruption robustness directly increases the lower bound on the left side of the inequality due to the monotonically increasing function of the right side of the inequality which may somewhat reduce adversarial robustness on the contrary. Meanwhile, most adversarial defense methods, e.g., PGD-based adversarial training (PGD-AT) (Madry et al. 2017), believes to improve adversarial robustness by searching the worst-case adversarial noises with much time consumption which indeed promises good adversarial robustness. However, due to the limited amount of adversarial noises model witnessed, the reduction of \mathcal{R}_{adv} will be in a restricted range on the left side of the inequality, which in turn reflects moderate robustness to corruptions on the right side of the inequality.

Progressive Adversarial Training

Inspired by the theoretical connections between adversarial and corruption robustness, we propose our *progressive adversarial training* (PAT) strategy that adds adversarial noises with huge complexity during training to shrink \mathcal{R}_{adv} significantly, which leads to the larger increase of $(\mathcal{C}_i^T)^\varepsilon \setminus \mathcal{C}_i^T$ with the fixed \mathcal{C}_i resulting in the decrease of \mathcal{R}_{cor} . Thus, both adversarial and corruption robustness are guaranteed.

Formulation

Sinha, Namkoong, and Duchi firstly proposed surrogate loss, which adds the Lagrangian constraint on the original loss function in order to approximate the worst-case perturbation:

$$\min_{\theta \in \Theta} \sup_P \{\mathbb{E}_P[\ell(\theta; x, y)] - \lambda d(P, P_0)\},$$

where P refers to the robust objective distribution, P_0 denotes original data distribution of training set, and $d(P, P_0)$ represents the distance between distribution P and P_0 . The process indirectly constrains $d(P, P_0)$ by optimizing the surrogate loss, meanwhile the augmented distribution approximates the robust distribution P .

Analogously, let $\|\cdot\|$ denotes the l_2 metric, we consider the surrogate loss with constraint as:

$$\phi_\lambda(\theta; x, y) = \min_{\theta} \sup_{\delta} [\ell(\theta; x, y, \delta) - \frac{\lambda}{2} \|\delta\|^2], \quad (3)$$

where δ represents the worst-case perturbation which approximates the objective robust distribution on current distribution P_i . In this sense, we could update the parameter in the direction of robustness with the optimization of surrogate loss. Our Progressive Adversarial Training method is developed to minimize robust surrogate loss.

Prior studies (Schmidt et al. 2018; Sun, Zhu, and Lin 2019) have shown that the sample complexity plays a critical role in training a robust deep model. Schmidt et al. concluded that the sample complexity of robust learning can be significantly larger than that of standard learning under adversarial robustness situation. Charles et al. believed that adversarial training may need exponentially more iterations to obtain large margins compared to standard training. Moreover, based on a risk decomposition theorem, (Zhai et al. 2019) theoretically and empirically showed that with more

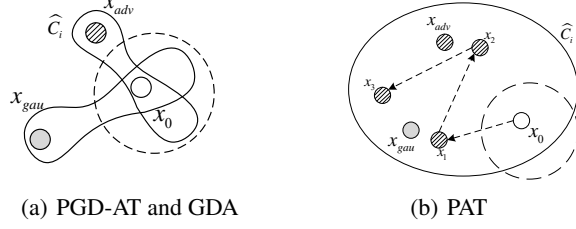


Figure 1: The effect of data augmentation on classification, where the dashed circles denote the partition of the empirical risk minimization. (a) The solid circles denote the partitions of PGD-AT and GDA. (b) The solid circle represents the partition of our proposed PAT.

unlabeled data, we can learn a model with better adversarially robust generalization. From the view of data complexity, we introduce Progressive Adversarial Training which injects adversarial noises progressively during several iterations within each training step, and thus we increase data complexity and diversity leading to robust models.

Specifically, given a training data $s = (x, y)$, we assume δ^* as the local optimal perturbation obtained through the gradient ascent. To get the local optimal perturbation δ^* , we calculate the perturbation gradient with respect to s as:

$$\nabla_{\delta^*} \phi_\lambda(\theta; s, \delta^*) = \nabla_{\delta^*} \ell(\theta; s, \delta^*) - \lambda \|\delta^*\|. \quad (4)$$

Since data complexity contributes significantly to adversarially robust model, progressive iteration process has been proposed to introduce diversified adversarial examples through k adversarial ascent steps. Thus, the perturbation δ^* can be updated in every step during the progressive process as:

$$\delta^* \leftarrow (1 - \beta)\delta^* + \frac{\varepsilon}{k} \cdot \frac{\nabla_{\delta^*} \ell(\theta; s, \delta^*)}{\|\nabla_{\delta^*} \ell(\theta; s, \delta^*)\|}, \quad (5)$$

where ε is the step size and it is normalized by the number of steps k , so that the overall magnitude of update equals ε . β here controls perturbation decay with the increase of progressive iterations as well as the contribution of the former perturbation δ^* .

During each perturbation computation step via gradient ascent, another batch of adversarial examples is obtained as $x \leftarrow x + \delta^*$, thus another one-step gradient descent is executed to promote the model parameters update. Obviously, $k\|\nabla_{\delta^*} \ell(\theta; s)\| = \|k\nabla_{\delta^*} \ell(\theta; s)\|$. With the surrogate loss optimizing, we have

$$\delta^* \leftarrow \delta^* + \frac{\varepsilon}{\|k\nabla_{\delta^*} \ell(\theta; s, \delta^*)\|} \cdot [\nabla_{\delta^*} \ell(\theta; s, \delta^*) - \lambda \delta^*], \quad (6)$$

where $\lambda = \beta\|k\nabla_{\delta^*} \ell(\theta; s)\|/\varepsilon$, according to equation (5) and (6). In this way, progressive adversarial training can be regarded as the optimization of robust surrogate loss in an iterative way.

In the case of PGD-AT, it attempts to update the model parameters with the approximated worst-case perturbations computed and added within current data for one step only.

Algorithm 1 Progressive Adversarial Training

Input: Training set $S = \{s_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$, Hyperparameters β, ε, k

Output: Updated parameters θ

```

1: for  $i$  in  $n$  training examples do
2:   for  $j$  in  $k$  iterative steps do
3:      $\delta^j \leftarrow (1 - \beta)\delta^j + \frac{\varepsilon}{k} \cdot \frac{\nabla_{\delta^j} \ell(\theta; x_i^j, y_i, \delta^j)}{\|\nabla_{\delta^j} \ell(\theta; x_i^j, y_i, \delta^j)\|}$ 
4:      $x_i^j \leftarrow x_i^j + \delta^j$ 
5:     Updates parameter  $\theta$  with augmented  $(x_i^j, y_i)$ 
6:   end for
7: end for
```

Meanwhile, GDA-based approach tries to improve model robustness by adding the average-case perturbations. Different from them, our method tends to augment current data and update model parameters for k steps, which significantly brings more data with higher complexity and diversity. Furthermore, these perturbations are much more intended and flexible when they are calculated in each step during training. Therefore, the models trained by PAT are expected to be robust against more types of noises, including both adversarial examples and corruptions (see the proof in Theorem 1).

Intuitively, as illustrated in Figure 1 (a), the data partition \hat{C}_i for the i -th class becomes moderately narrow after traditional data augmentation (GDA and PGD-AT), which indicates that the partition space is “stretched” to cover more adversarial examples or Gaussian noises in some specific directions (Shaham, Yamada, and Negahban 2018). However, it directly declines the model generalization ability to benign examples as partition \hat{C}_i fails to cover some portions of benign instances that are used to be classified correctly by empirical risk minimization (ERM). On the contrary, after PAT, as shown in Figure 1 (b), the capacity of data partition \hat{C}_i becomes larger when more data with higher diversities is witnessed by the model during training. Thus, PAT promises a model with excellent generalization ability on benign examples as well as polluted ones.

Generalization and Robustness

Now, we further explain and theoretically prove the robustness and the upper bound on the expected generalization error for our algorithm. For the loss function $\ell(\theta; z)$, obviously there exist upper bounds M_0 and M_1 , such that $|\ell(\theta; z)| \leq M_0$ and $\|\nabla_z \ell(\theta; z)\| \leq M_1$.

Assumption 1 For the test set \mathcal{Z} and l_2 metric $\|\cdot\|$, assuming that the loss function is smooth. Then there exist L_0 and L_1 , such that $|\ell(\theta; z) - \ell(\theta; z')| \leq L_0 \|z - z'\|$, $\|\nabla_z \ell(\theta; z) - \nabla_{z'} \ell(\theta; z')\| \leq L_1 \|z - z'\|$.

Lemma 1 ([Xu and Mannor 2012], Theorem 14). For the set \mathcal{Z} and l_2 metric $\|\cdot\|$, if $\ell(\theta; z)$ satisfies L_0 -Lipschitzian, then for augmented training set \mathcal{S} and $\forall \gamma > 0$, $f_S(\theta; \cdot)$ is $(N(\gamma/2, \mathcal{Z}, \|\cdot\|), \gamma L_0)$ -robust. That is to say, \mathcal{Z} can be partitioned into $K = N(\gamma/2, \mathcal{Z}, \|\cdot\|)$ disjoint sets $\{\mathcal{C}_i\}_{i=1}^K$, such that $\forall s \in \mathcal{S}$,

$$s, z \in \mathcal{C}_i \Rightarrow |\ell(\theta; s) - \ell(\theta; z)| \leq \gamma L_0.$$

Theorem 2 If loss function $\ell(\theta; z)$ satisfies L_0 -Lipschitzian and $\nabla_z \ell(\theta; z)$ satisfies L_1 -Lipschitzian, then for existing upper bound M_0 and M_1 , the surrogate loss $\phi_\lambda(\theta; z)$ satisfies L_λ -Lipschitzian. For augmented training set \mathcal{S} and $\forall \gamma > 0$, $f_{\mathcal{S}}(\theta; \cdot)$ is $(K, \gamma L_\lambda)$ -robust, where $L_\lambda = L_0 + \frac{2M_1 L_1 + 1}{\lambda}$, $K = N(\gamma/2, \mathcal{Z}, \|\cdot\|)$.

The robust property in Lemma 1 reveals that, for the number of partition K and a sample z , if z falls into the same partition as s , then the deviation between their losses could be controlled by γL_0 . Here we obtain the $(N(\gamma/2, \mathcal{Z}, \|\cdot\|), \gamma L_0)$ -robust by Lipschitzian property in Lemma 1. If γ increases, the covering number $N(\gamma/2, \mathcal{Z}, \|\cdot\|)$ will decrease and γL_0 will increase for a fixed Lipschitz constant, which promises the balance of the robust property. In Theorem 2 we derive the robust property of learning function for surrogate loss. The proof is shown in the supplementary material.

Theorem 3 Given augmented training set \mathcal{S} and $f_{\mathcal{S}}(\theta; \cdot)$ satisfying $(K, \gamma L_\lambda)$ -robust, then for any probability $p > 0$, there is at least probability $1 - p$ such that surrogate loss satisfies

$$\phi_\lambda(\theta; z) \leq \frac{1}{n} \sum_{i=1}^n \hat{\ell}(\theta; z_i) + \gamma L_\lambda + M_0 \left(\sqrt{\frac{2K \ln 2 - 2 \ln p}{n}} \right) + \frac{M_1^2}{\lambda - L_1},$$

where $L_\lambda = L_0 + \frac{2M_1 L_1 + 1}{\lambda}$, $K = N(\gamma/2, \mathcal{Z}, \|\cdot\|)$, $n = |\mathcal{S}|$ denotes the volume of \mathcal{S} .

In Theorem 3, we analyze the generalization error upper bound for the surrogate loss of our proposed PAT. The proof is in the supplementary material. The difference between losses could be constrained in γL_λ with similar input examples. There is a balance between K and γL_λ . If γ increases, then the former K decreases and γL_λ increases at the same time, which promises a reasonable generalization bound with the robust property. Thus, according to the theoretical analysis above, we can draw the conclusion that PAT builds robust model with proved upper bound on the generalization error.

Experiment and Evaluation

In this section, we will evaluate our Progressive Adversarial Training with other compared methods on the image classification task against both adversarial and corrupted examples.

We adopt the popular MNIST (LeCun 1998), CIFAR-10 (Krizhevsky and Hinton 2009) and SVHN (Netzer et al. 2011) as the evaluation datasets in our experiments. As for the deep models, we also choose the widely-used models, such as LeNet-5 (LeCun et al. 1998) with simple architecture for MNIST, VGG-16 (Simonyan and Zisserman 2014) for CIFAR-10 and standard ResNet-18 (He et al. 2016) for SVHN, respectively. For the compared methods, different data augmentation methods are employed including PGD-AT and GDA.

Evaluation Protocol

We first evaluate the model robustness with top-1 classification accuracy respectively on benign, adversarial and corrupted datasets, following the guidelines from (Carlini et al. 2019). Meanwhile, we also use our proposed criterion *Mixed*

Test to evaluate the generalization ability of model, where each test set is mixed with **benign, adversarial and corrupted examples** in the same proportion.

Adversarial examples. Following (Kurakin, Goodfellow, and Bengio 2017; Athalye, Carlini, and Wagner 2018; Carlini et al. 2019), we construct adversarial examples with the strongest attacks in different norms from both **gradient-based attack** and **optimization-based attack** including: Projected Gradient Descent (PGD) (Madry et al. 2017) attack and C&W attack (Carlini and Wagner 2017). For CIFAR-10 and SVHN, we set the l_{inf} -perturbation size ϵ of PGD as 2/255 and 4/255, and set the l_2 -perturbation size c of C&W as 0.2 and 0.5, which is similar to (Tsuzuku, Sato, and Sugiyama 2018).

Corruptions. Following (Hendrycks and Dietterich 2019), we constructed corrupted datasets of CIFAR-10 and SVHN consisting of various types corruption, e.g., noise, blur, weather, digital, etc. More precisely, we divide all 15 corruptions into 3 main types as *noise*, *blur* and *other*, with 6 severity levels. Due to the simplicity of MNIST, we do not consider the corruptions on this dataset.

Experiment settings. For CIFAR-10 we set the volume of the sample set as 10000, and for SVHN we set the volume as 26032, which are the same as the official dataset. We use SGD with momentum as an optimizer for ERM baseline, the compared methods and our method. The training details of the compared methods and PAT are reported in the supplementary material.

Corruption and Adversarial Robustness

To comprehensively evaluate model robustness, we conduct the experiment on clean, adversarial and corrupted datasets, respectively, and we compare our PAT (iteration step $k=3$) with ERM (Naive), PGD-AT and GDA.

Figure 2 shows the accuracy performance of different methods, from which we can get the following observations: (1) It is obvious that GDA achieves good performance for all types of corruptions, which meets part of the assumption in Theorem 1. However, the \mathcal{R}_{adv} for GDA could be less than PAT due to the lack of adversarial robust augmentation, and thus leads to weak adversarial robustness. (2) PGD-AT performs poorly on corruptions and clean examples, which might derive from the decrease of \mathcal{C}_i^T in Theorem 1, leading to the looser upper bound and weaker corruption robustness compared to PAT. (3) For all cases, models trained by PAT consistently achieve the most robust performance under both adversarial examples and corruptions, which means that PAT brings more data with higher complexity. In this way, \mathcal{R}_{adv} shrinks significantly and $(\mathcal{C}_i^T)^\epsilon \setminus \mathcal{C}_i^T$ increases, which leads to a tighter upper bound and the stronger adversarial robustness as well as corruption robustness.

Generalization Evaluation with Mixed Test

Xu and Mannor firstly explored a deep connection between model robustness and generalization. They proved that a weak notion of robustness is both sufficient and necessary for generalization ability. Intuitively, if the test data is similar to training data, then the corresponding test error should

Table 1: Experiment results using *Mixed Test* on CIFAR-10 and SVHN with VGG-16 and ResNet18, respectively. We combine benign, adversarial and corrupted examples into the test set in the same proportion.

VGG16 (l_{inf})	Cln+($\epsilon=2$)+Blur	Cln+($\epsilon=4$)+Blur	Cln+($\epsilon=2$)+Noise	Cln+($\epsilon=4$)+Noise	Cln+($\epsilon=2$)+Other	Cln+($\epsilon=4$)+Other	Cln+($\epsilon=2$)+All	Cln+($\epsilon=4$)+All
Naive	69.04%	60.53%	65.78%	57.27%	72.20%	63.69%	69.01%	60.50%
PGD-AT	80.75%	75.40%	81.06%	75.71%	81.04%	75.69%	80.95%	75.60%
GDA	79.45%	71.00%	81.40%	72.95%	81.32%	72.87%	80.72%	72.27%
PAT	84.80%	79.75%	85.48%	80.43%	85.53%	80.48%	85.27%	80.22%
VGG16 (l_2)	Cln+($c=0.2$)+Blur	Cln+($c=0.5$)+Blur	Cln+($c=0.2$)+Noise	Cln+($c=0.5$)+Noise	Cln+($c=0.2$)+Other	Cln+($c=0.5$)+Other	Cln+($c=0.2$)+All	Cln+($c=0.5$)+All
Naive	60.86%	60.84%	57.60%	57.58%	64.02%	64.00%	60.83%	60.81%
PGD-AT	69.55%	67.40%	69.86%	67.72%	69.84%	67.69%	69.75%	67.60%
GDA	62.48%	61.03%	64.44%	62.99%	64.35%	62.90%	63.76%	62.31%
PAT	74.09%	71.55%	74.77%	72.22%	74.82%	72.27%	74.56%	72.01%
ResNet18 (l_{inf})	Cln+($\epsilon=2$)+Blur	Cln+($\epsilon=4$)+Blur	Cln+($\epsilon=2$)+Noise	Cln+($\epsilon=4$)+Noise	Cln+($\epsilon=2$)+Other	Cln+($\epsilon=4$)+Other	Cln+($\epsilon=2$)+All	Cln+($\epsilon=4$)+All
Naive	86.42%	77.53%	83.84%	74.94%	86.50%	77.61%	85.59%	76.69%
PGD-AT	89.51%	82.28%	89.17%	81.94%	89.37%	82.15%	89.35%	82.12%
GDA	89.49%	82.08%	89.05%	81.65%	88.77%	81.36%	89.10%	81.70%
PAT	92.62%	87.68%	92.20%	87.26%	91.84%	86.89%	92.22%	87.28%
ResNet18 (l_2)	Cln+($c=0.2$)+Blur	Cln+($c=0.5$)+Blur	Cln+($c=0.2$)+Noise	Cln+($c=0.5$)+Noise	Cln+($c=0.2$)+Other	Cln+($c=0.5$)+Other	Cln+($c=0.2$)+All	Cln+($c=0.5$)+All
Naive	65.62%	65.99%	63.03%	63.40%	65.70%	66.07%	64.78%	65.15%
PGD-AT	71.58%	67.33%	71.24%	66.99%	71.44%	67.19%	71.42%	67.17%
GDA	72.14%	67.58%	71.70%	67.14%	71.42%	66.86%	71.75%	67.19%
PAT	83.28%	70.76%	82.86%	70.34%	82.49%	69.97%	82.88%	70.36%

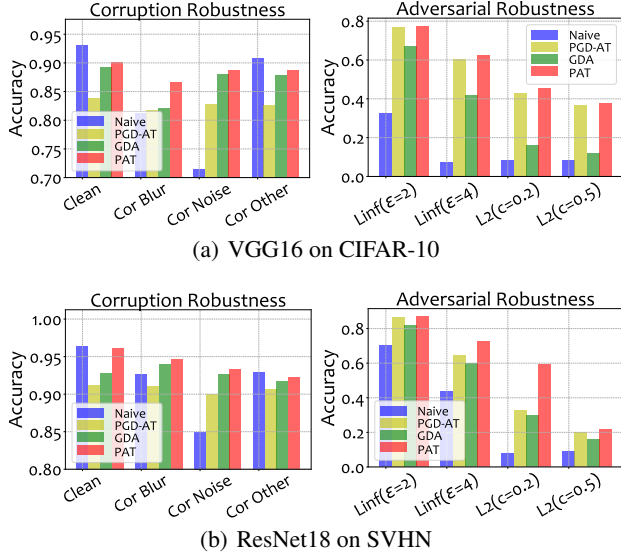


Figure 2: Model robustness evaluation on adversarial and corrupted examples of CIFAR-10 and SVHN. We evaluate corruption robustness on the clean and corrupted datasets, and adversarial robustness under PGD and C&W attack. The parameter setting and the experimental results of MNIST are reported in the supplementary material.

also be close to empirical training error as well. Previous adversarial related studies (Athalye, Carlini, and Wagner 2018;

Xie et al. 2018) usually evaluate the accuracy of model on individual type of examples, i.e., test set only contains adversarial examples or clean examples, and observe the gap between test and training error to obtain generalization ability. However, for models trained on dataset augmented with noises, e.g., adversarial examples, it is prejudiced and biased to evaluate model generalization on any specified type of test data, which turns out to be very different from the training set. Therefore, we believe that it is reasonable to evaluate model generalization and robustness with mixed types of samples under the noise situation, especially for those models trained with data augmentation containing noises. Thus, we propose *Mixed Test* to fairly evaluate model generalization ability, which combines clean, adversarial and corrupted examples in an equal proportion.

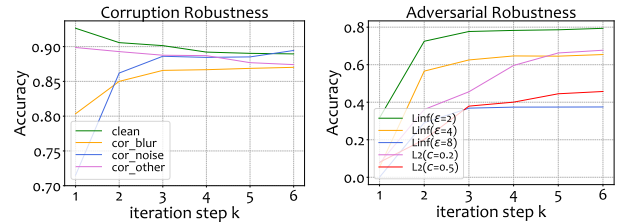


Figure 3: The corruption and adversarial robustness evaluation of Progressive Adversarial Training on CIFAR-10, with the iteration step k ranging from 1 to 6.

The experimental results using *Mixed Test* on CIFAR-10

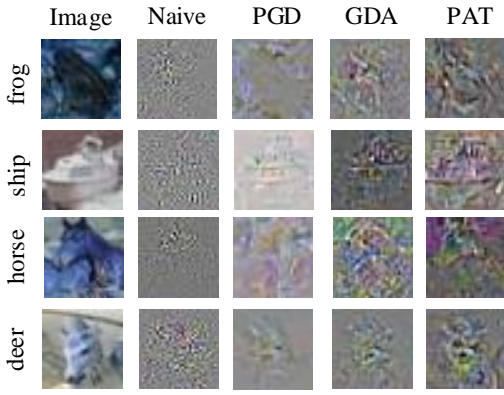


Figure 4: Visualization of the gradients with respect to the input images via different training strategies on CIFAR-10.

and SVHN are shown in Table 1. In order to conduct fair experiment, we keep all adversarial attack methods and corruption types the same as the section above. Obviously, PAT outperforms other data augmentation methods, i.e., PGD-AT and GDA, with big margins, showing the strongest generalization ability. As discussed before in Figure 1, since PGD-AT and GDA could increase model robustness against specific noise, e.g., obtaining higher adversarial accuracy, they fail to cover some portions of benign examples that are used to be classified by ERM, leading to weaker generalization ability. Also, it is interesting to see that the accuracy of GDA and ERM (Naive) drops more drastically compared to PGD-AT and PAT, when adversarial perturbation (e.g. ϵ or c) increases. This phenomenon further demonstrates the strong adversarial defense ability of PGD-AT and PAT.

Effects of Progressive Iteration

Our PAT relies on multiple progressive iterations to largely improve the noise diversity injected into deep models. To further investigate the effect of progressive iteration on model robustness and generalization, we conduct an experiment on PAT with different iterative step k . In fact, the iteration step k in our method can be regarded as the data augmentation times. From the results in Figure 3, we can draw two meaningful conclusions. Firstly, as the accuracy increases on clean examples, model robustness drops, which indicates that there exists a trade-off between accuracy and robustness in practice. Secondly, for a fixed perturbation threshold ϵ , the larger k , bringing larger data complexities and diversities during training, indeed supplies models with stronger adversarial and corruption robustness. Though the model robustness improves with more iteration steps, the training time consumption rises sharply at the same time. Moreover, the model robustness gradually saturates with a larger k , e.g., 5 or 6. Therefore, in practice we can simply choose a moderate value for k (e.g., 3) in order to keep the balance between robustness, accuracy and training time.

Feature Alignment with Human Vision

Beyond measuring the performance of model robustness and generalization with digital indicators (e.g., accuracy), we also try to discover model robustness more intuitively and perceptually using visualization techniques from the view of feature alignment with human vision. Since the stronger model generates gradients that are better aligned with human visual perception (Tsipras et al. 2018), we normalize and visualize the gradient of loss with respect to the input images on CIFAR-10. As shown in Figure 4, the gradients for Naive model turn out to be noises and appear meaningless to human, and those for PGD-AT and GDA are less perceptually aligned with human vision, e.g., too light, fewer outlines, etc. Among all methods, it is easy to observe that the gradients for PAT align much better with perceptually relevant features to human visual perception. This is mainly because that, in order to make consistent predictions, salient objects and critical features must be recognized more easily which result in robust and stable representations obtained from robust model itself. Thus, more broadly, robustness is more likely to be a reliable property for models which offers a path to build more human-aligned applications.

Conclusions

In this paper, we first theoretically established the connections between adversarial robustness and corruption robustness, by introducing a formal and uniform definition of the model robustness. Based on the surprising finding, we proposed a powerful training method named Progressive Adversarial Training (PAT), where diversified adversarial noises are aggregated, augmented and injected progressively to simultaneously guarantee both adversarial and corruption robustness. Theoretical analyses have been dedicated to proving the upper bound of model generalization and robustness, which in turn promise model with strong generalization ability and robustness against noises. Extensive experiments on MNIST, CIFAR-10 and SVHN show that PAT comprehensively performs well compared to various augmentation methods.

References

- [Athalye, Carlini, and Wagner 2018] 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*.
- [Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- [Carlini and Wagner 2017] 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.
- [Carlini et al. 2019] Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; and Madry, A. 2019. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.

- [Charles et al. 2019] Charles, Z.; Rajput, S.; Wright, S.; and Papailiopoulos, D. 2019. Convergence and margin of adversarial training on separable data. *arXiv preprint arXiv:1905.09209*.
- [Dodge and Karam 2017] Dodge, S., and Karam, L. 2017. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, 1–7. IEEE.
- [Fawzi, Fawzi, and Fawzi 2018] Fawzi, A.; Fawzi, H.; and Fawzi, O. 2018. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, 1178–1187.
- [Fawzi, Moosavi-Dezfooli, and Frossard 2016] Fawzi, A.; Moosavi-Dezfooli, S.-M.; and Frossard, P. 2016. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, 1632–1640.
- [Ford et al. 2019] Ford, N.; Gilmer, J.; Carlini, N.; and Cubuk, D. 2019. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*.
- [Goodfellow, Shlens, and Szegedy 2014a] Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014a. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- [Goodfellow, Shlens, and Szegedy 2014b] Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014b. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- [He et al. 2016] 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.
- [Hendrycks and Dietterich 2019] Hendrycks, D., and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.
- [Hinton et al. 2012] Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; and Sainath, T. N. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6):82–97.
- [Krizhevsky and Hinton 2009] Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.
- [Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, 1097–1105.
- [Kurakin, Goodfellow, and Bengio 2016] Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.
- [Kurakin, Goodfellow, and Bengio 2017] Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial machine learning at scale. In *International Conference on Learning Representations*.
- [LeCun et al. 1998] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.
- [LeCun 1998] LeCun, Y. 1998. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.
- [Liu et al. 2019] Liu, A.; Liu, X.; Fan, J.; Ma, Y.; Zhang, A.; Xie, H.; and Tao, D. 2019. Perceptual-sensitive gan for generating adversarial patches. In *33rd AAAI Conference on Artificial Intelligence*.
- [Madry et al. 2017] 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*.
- [Metz et al. 2019] Metz, L.; Maheswaranathan, N.; Shlens, J.; Sohl-Dickstein, J.; and Cubuk, E. D. 2019. Using learned optimizers to make models robust to input noise. *arXiv preprint arXiv:1906.03367*.
- [Netzer et al. 2011] Netzer, Y.; Wang, T.; Coates, A.; Bischoff, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.
- [Sankaranarayanan et al. 2018] Sankaranarayanan, S.; Jain, A.; Chellappa, R.; and Lim, S. N. 2018. Regularizing deep networks using efficient layerwise adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [Schmidt et al. 2018] Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 5014–5026.
- [Shaham, Yamada, and Negahban 2018] Shaham, U.; Yamada, Y.; and Negahban, S. 2018. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing* 307:195–204.
- [Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- [Sinha, Namkoong, and Duchi 2017] Sinha, A.; Namkoong, H.; and Duchi, J. 2017. Certifiable distributional robustness with principled adversarial training. *stat* 1050:29.
- [Sun et al. 2018] Sun, Z.; Ozay, M.; Zhang, Y.; Liu, X.; and Okatani, T. 2018. Feature quantization for defending against distortion of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7957–7966.
- [Sun, Zhu, and Lin 2019] Sun, K.; Zhu, Z.; and Lin, Z. 2019. Towards understanding adversarial examples systematically: Exploring data size, task and model factors. *arXiv preprint arXiv:1902.11019*.
- [Tsipras et al. 2018] Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. Robustness may be at odds with accuracy. *stat* 1050:11.
- [Tsuzuku, Sato, and Sugiyama 2018] Tsuzuku, Y.; Sato, I.; and Sugiyama, M. 2018. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural

networks. In *Advances in Neural Information Processing Systems*, 6541–6550.

[Wellner and others 2013] Wellner, J., et al. 2013. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

[Xie et al. 2018] Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2018. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*.

[Xu and Mannor 2012] Xu, H., and Mannor, S. 2012. Robustness and generalization. *Machine learning* 86(3):391–423.

[Zhai et al. 2019] Zhai, R.; Cai, T.; He, D.; Dan, C.; He, K.; Hopcroft, J.; and Wang, L. 2019. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*.

[Zheng et al. 2016] Zheng, S.; Song, Y.; Leung, T.; and Goodfellow, I. 2016. Improving the robustness of deep neural networks via stability training. In *IEEE conference on computer vision and pattern recognition*.

Supplementary Material: Towards Noise-Robust Neural Networks via Progressive Adversarial Training

Hang Yu, Aishan Liu, Xianglong Liu*, Jichen Yang, Chongzhi Zhang

State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
 {hyu0829,liuaishan, sxxayjc}@buaa.edu.cn, {xlliu,chongzhizhang}@nlsde.buaa.edu.cn

Proof of Theorem 1

Lemma 2 (Gaussian Isoperimetric Inequality).

$$\Phi^{-1}(\gamma(\mathcal{A}^h)) \geq \Phi^{-1}(\gamma(\mathcal{A})) + h, \quad (1)$$

where $\gamma(\cdot)$ denotes Gaussian measure, and the set \mathcal{A}^h is called the h -extension respect to set \mathcal{A} . Here, we give the proof of Theorem 1.

Recall the definitions of \mathcal{C}_i^T , \mathcal{C}_i^E and \mathcal{C}_i in subsection Model Robustness and Theorem 1, the definition and illustration of ε -extension in the same subsection and Figure 1, as well as the robustness sets \mathcal{R}_{cor} and \mathcal{R}_{adv} . Intuitively, the smaller volume of \mathcal{R}_{cor} and \mathcal{R}_{adv} , the stronger robustness the model performs.

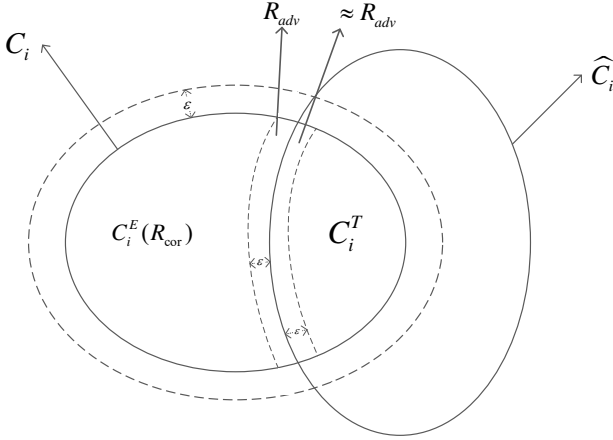


Figure 1: The relationship between adversarial and corruption robustness set.

As illustrated in Figure 1, we define the ε -band of set \mathcal{A}^E as $\mathcal{A}^E \setminus \mathcal{A}$. With Lemma 2, we could obtain that,

$$\begin{aligned} \mathbb{P}(\mathcal{R}_{cor}^E) &= \mathbb{P}(\mathcal{R}_{cor}) + \mathbb{P}(\mathcal{R}_{cor}^E \setminus \mathcal{R}_{cor}) \\ &= \mathbb{P}(\mathcal{R}_{cor}) + \mathbb{P}(\mathcal{C}_i^E \setminus \mathcal{C}_i) - \mathbb{P}((\mathcal{C}_i^T)^E \setminus \mathcal{C}_i^T) + 2\mathbb{P}(\mathcal{R}_{adv}) \\ &\geq \Phi(\Phi^{-1}(\mathbb{P}(\mathcal{R}_{cor})) + \varepsilon). \end{aligned}$$

*Corresponding author.

In the training process, if \mathcal{R}_{adv} shrinks with data augmentation, the intersection of $\hat{\mathcal{C}}_i$ and \mathcal{C}_i increases, while \mathcal{C}_i stays the same. This will lead to the increase in $(\mathcal{C}_i^T)^E \setminus \mathcal{C}_i^T$ which can be written as an inequality shown below:

$$\mathbb{P}(\mathcal{C}_i^E \setminus \mathcal{C}_i) - \mathbb{P}((\mathcal{C}_i^T)^E \setminus \mathcal{C}_i^T) + 2\mathbb{P}(\mathcal{R}_{adv}) \geq \Phi(\Phi^{-1}(\mathbb{P}(\mathcal{R}_{cor})) + \varepsilon) - \mathbb{P}(\mathcal{R}_{cor}).$$

Therefore, the left hand side of the above inequality decreases, i.e. the upper bound of right hand side decreases. We will prove the monotone interval in the following paragraph.

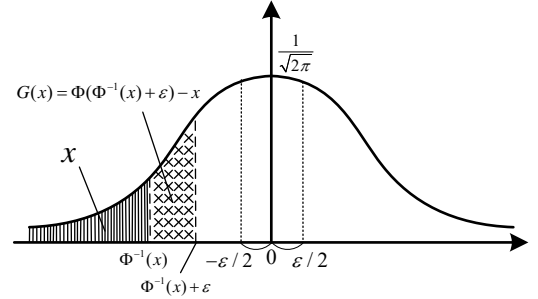


Figure 2: The monotone interval of $G(x)$.

Given the function $G(x) = \Phi(\Phi^{-1}(x) + \varepsilon) - x$, we will prove the monotone interval of $G(x)$ in a geometric view. Figure 2 represents the probability density function $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ of $N(0, 1)$. According to the definition, the left area of shaded lines denotes x in $G(x)$ with the corresponding abscissa of shaded area denoted as $\Phi^{-1}(x)$. Then we add ε to $\Phi^{-1}(x)$ and obtain $\Phi(\Phi^{-1}(x) + \varepsilon)$, which is the sum of left shaded area and crossed area, i.e., the integral of the right crossed area's abscissa. Thus, it is reasonable to say the crossed area denotes $G(x)$. In this way, with the increasing of $\Phi^{-1}(x)$ on negative half axis, the area of $G(x)$ increases and achieve maximum value until $\Phi^{-1}(x) = -\frac{\varepsilon}{2}$, where $G(x)$ reaches the maximum when $x = \Phi(-\frac{\varepsilon}{2})$. Therefore, $G(x)$ is monotonic increasing when $x \in (0, \Phi(-\frac{\varepsilon}{2})]$, which explains Theorem 1 that corruption robustness increases with the ascent of adversarial robustness.

Proof of Theorem 2

Lemma 3 ([Volpi et al. 2018], Claim 5). *For surrogate loss satisfying Assumption 1, we have*

$$\phi_\lambda(\theta; z) = \ell(\theta; z) + \frac{1}{\lambda} \|\nabla_z \ell(\theta; z)\|^2 + O\left(\frac{1}{\lambda^2}\right), \quad (2)$$

and

$$\phi_\lambda(\theta; z) \leq \ell(\theta; z) + \frac{1}{\lambda - L_1} \|\nabla_z \ell(\theta; z)\|^2. \quad (3)$$

According to Lemma 3 and Assumption 1, for $z, z' \in \mathcal{Z}$ we have:

$$\begin{aligned} |\phi_\lambda(\theta; z) - \phi_\lambda(\theta; z')| &= \ell(\theta; z) + \frac{1}{\lambda} \|\nabla_z \ell(\theta; z)\|^2 + O\left(\frac{1}{\lambda^2}\right) \\ &\quad - \ell(\theta; z') - \frac{1}{\lambda} \|\nabla_{z'} \ell(\theta; z')\|^2 - O\left(\frac{1}{\lambda^2}\right) \\ &\leq |\ell(\theta; z) - \ell(\theta; z')| + \frac{2M_1}{\lambda} \|\nabla_z \ell(\theta; z)\| \\ &\quad - \|\nabla_{z'} \ell(\theta; z')\| + O\left(\frac{1}{\lambda^2}\right) \\ &\leq L_0 \|z - z'\| + \frac{2M_1}{\lambda} L_1 \|z - z'\| + O\left(\frac{1}{\lambda^2}\right) \\ &\leq \left(L_0 + \frac{2M_1 L_1 + 1}{\lambda}\right) \|z - z'\|. \end{aligned}$$

Thus, surrogate loss satisfies L_λ -Lipschitzian. According to Lemma 1, for all $\gamma > 0$ we obtain $(K, \gamma L_\lambda)$ robust, where $K = N(\gamma/2, \mathcal{Z}, \|\cdot\|)$ denotes $\gamma/2$ -covering number of \mathcal{Z} in metric $\|\cdot\|$.

Proof of Theorem 3

Lemma 4 ([Xu and Mannor 2012], Theorem 3). *If $f_S(\theta; \cdot)$ satisfies $(K, \gamma L_\lambda)$ robust, for probability $p > 0$, there exists at least probability of $1 - p$, such that*

$$\ell(\theta; \cdot) \leq \frac{1}{n} \sum_{i=1}^n \hat{\ell}(\theta; \cdot) + \gamma L_\lambda + M_0 \left(\sqrt{\frac{2K \ln 2 + 2 \ln(\frac{1}{p})}{n}} \right), \quad (4)$$

where $\hat{\ell}$ and K denotes empirical training loss and the number of partitions after classification on \mathcal{Z} , respectively. According to Assumption 1 and Lemma 3, then we have inequality below:

$$\phi_\lambda(\theta; z) \leq \ell(\theta; z) + \frac{1}{\lambda - L_1} \|\nabla_z \ell(\theta; z)\|^2 \leq \frac{M_1^2}{\lambda - L_1}.$$

Meanwhile, according to Lemma 4 and Theorem 2, for $z, z' \in \mathcal{Z}$ and $\forall \gamma > 0$, we could obtain the upper bound of

our surrogate loss as follows:

$$\begin{aligned} \phi_\lambda(\theta; z) &\leq \ell(\theta; z) + \frac{1}{\lambda - L_1} \|\nabla_z \ell(\theta; z)\|^2 \\ &\leq \ell(\theta; z) + \frac{M_1^2}{\lambda - L_1} \\ &\leq \frac{1}{n} \sum_{i=1}^n \hat{\ell}(\theta; z) + \left(L_0 + \frac{2M_1 L_1 + 1}{\lambda}\right) \gamma \\ &\quad + M_0 \left(\sqrt{\frac{2N(\gamma/2, \mathcal{Z}, \|\cdot\|) \ln 2 - 2 \ln p}{n}} \right) + \frac{M_1^2}{\lambda - L_1} \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\ell}(\theta; z) + \gamma L_\lambda + M_0 \left(\sqrt{\frac{2K \ln 2 - 2 \ln p}{n}} \right) + \frac{M_1^2}{\lambda - L_1}, \end{aligned}$$

where $K = N(\gamma/2, \mathcal{Z}, \|\cdot\|)$ denotes $\gamma/2$ -covering number of \mathcal{Z} in distance metric $\|\cdot\|$, and $n = |\mathcal{S}|$ stands for the volume of augmented dataset.

More Experimental Results

In this section, we give more experimental results including the robustness evaluation on MNIST, the effects of progressive iteration and the feature alignment with human vision.

As for the experiment settings, we set the SGD momentum factor in optimizer as 0.9. We train 80 epochs for each method on CIFAR-10 and SVHN, and 40 epochs on MNIST. Specifically, for augmentation method PGD-AT in the training process, we mixed benign and adversarial examples in equal proportion, following the recommendation of (Kurakin, Goodfellow, and Bengio 2016).

For PGD-AT on l_{inf} metric, we set perturbation $\epsilon=8/255$, step size $\alpha=\epsilon/10$ and iteration step $k=10$. For GDA, we add noise $N(0, \sigma)$ on training set and fix $\sigma=0.1$. For our progressive adversarial training method on l_2 metric, we set perturbation $\epsilon=1.0$, $\beta=0.5$ and iteration step $k=3$. Motivated by Curriculum Learning (Bengio et al. 2009), we gradually add perturbation ϵ in PAT.

Robustness Evaluation on MNIST

For the simplicity of MNIST, corruption robustness is not considered. We only consider the robustness of model on adversarial examples. Among those attack methods, for PGD attack, we set the parameters as $\epsilon=0.1$ and 0.2 , step size $\alpha=\epsilon/10$ and iteration $k=10$. As for C&W attack, we set perturbation to $c=0.1$ and 0.2 . The evaluation results can be found in Figure 3.

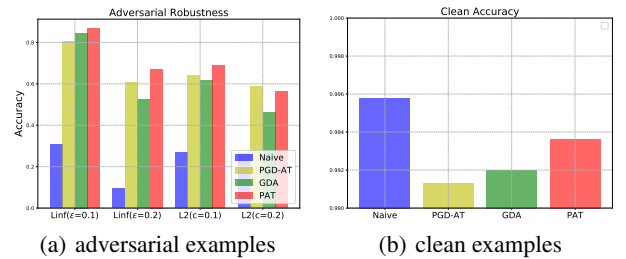


Figure 3: Robustness evaluation on MNIST with LeNet-5.

Table 1: Generalization evaluation of mixed results on CIFAR-10, with the different iteration step k ranging from 1 to 6.

VGG16(l_{inf})	Cln+($\epsilon=2$)+Blur	Cln+($\epsilon=4$)+Blur	Cln+($\epsilon=2$)+Noise	Cln+($\epsilon=4$)+Noise	Cln+($\epsilon=2$)+Other	Cln+($\epsilon=4$)+Other	Cln+($\epsilon=2$)+All	Cln+($\epsilon=4$)+All
PAT(k=1)	68.22%	59.46%	65.27%	56.51%	71.39%	62.63%	68.30%	59.54%
PAT(k=2)	85.02%	79.72%	85.42%	80.12%	85.78%	80.48%	85.41%	80.10%
PAT(k=3)	84.80%	79.75%	85.48%	80.43%	85.53%	80.48%	85.27%	80.22%
PAT(k=4)	84.75%	80.22%	85.31%	80.78%	85.39%	80.86%	85.15%	80.62%
PAT(k=5)	84.44%	79.82%	84.99%	80.37%	85.04%	80.42%	84.82%	80.20%
PAT(k=6)	84.72%	80.13%	85.52%	80.93%	85.58%	80.99%	85.27%	80.68%
VGG16(l_2)	Cln+(c=0.2)+Blur	Cln+(c=0.5)+Blur	Cln+(c=0.2)+Noise	Cln+(c=0.5)+Noise	Cln+(c=0.2)+Other	Cln+(c=0.5)+Other	Cln+(c=0.2)+All	Cln+(c=0.5)+All
PAT(k=1)	60.12%	60.32%	57.17%	57.38%	63.29%	63.50%	60.20%	60.40%
PAT(k=2)	72.90%	67.56%	73.30%	67.96%	73.66%	68.32%	73.29%	67.95%
PAT(k=3)	74.09%	71.55%	74.77%	72.22%	74.82%	72.27%	74.56%	72.01%
PAT(k=4)	78.53%	72.01%	79.08%	72.57%	79.17%	72.65%	78.93%	72.41%
PAT(k=5)	80.38%	73.13%	80.93%	73.68%	80.99%	73.74%	80.77%	73.52%
PAT(k=6)	79.89%	71.76%	80.70%	72.57%	80.76%	72.63%	80.45%	72.32%

Effects of Progressive Iteration

In this part, we give more details of the experiment on models with different progressive iterations. For the robustness and generalization evaluation in Table 1, we observe that as iteration step k increases, PAT performs stronger adversarial robustness on both l_{inf} and l_2 adversarial examples, as well as stronger corruption robustness on noise corruption set. However, no obvious increase is witnessed on blur and other corruption sets after $k > 2$. In conclusion, the progressive iteration mechanism can significantly improve model robustness on both adversarial and corruption conditions.

et al. 2018). Thus, we visualize the gradient of loss with respect to inputs on CIFAR-10. Here is the additional visualization results of feature alignment.

From the visualization of gradients shown in Figure 4, we can observe distinctly that gradients become more semantically meaningful as model becomes more robust, i.e., from PGD-AT to PAT incrementally.

References

- Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48. ACM.
- Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.
- Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. Robustness may be at odds with accuracy. *stat* 1050:11.
- Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 5334–5344.
- Xu, H., and Mannor, S. 2012. Robustness and generalization. *Machine learning* 86(3):391–423.

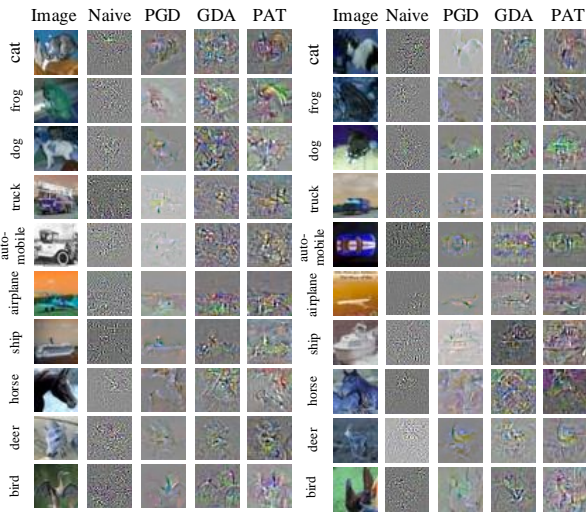


Figure 4: Feature alignment with human vision.

Feature Alignment with Human Vision

We believe that, gradient features become more perceptually relevant with human perception as the model becomes stronger, which we can also find similar points from (Tsipras