Towards Interpreting and Utilizing Symmetry Property in Adversarial Examples

Shibin Mei*, Chenglong Zhao*, Bingbing Ni†, Shengchao Yuan

Shanghai Jiao Tong University, Shanghai 200240, China
{adair327, cl-zhao, nibingbing, sc_yuan}@sjtu.edu.cn

Abstract

In this paper, we identify symmetry property in adversarial scenario by viewing adversarial attack in a fine-grained manner. A newly designed metric called attack proportion, is thus proposed to count the proportion of the adversarial examples misclassified between classes. We observe that the distribution of attack proportion is unbalanced as each class shows vulnerability to particular classes. Further, some class pairs correlate strongly and have the same degree of attack proportion for each other. We call this intriguing phenomenon symmetry property. We empirically prove this phenomenon is widespread and then analyze the reason behind the existence of symmetry property. This explanation, to some extent, could be utilized to understand robust models, which also inspires us to strengthen adversarial defenses.

Introduction

Remarkable success has been achieved in computer vision community due to the significant development of deep neural networks (DNNs). However, recent researches (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017) have shown the fatal vulnerability of DNNs based systems. Namely, natural images added with human-imperceptible perturbations, dubbed *adversarial examples*, can fool the well-pretrained classifier into predicting incorrect labels. As the increasing trend of deploying DNNs based approaches in safety- and security-critical devices, such as face verification and self-driving, it is a crucial topic to understand and explore some of the properties and reasons for the existence of adversarial examples.

Instead of explaining the adversarial examples as a whole (Goodfellow, Shlens, and Szegedy 2014; Athalye, Carlini, and Wagner 2018), we consider the adversarial examples from the perspective of the mutual relationship between different classes. In particular, we give a new metric of adversarial attack, called *attack proportion*, to count the proportion of the adversarial examples misclassified between classes. Instead of simply using attack success rate (Madry et al. 2017; Goodfellow, Shlens, and Szegedy 2014), the defined *attack proportion* allows us to investigate the adversarial examples from a fine-grained perspective. Based on attack proportion, we observe an intriguing phenomenon that adversarial examples crafted from a certain class are prone to be misclassified as particular classes. Besides, we further find that the attack proportion of adversarial examples crafted from class i misclassified as class j is almost equal to that of crafted from class j misclassified as class i. In this paper, we define the observed phenomenon as **symmetry property** in adversarial examples. Through a lot of experimental results under recent popular adversarial attacks, we ensure that the symmetry property is common in various kinds of attack methods and neural networks (See Fig. 3). More importantly, we also find that the maximum attack proportion is bi-directional for some class pairs, and we define this kind of pairs as **maximum symmetry pair** (See Fig. 1). It means that these class pairs possess a strong relationship in adversarial examples. Therefore, we pose the following question:

> *Why does symmetry property exist in adversarial examples? Is there any relation between the property and adversarial robustness? And could we take advantage of it to defend against adversarial attacks?*

We observe that classes belonging to maximum symmetry pair show similar appearance, e.g., automobile and truck in CIFAR-10 (See Fig. 1). Based on this, we assume that samples with similar appearance, though different classes, have similar mappings in feature space. These feature points lay aside the class-wise decision boundary, which can be easily made into adversarial examples by adding a small noise. In other words, the features of samples extracted by DNNs are not discriminative enough, and there lacks a large support margin between different classes. Recent works (Liu et al. 2016; Pang et al. 2019) also show that the common training scheme of DNNs, i.e., the softmax cross-entropy loss, merely builds the mapping between samples and their corresponding labels while ignores the relationship among samples in the feature space. To this end, we explore the similarity matrix (See Fig. 5) of weights of the classifier, where the weights are considered as the prototype of features (Wang et al. 2018; Deng et al. 2019). We note that the similarity matrix is corresponding to the symmetry property, i.e., more degree of similarity between different classes leads to more attack proportion.

This inspires us to develop a new adversarial defense

*These authors contributed equally.
†Corresponding author.

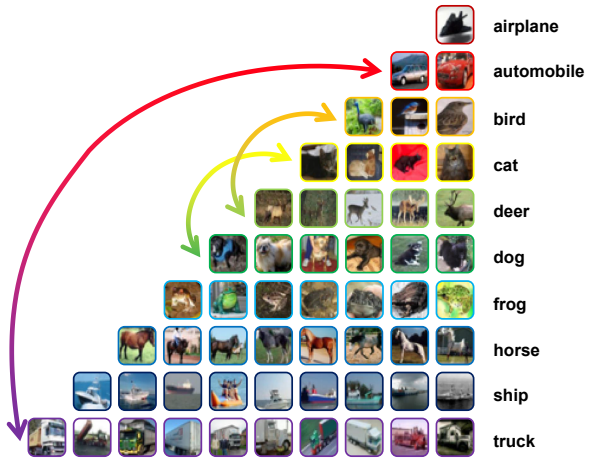


Figure 1: Maximum Symmetry Pair. We mark the maximum symmetry pairs in CIFAR-10 dataset with color arrows.

method by decoupling the similarity between different classes. We consider the symmetry pairs deserve special treatment in robust training, that is, much stronger regularization should be attached to these pairs, which is very different from the previous methods (Mustafa et al. 2019; Pang et al. 2019, 2020). Prior methods, even if they are designed to learn more discriminative representations, treat the relationship between different class pairs equally. In this paper, we focus on the category preference in adversarial attacks inspired by symmetry property, and adjust the constraint to inter-class and intra-class representations with a fine-grained manner. Moreover, the awareness of the symmetry property brings a novel view to adversarial attacks, and can serve as a general regularization strategy.

Related Work

Adversarial Examples. Szegedy *et al.* (Szegedy et al. 2013) firstly find the intriguing phenomenon that clean samples added with human-imperceptive noise can mislead the DNNs based classifier, and also use an optimization-based algorithm to craft adversarial examples. FGSM(Fast Gradient Sign Method) (Goodfellow, Shlens, and Szegedy 2014) is proposed to utilize a single-step gradient to manufacture adversarial examples, which is more efficient than (Szegedy et al. 2013). Then PGD(Projected Gradient Descent) (Madry et al. 2017), an iterative variant of FGSM, is proposed to strengthen the adversarial attack. Moreover, Dong *et al.* (Dong et al. 2018) introduce a moment term to enhance the transfer attack ability of gradient-based attack methods (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017). All these popular attack methods depend on available access to the gradient of models, which is not practical in the physical world. To this end, some black-box attack methods are proposed without requirement of the full knowledge of neural networks, such as transfer-based methods (Papernot et al. 2017) and query-based methods. (Ilyas et al. 2018; Li et al. 2019; Guo, Yan, and Zhang 2019; Guo et al. 2019)

Adversarial Defense. To defend against attacks, many methods are proposed to enhance the adversarial robustness of models. From the perspective of min-max optimization, Madry *et al.* (Madry et al. 2017) propose to optimize the models with the crafted adversarial examples, and this setting is then followed by AVmixup (Lee, Lee, and Yoon 2020), TRADES (Zhang et al. 2019), MART (Wang et al. 2019). AVmixup introduces a soft-label trick to narrow the large gap between test and training accuracy in adversarial training. TRADES identified the trade off between model robustness and clean sample performance and MART proposed defense method by revisiting misclassified examples. By introducing non-differential units into networks to block the gradient back-propagation, Gradient masking based methods (Song et al. 2017) are proposed to prevent the generation of adversarial examples. Moreover, some methods (Mustafa et al. 2019; Pang et al. 2019) focus on achieving a large margin between different classes in latent feature space to defend against adversarial attacks.

Methodology

Preliminary

Suppose a parameterized K-class DNNs classifier $f_\theta(x) : x \mapsto y$, where $x \in \mathcal{X}$ denotes input data, $y \in \mathcal{Y} = \{1, 2, ..., K\}$ denotes output label and θ indicates the learnable parameter of the classifier. A well-trained classifier $f_\theta(x)$ can predict the label of the input correctly, i.e., $f_\theta(x) = y^*$ (ground-truth label). The goal of adversarial attack is to mislead the classifier $f_\theta(\cdot)$ and result in wrong output, which can be implemented by injecting malicious, human-imperceptive noise δ to the input data. In general, adversarial attacks can be divided into two ways, i.e., *non-targeted* and *targeted* ones. For non-targeted attack, adversarial example $x + \delta$ is crafted to mislead classifier, i.e., $f_\theta(x + \delta) \neq y^*$, where $x + \delta$ stays in the vicinity of x in L_p ball (2 or ∞). For *targeted attack* ones, adversarial example is crafted to fool the classifier into outputting undesirable specific label y_{adv}, where $y_{adv} \neq y^*$. In this paper, we only discuss non-targeted attack considering the meaningful of symmetry property.

Attack Proportion

Recall that natural samples belonging to class i added with adversarial perturbation will be misclassified as $y \in \mathcal{Y} \setminus \{i\}$. *Attack success rate* (ASR) (Madry et al. 2017; Dong et al. 2018) is utilized as a general metric to evaluate the performance of adversarial attack,which is formulated as follows:

$$ASR = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}[f_\theta(x_k + \delta) \neq y^*], \tag{1}$$

where δ indicates the crafted adversarial noise and y^* indicates the ground-truth label. As shown in Eqn. 1, ASR is a very summary metric, where all successful adversarial examples are regarded equally important without distinction. So, it is hard to conduct more exploration and investigation of adversarial attacks by ASR.

To this end, we propose a new evaluation metric in this paper to further analyze the results of adversarial attack. In

particular, we divide the input samples into several subsets $\{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$ by their labels, and the subscript of $\mathcal{I}_i$ indicates the label of samples in this subset. For each subset $\mathcal{I}_i$, we count the proportion of the adversarial samples classified as different classes, which can be depicted as follows:

$$\mathcal{P}(i \mapsto j) = \frac{1}{|\mathcal{I}_i|} \sum_{x_k \in \mathcal{I}_i} \mathbb{1}[f_\theta(x_k + \delta) = j], \tag{2}$$

where $|\cdot|$ indicates the sample number of the subset. The new metric $\mathcal{P}(i \mapsto j)$, dubbed *attack proportion*, reveals the distribution of the adversarial examples misclassified as different classes. From our experiments, we find that the attack proportion $\mathcal{P}(i \mapsto j)$ is unbalanced in subset I_i. Namely, some classes are vulnerable under adversarial attack, while others are not. Two demos are reported in Fig. 2, where we calculate the metric, attack proportion, on two classes on MNIST and CIFAR-10 datasets. It can be seen that 44.75% adversarial examples of *digit 3* are classified as *digit 5* in MNIST, while 32.10% that of *airplane* are classified as *bird* in CIFAR-10. One intuitive explanation for this phenomenon is that *digit 3* and *digit 5*, *airplane* and *bird* are very similar in appearance. Correspondingly, after mapping to feature space, they will distribute near the decision boundary. Small magnitude noise will encourage these samples to cross the the decision boundary and cause misclassification. Note that the symmetry property will not be changed under different valid attack budgets, that is, the adversarial noise does not damage the semantics. More validations can be found in the supplementary materials.

Symmetry Property Analysis

For a K-class dataset $\mathcal{D}$ and an attack method, we construct an attack matrix $\mathcal{M} \in \mathbb{R}^{K \times K}$, where each element $\mathcal{M}_{ij}$ in the matrix is equal to attack proportion $\mathcal{P}(i \mapsto j)$, i.e,

$$M_{ij} = \mathcal{P}(i \mapsto j). \tag{3}$$

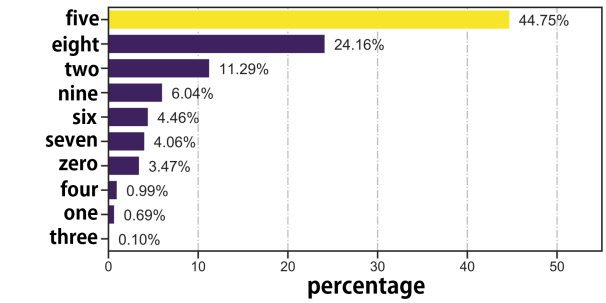
Definition 1 *(Symmetry Property) Given an attack matrix $\mathcal{M}$, there exits the symmetry property between class i and class j ($i \neq j$), if*

$$|\mathcal{M}_{ij} - \mathcal{M}_{ji}| < \epsilon, \tag{4}$$

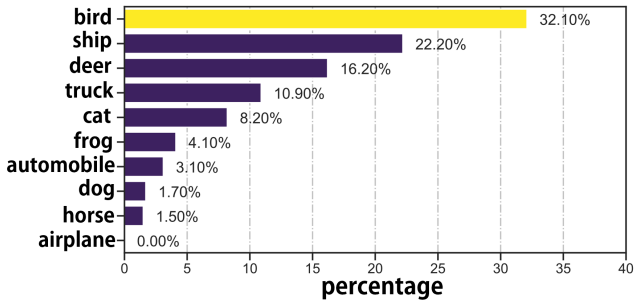
ϵ denotes an error term.

If the error term ϵ is set as zero, the restriction of the symmetry property is strict. However, it is hard to find any case which can satisfy this strictly restricted symmetry property in our empirical experiments, so we introduce a small error term ϵ to relax the definition of symmetry property, which facilities further exploration.

As shown in Fig. 3, we report the attack matrix $\mathcal{M}$ on CIFAR-10 dataset under various attack methods and models, i.e, FGSM and PGD attacks, VGG-16 and ResNet-34 models. These attack matrices are described by heatmap, and the depth of the color represents the attack proportion. Intuitively, these matrices are almost symmetric. That is, classes on CIFAR-10 satisfy the defined symmetry property. Moreover, we count the proportion that classes satisfy symmetry property under different values of the error term, as shown in Tab.1. Thus, the symmetry property is common in CIFAR-10, and has nothing to do with attack methods and DNN models. Furthermore, we also give the chord diagrams to visualize the results of classes that satisfy symmetry property in CIFAR-10 intuitively, as shown in Fig 3. Similar results can be drawn on CIFAR100 and ImageNet, as shown in Fig.4. We also put the experiments on train set and test set and on more wider model like WideResNet in the supplementary materials. The symmetry property also exists on defense model and we put the related results in supplementary materials. As mentioned in above section, the attack proportion is unbalanced, and most of the adversarial examples crafted from one subset will be misclassified as a certain class. In order to further discuss and analyze this unbalanced phenomenon based on the symmetry property, we give the definition of maximum symmetry pair as follows:



(a) MNIST: Digit 3



(b) CIFAR-10: Airplane

Figure 2: We report the results of attack proportion on MNIST and CIFAR-10 datasets. Two classes, i.e., digit 3 (a) and airplanes (b), are picked to display in this figure. The yellow color bar marks the maximum attack proportion.

Definition 2 *(Maximum Symmetry Pair) Let samples of class i and class j satisfy the symmetry property, then i and j are maximum symmetric pair, if*

$$\arg\max_j \mathcal{M}_{kj} = \arg\max_i \mathcal{M}_{ik}, \tag{5}$$

where k is an arbitrary class.

The *maximum symmetry pair* indicates a strong relationship between two classes. In particular, if most of the adversarial examples of *class i* are classified as *class j*, then the most of the adversarial examples of *class j* are classified as *class i*, and vice versa. As depicted in Fig. 1, we have picked up

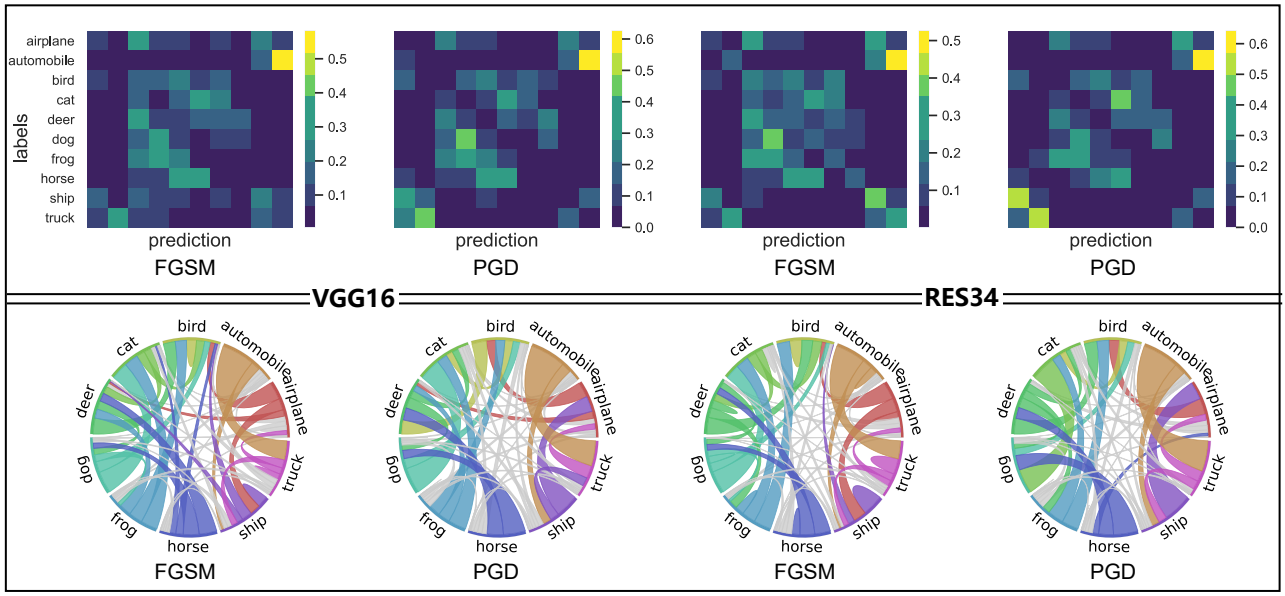
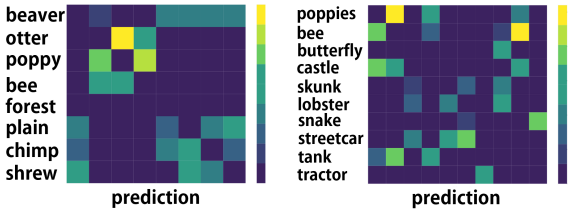
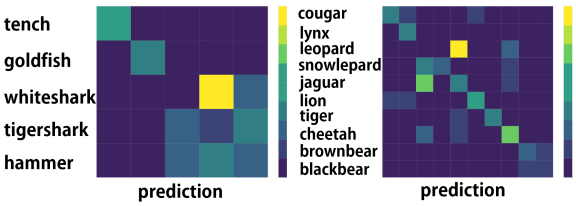


Figure 3: We report the attack results on CIFAR-10 dataset under different settings, i.e, FGSM and PGD attacks, VGG-16 and ResNet-34 models. The fist row shows the attack matrix, and the second row shows the chord diagram to visualize the relationship between different classes intuitively.



(a) CIFAR100 (ResNet50)



(b) ImageNet (Inception-v3)

Figure 4: Examples of symmetry pairs in CIFAR100 and ImageNet under PGD attack. We select 2 groups of categories respectively due to the large number of total categories.

Models	Methods	$\epsilon \leq$ 0.05	0.1	0.2
VGG-16	FGSM	55.5%	86.7%	95.6%
	PGD	71.1%	84.4%	97.8%
ResNet-34	FGSM	71.1%	84.4%	95.6%
	PGD	71.1%	91.1%	100%

Table 1: We count the ratio of class pairs which satisfy the symmetry property on CIFAR-10 dataset.

several pairs of classes on CIFAR-10 dataset which satisfy the defined *maximum symmetry pair*.

Why Does Symmetry Property Exist?

It is crucial to discuss this question seriously for developing further research. As shown in Fig. 1, we have marked several maximum symmetry pairs, such as (*automobile*, *truck*), (*cat*, *dog*). Obviously, these pairs show similar appearance and shape. An intuitive explanation is that images with similar shapes have similar mappings in feature space. That is, the models don't learn how to extract discriminative features especially for similar images. As a result, these features lay aside the decision boundary, which can be easily made into *adversarial examples* by adding small perturbation. To further verify this explanation, we conduct experiments to analyze the relationship between adversarial examples and features.

Following prior arts (Deng et al. 2019; Wang et al. 2018) which have shown that the weights of classifier of DNNs can represent as the prototype (or center) of features for each class. The key intuition is that the weights of classifier preserve the most abundant and semantic information for classification task. Suppose the weights of classifier are $\mathcal{W} = \{w_j\}_{j=1}^{K}$, where each vector w_j corresponds to a class and K is the number of classes. Then the similarity matrix $\mathcal{S} \in \mathbb{R}^{K \times K}$ of features for different classes can be computed

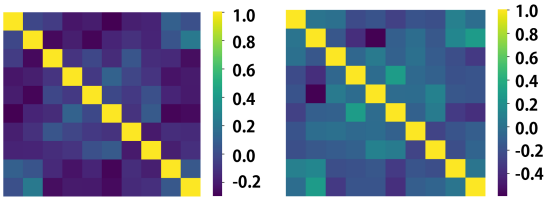


Figure 5: Similarity Matrices. The two matrices are computed on VGG-16 and ResNet-34 model which trained on CIFAR-10 dataset.

based on the $\mathcal{W}$, which is formulated as follows:

$$\mathcal{S}_{ij} = \frac{w_i \cdot w_j}{\|w_i\|_2 \|w_j\|_2} \tag{6}$$

As shown in Fig 5, we report the similarity matrix of VGG-16 and ResNet-34 model trained on CIFAR-10.

We observe that the similarity matrix $\mathcal{S}$ is corresponding to attack matrix $\mathcal{M}$ (see Fig. 3) for the same model. In particular, the more similarity in matrix $\mathcal{S}$ between different classes, the more attack proportion in matrix $\mathcal{M}$, and vice versa. Namely, less separability of inter-class features leads to a strong relationship between adversarial examples. We argue that the original loss function, i.e. cross-entropy, of DNNs does not explicitly impose any constraints to learn discriminative representation for classification, and more importantly, treats all categories equally by merely focusing on building the mapping between the input and its label and ignoring the uneven distribution of different categories in feature space. As a result, samples of different classes, especially for classes of symmetry pairs, will be mapped into the similar region, and there lacks enough support margin to distinguish them. It is hard to obtain an accurate and robust decision boundary to distinguish these samples, so inputs disturbed by small perturbation can easily go across to the region of other classes and cause misclassification.

Defense via Symmetry Property

Based on the analysis of symmetry property in adversarial examples, we propose a novel regularizer scheme to achieve special constraint for symmetry pairs. Some recent defense works (Liu et al. 2016; Pang et al. 2019, 2020) also starts from learning discriminative features by enlarging inter-class distance and reducing intra-class distance simultaneously but all class pairs are equally treated. They have shown that imposing a proper constraint on the weights of classifier can obtain structured features, which will enhances the adversarial robustness of neural networks. The difference between the previous methods is the constraint of the cosine angular Θ of inter-class. MMC (Pang et al. 2019) pre-calculate Θ by maximizing Mahalanobis center loss; WP (Xu, Li, and Yang 2020) maximize the hypersphere distance and set Θ as $\frac{1}{1-K}$(K represents the number of classes); PCL (Mustafa et al. 2019) and HE (Pang et al. 2020) adaptively adjust Θ during robust training. Instead of treating the all cosine angular of inter-class equally, we propose to introduce the prior knowledge of *maximum symmetry pairs* to construct the symmetry-aware loss function.

For a dataset $\{x_i, y_i\}_{i=1}^N$, the features $z = g(x)$ are extracted from CNN and then fed into the classifier to obtain the predicted label. The weight matrix W of classifier is normalized to facilitate computing the cosine angular of inter-class. Then the inter-class constraint loss is defined as follows:

$$\mathcal{L}_{inter} = \|W^T W - \Sigma\|_2, \tag{7}$$

$$where \quad \Sigma = \begin{cases} 1, & i = j \\ \Theta - \zeta, & i \neq j \wedge (i,j) \in \mathcal{O} \\ \Theta + \widetilde{\zeta}, & i \neq j \wedge (i,j) \notin \mathcal{O} \end{cases}, \tag{8}$$

$\mathcal{O}$ indicates the set of *maximum symmetry pairs*. ζ is an adjustable parameter, while $\widetilde{\zeta}$ is a compensation for ζ, which guarantees the existence of optimal solution. The defined loss encourages to decrease the cosine similarity of different weights and give full attention to maximum symmetry pairs. We pick the most succinct Θ scheme for simplicity as (Xu, Li, and Yang 2020) and achieve state-of-the-art defense performance, which demonstrate the effectiveness of our method to add an angular parameter ζ to further enlarge the angular of maximum symmetry pairs. We have also tried to apply constraints to all symmetry pairs but obtain little improvement. Moreover, we propose to use MSE (Mean Square Error) loss to construct an intra-class constraint, which is formulated as follows:

$$\mathcal{L}_{intra} = \|zW - \Sigma(y,:)\|_2 \tag{9}$$

The defined intra-class loss encourages the extracted features to approach the prototype feature of its corresponding class, while keeping a certain angular margin from the prototype features of other classes. We combine the defined intra-class and inter-class loss to get the final objective function as follows:

$$\mathcal{L} = \mathcal{L}_{intra} + \alpha \mathcal{L}_{inter}, \tag{10}$$

where α is a trade-off hyper-parameter. As previous works (Wang et al. 2019; Zhang et al. 2019; Pang et al. 2020, 2019), we combine our newly designed loss with the framework of adversarial training. Moreover, we find during the training process, although the distribution of attack proportion may change, the symmetry property and the maximum symmetry pair remain unchanged. Therefore, we fix the maximum symmetry pairs during the training process.

Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed adversarial defense method, which well demonstrates the favorable performance of our method compared with recent prior arts.

Architectures and Datasets. We conduct experiments on several benchmark datasets including CIFAR-10 and SVHN. We use ResNet34 (He et al. 2016) and WideResNet28×10 for experiments on CIFAR-10, ResNet-34 for experiments on SVHN. We also conduct experiments on larger datasets such as CIFAR100 and TinyImageNet. All these networks are trained for 200 epochs with batch size 128, and optimized by stochastic gradient descent (SGD), with Nesterov

Method	Clean	FGSM	PGD50	CW	AA
ResNet34					
Standard	**92.96**	14.24	1.00	0.34	0.00
PCL	83.12	57.65	44.15	43.44	37.47
WP	84.79	59.33	45.86	45.51	41.13
LS	81.87	57.43	43.60	44.05	38.15
A.T.	80.67	57.31	45.02	44.59	39.75
MMC	82.29	59.17	46.24	42.40	40.31
HE	83.22	61.26	46.26	44.23	42.19
TRADES	80.94	59.43	48.18	47.24	46.90
MART	82.47	61.16	48.19	48.37	46.34
Ours	84.00	**63.45**	**49.62**	**49.72**	**47.63**
WideResNet28×10					
Standard	**93.50**	18.79	0.00	0.37	0.00
PCL	86.44	63.18	47.19	46.43	43.98
WP	85.74	64.85	48.95	47.92	46.14
LS	84.17	63.39	46.76	46.06	43.84
A.T.	83.14	63.47	46.48	46.34	44.07
MMC	81.89	64.83	48.11	47.31	45.48
HE	85.48	65.16	49.98	49.03	47.25
TRADES	83.90	64.33	49.87	49.69	48.20
MART	84.78	65.12	50.26	50.13	49.16
Ours	86.53	**66.38**	**52.48**	**51.92**	**50.47**

Table 2: Evaluation of the defense ability of the proposed method on CIFAR-10 dataset. All experiments are conducted on ResNet34 and WideResNet28×10 model. We also compare our method with several state-of-the-art defense methods, i.e., PCL, WP, LS, A.T., MMC, HE, TRADES, MART. Classification accuracy(%) is used as the evaluation metric.

momentum 0.9 and weight decay 5×10^{-4}. The learning rate is set as 0.1 and divided by 10 at 100 and 150 epochs, respectively.

Defense Methods. We compare the proposed method with recent defense methods, including standard trained model with softmax-cross-entropy loss(Standard), prototype conformity loss (PCL) (Mustafa et al. 2019), Weight Penalization (WP) (Xu, Li, and Yang 2020), Large margin softmax(LS) (Liu et al. 2016), adversarial training (A.T.) (Madry et al. 2017), Hypersphere Embedding (HE) (Pang et al. 2020), Max-Mahalanobis center (MMC) (Pang et al. 2019), TRADES (Zhang et al. 2019) and MART (Wang et al. 2019). All these defense methods are implemented as their official settings.

We empirically set the parameter $\zeta = 0.1$ and $\alpha = 2$ in our experiments. Parameter analysis can be found in the supplementary materials.

Results under White-box Setting

In this section, we evaluate the defense ability of the proposed method under several white-box attacks. Several state-of-the-art white-box adversarial attacks, including FGSM (Goodfellow, Shlens, and Szegedy 2014), PGD (Madry et al. 2017), CW (Carlini and Wagner 2017) and AutoAttack (AA) (Croce and Hein 2020), are introduced to conduct defense experiments. We set the l_∞ distortion parameter, i.e., adversarial perturbation budget, as $\epsilon = 8/255$ for FGSM, PGD50, CW, and AA. As shown in Tab. 2, the experiment results on CIFAR-10 are reported on two neural networks, i.e, ResNet34 and WideResNet28×10. The proposed method has achieved favorable defense performance against various attacks, with relatively little accuracy degradation on clean samples. Besides, compared with recent state-of-the-art defense methods, i.e, PCL, WP, LS, A.T., MMC, HE, TRADES and MART, the proposed method exhibits outstanding performance, even under iterative-based attack methods with stronger attack ability.

Methods	Clean	FGSM	PGD10	PGD20	PGD50
Standard	96.10	15.44	1.15	0.51	0.33
PCL	96.27	77.24	58.75	56.99	52.03
WP	96.28	76.39	58.11	57.80	57.77
LS	96.31	76.67	57.80	55.29	53.33
A.T.	92.40	74.96	58.27	57.74	55.88
MART	94.59	78.03	59.29	59.17	58.13
Ours	**96.55**	**79.31**	**60.87**	**60.16**	**59.78**

Table 3: Evaluation of the defense ability on the SVHN dataset. All experiments are conducted on the ResNet-34 model. Classification accuracy(%) is used as the evaluation metric.

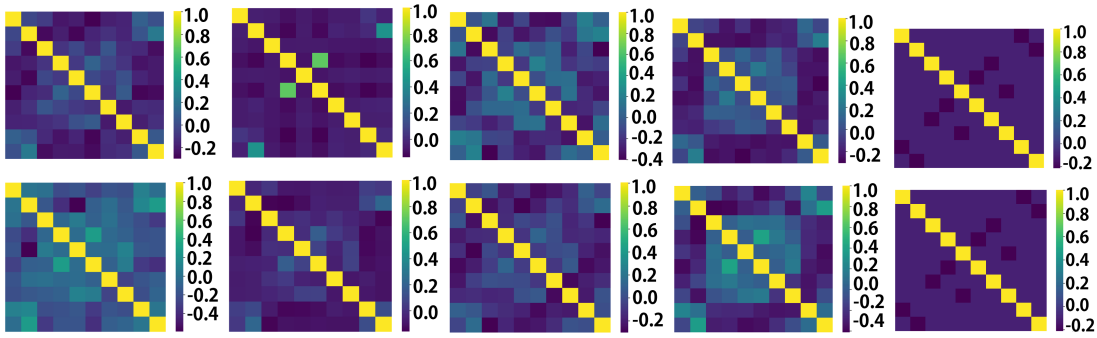
Tab. 3 shows our experimental results on the SVHN dataset. We present classification accuracy compared with PCL, WP, LS, A.T. and MART under FGSM, PGD10, PGD20, PGD50 attacks. As can be seen in Tab. 3, we achieve favorable performance under strong white-box attacks, similar with that of on CIFAR-10. For margin based defense methods, we only choose the representative PCL, WP and LS for comparison.

Results under Black-box Setting

In order to further validate the robustness of the proposed method under the black-box setting, we perform black-box attacks on CIFAR-10, which is more challenging for attacks without access to the parameters and structures of the models. In this section, we apply four black-box attack algorithms, i.e., NES (Ilyas et al. 2018), Nattack (Li et al. 2019), SimBA (Guo et al. 2019), and Subspace attack (Guo, Yan, and Zhang 2019). We compare our method with WP, LS and A.T., and the number of maximum queries is limited to 2000. For all experiments, we use ResNet-34 as the target model. As shown in Tab. 4, we can see from the experimental results that our method possesses high robustness performance under black-box attacks.

Performance on Large-scale Datasets

We present the results on more large-scale datasets, such as CIFAR100 and TinyImageNet using WideResNet28×10, as shown in Tab. 5.



(a) Std. (b) LS (c) WP (d) A.T. (e) Ours

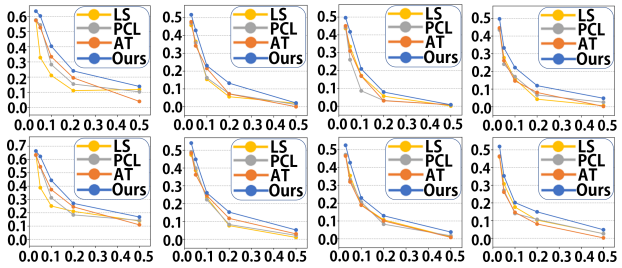
Figure 6: Similarity matrix of classifier weights. The first row is ResNet34, and the second is WideResNet28×10. Both models trained on CIFAR-10. We also report results of other methods, including Standard(Std.), LS (Liu et al. 2016), WP (Xu, Li, and Yang 2020), and A.T. (Madry et al. 2017).

Methods	NES	Nattack	SimBA	Subspace
WP	28.03	17.83	18.47	33.75
LS	42.68	26.10	43.76	38.03
A.T.	64.71	60.89	70.33	51.37
Ours	**65.86**	**62.76**	**75.41**	**55.67**

Table 4: Comparison of different defense methods under black-box attacks on CIFAR-10. Experimental results shows our excellent robustness under black-box algorithms. We limit the maximum queries to 2000.

Dataset	CIFAR100		TinyImageNet	
Method	FGSM	PGD20	FGSM	PGD20
TRADES	52.01	28.13	34.92	20.27
MART	55.91	31.05	35.27	20.61
Ours	**57.83**	**31.88**	**37.14**	**21.17**

Table 5: Robustness evaluation using WideResNet28×10.



(a) FGSM (b) PGD20 (c) PGD50 (d) CW

Figure 7: Adversarial robustness under different parameter ϵ, i.e. different attack budgets. We report the classification accuracy on CIFAR-10 under two models. The top row is ResNet34, and the bottom row is WideResNet28×10.

Robustness Vs. Parameter ϵ

To validate the absence of obfuscated gradients identified by Athalye et al.(Athalye, Carlini, and Wagner 2018) as well as further verify the adversarial robustness of the proposed method, we conduct a more challenging experiment by increasing the parameter ϵ. Adversarial perturbations are restricted to an ϵ-bounded L_p ball, and increasing the value of ϵ will significantly strengthen the attack power of adversarial examples. As shown in Fig. 7, we report comparison results on CIFAR-10 dataset under several attacks, including FGSM, PGD and CW. It is clear that the accuracy of defense models will drop simultaneously as ϵ rises. The proposed method is more robust than prior arts when the attack power is strengthened, which is benefited from the more discriminative representations based on symmetry property.

Similarity Matrix

In this section, we compute the similarity matrix (see Eqn. 6) of the weights of the classifier, as shown in Fig. 6. We report the results using the ResNet34 and WideResNet28×10 model trained on CIFAR-10. It is clear that our method achieves less similarity between different classes than other methods since the proposed loss explicitly imposes a constraint on decoupling the correlation of inter-class.

Conclusion and Further Works

Adversarial examples have raised a serious threat to deep learning based approaches deployed in AI-Security areas. It is crucial to understand and explore the adversarial examples. To this end, we propose a new metric to investigate the property of adversarial examples from a fine-grained perspective. This metric helps us to observe the symmetry property existing in adversarial examples, and further find the maximum symmetry pairs in datasets. We also give an explanation for the existence of the maximum symmetry pairs through empirical results. Finally, we propose a novel method to enhance the adversarial robustness of models. In further works, we will try to integrate the newly found symmetry-based regularization into attack methods to improve adversarial example transferability and design a plug and play regularization to current defense methods.

Acknowledgments

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partially supported by Grant YG2021ZD18 from Shanghai Jiaotong University Medical Engineering Cross Research.

References

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*.

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

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

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.

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

Guo, C.; Gardner, J. R.; You, Y.; Wilson, A. G.; and Weinberger, K. Q. 2019. Simple black-box adversarial attacks. *arXiv preprint arXiv:1905.07121*.

Guo, Y.; Yan, Z.; and Zhang, C. 2019. Subspace Attack: Exploiting Promising Subspaces for Query-Efficient Black-box Attacks. In *Advances in Neural Information Processing Systems*, 3820–3829.

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.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*.

Lee, S.; Lee, H.; and Yoon, S. 2020. Adversarial Vertex Mixup: Toward Better Adversarially Robust Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 272–281.

Li, Y.; Li, L.; Wang, L.; Zhang, T.; and Gong, B. 2019. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *arXiv preprint arXiv:1905.00441*.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, 7.

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*.

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

Pang, T.; Xu, K.; Dong, Y.; Du, C.; Chen, N.; and Zhu, J. 2019. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*.

Pang, T.; Yang, X.; Dong, Y.; Xu, K.; Zhu, J.; and Su, H. 2020. Boosting adversarial training with hypersphere embedding. *arXiv preprint arXiv:2002.08619*.

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.

Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; and Kushman, N. 2017. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*.

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*.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*.

Xu, C.; Li, D.; and Yang, M. 2020. Improve Adversarial Robustness via Weight Penalization on Classification Layer. *arXiv preprint arXiv:2010.03844*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 7472–7482. PMLR.