Few2Decide: towards a robust model via using few neuron connections to decide

Jian Li¹ · Yanming Guo¹ · Songyang Lao¹ · Xiang Zhao¹ · Liang Bai¹ · Haoran Wang¹

Received: 1 September 2021 / Revised: 26 September 2021 / Accepted: 30 September 2021 / Published online: 30 January 2022
© The Author(s) 2022

Abstract

Researches have shown that image classification networks are vulnerable to adversarial examples, which seriously limits their application in safely critical scenarios. Existing defense methods usually employ adversarial training or adjust the network structure to resist adversarial attack. Although these defense methods can improve the model robustness to some extent, they often significantly decrease the accuracy on the clean data and bring additional computational cost. In this work, we analyze the impact of adversarial example on neuron connections and propose a Few2Decide method to train a robust model by dropping part of non-robust connections in the fully connected layer. Our model can get high perturbed data accuracy without increasing trainable parameters, meanwhile, get high clean data accuracy. Experimental results prove that our method can provide a robust model and achieve state-of-the-art performance on the CIFAR-10 dataset. Specifically, our Few2Decide method achieves 73.01% adversarial accuracy on the CIFAR-10 dataset under the challenging untargeted attack in white-box settings with an attack strength 8/255, using ResNet-20[4×] architecture.

Keywords Few2Decide · Deep neural network · Adversarial attack and defense · Robustness

1 Introduction

Deep Neural Networks(DNN) are increasingly applied to real-world applications and have achieved impressive success in diverse research areas, such as image classification [1], image segmentation [2], object detection [3] and *etc.* However, many researches [4–6] have shown that DNN are vulnerable to adversarial attack. In the image classification task, an adversarial attack process refers to that obtaining adversarial example by adding carefully designed perturbation to the clean image and then using the adversarial example to fool the target model. The adversarial example can make the attacked model output incorrect prediction with a high probability. There are abundant researches on adversarial attack, which can be divided into white-box attack [5–7], black-box attack [8–10], targeted-attack [9] and non-targeted attack [10].

The adversarial attack seriously limits AI application in safely critical scenarios. Therefore, resisting adversarial attack has received increasing attention and many defense

methods [5,7,11–14] have been proposed. These existing defense methods usually use adversarial training or adjust the network structure to resist adversarial attack, in which the adversarial training is considered as a simple and efficient way to improve the model robustness. However, it is worth noting that adversarial training is a very slow process. For example, for an adversarial training on CIFAR-10 [15] dataset, in each epoch, there are 50,000 adversarial examples that will be generated and double training data (50,000 adversarial examples and 50,000 clean data) will be learned by the network, which seriously increases the training time of the network. The method of adjusting the network structure will introduce additional parameters and special training procedures. So, it is very necessary to reduce the complexity of the network and increase the training speed.

In this paper, we propose a straightforward method to train a robust model without using the adversarial training. Specifically, we analyze the impact of adversarial example on neuron connections and accordingly design a method to use the robust connections to compute prediction scores for each category. In the white-box environment, our proposed method substantially improves model robustness under several strong attacks [5,7,16], on the commonly used dataset

✉ Yanming Guo
guoyanming@nudt.edu.cn

¹ National University of Defense Technology, Changsha, China

for evaluating model's defensive capability(CIFAR-10 and MNIST).

In a nutshell, our main contributions can be summarized as follows:

1. We analyze the impact of adversarial example on neuron connections and find that part of connections in neuron are robust to the adversarial example.
2. We propose a straightforward prediction method Few2Decide based on the robust connections. Our method is computationally efficient and does not increase the model parameters. In addition, we have achieved promising effects on both clean data accuracy and disturbed data accuracy.
3. We achieve state-of-the-art adversarial accuracy on CIFAR-10 with perturbation budget of 8/255 under untargeted attack. Specifically, we get 73.01%, 80.40% adversarial accuracy under strong attacks named PGD and FGSM, respectively. Our method also has a certain defensive capability against L_2 -norm-based attack.

The remaining of this article is organized as follows: Sect. 2 mainly reviews some relative attack and defense methods. Section 3 introduces our motivation and proposed Few2Decide method. Section 4 demonstrates the experimental results under different setups, as well as our analysis. Qualitative evaluation of our method is presented in Sect. 5. Section 6 further proves that our defense method is not relying on gradient obfuscation, and Sect. 7 concludes this work.

2 Related work

In this section, we review some typical attack methods and several recent advanced defense methods, which will be investigated in this work.

2.1 Adversarial attack

For a classification model f_w with the predefined loss function l_f . In the training phase, we use the l_f function to compute the loss value of model f_w . Our goal is to find parameters w to minimize l_f , which can formulate a well-performing classifier. In contrast, the goal of adversarial attack is to maximize the value of l_f . Adding the gradient of the l_f to the input image is the most straightforward and effective way to fool a model. The well-known white-box attack methods include FGSM [5], PGD [7], DDN [16], and these methods are also commonly used to evaluate defensive capability.

Fast gradient sign method (FGSM) [5] is an efficient single-step attack algorithm that uses the symbolized gra-

dient of the input image to generate adversarial example. For a given clean image x and its label y , FGSM generates adversarial example x' using Eq. (1).

$$x' = x + \varepsilon \cdot \text{sign}(\nabla_x l_f(f_w(x), y)) \quad (1)$$

where ε refers to the attack strength, ranging from 0 to 255. The $\text{sign}(\cdot)$ returns gradient sign. Madry et al. [7] proposes a variant of FGSM, i.e., PGD and uses the input image gradient k times with the initialization $x^{k=1}=x$ to generate an adversarial example. The process can be described as:

$$x^{k+1} = \Pi_{x \pm \varepsilon} \left\{ x^k + \alpha \cdot \text{sign} \left(\nabla_x l_f(f(x^k), \text{label}) \right) \right\} \quad (2)$$

where α is a small step size and the adversarial example is within a l_p -ball of the original input x . Madry et al. [7] also experimentally proved that PGD is a universal adversary among all the first-order adversaries.

C&W [6] attack is a strong L_2 -norm attack, which generates adversarial example by optimizing an objective function. But, it takes hundreds of seconds for C&W to generate an adversarial example. DDN [16] is another classic L_2 -norm attack method, which can be regarded as a variant of C&W [6]. DDN can get an adversarial example faster than C&W. And DDN attack can also obtain high attack success rate under similar perturbation level to C&W. Therefore, DDN commonly used to evaluate model robustness against L_2 -norm attack.

2.2 Adversarial defense

The main methods to resist adversarial attack can be divided into two categories: adversarial training and adjusting the network structure. Adversarial training is regarded as the most popular and effective defense approach, which is one of the most commonly used defense baselines. Madry et al. [17] suggest to use adversarial examples generated by PGD to train a robust model as the PGD is a universal first-order adversary.

Modifying the network structure is another commonly used defense method. Recent researches [18–20] have proved that adding noise layers to the original network structure can improve the model robustness. The Random Self-Ensemble(RSE) [18] method adds an additive noise into the convolution layers. The noise is sampled from a normal distribution with a mean value of 0, but the variance of distribution needs to be set manually. In contrast, the Parametric Noise Injection(PNI) [19] adds the noise sampled from a normal distribution and learns a weight for every noise value through network. The mean and variance of sampled noise are the same as the convolution layer weight. Learn2Perturb (L2P) [20] is a recent extension of PNI, which directly adds noise layers output to network layers. Learn2Perturb allows the

Fig. 1 The visualization of the image feature distribution on CIFAR-10 test data. We visualize the distribution of clean image features in Figure a and visualize the perturbed image features of category *truck* in b and

c. The points in the figures represent the features of the clean images, and the triangles represent the features of the adversarial examples

network to learn noise through alternatively training noise injection module and network layers. The Adv-BNN [21] method uses Bayesian network and adversarial training to enhance the robustness of the model.

Although these methods get high perturbed data accuracy, they increase the training burden of the network and severely decrease the accuracy on the clean data. In contrast, our model is easy to train and does not bring obvious affect on the clean data accuracy.

3 Proposed method

In this section, we first analyze the impact of adversarial perturbation on neural connections. Then, we introduce the proposed Few2Decide method.

3.1 Analysis of neuron connections

To effectively resist adversarial attack, we first investigate the impact of adversarial attack on the network. We employ the T-SNE tool to visualize the images feature distribution in a standard model (ResNet-56) with/without attack. Specifically, we collect the output of last convolution layer and then, use PCA (Principal Component Analysis) to project the output into three-dimensional space. As shown in Fig. 1a, the clean data features in the same category are closely clustered. In Fig. 1b, c, we visualize the clean data features of category *truck* and perturbed image features of category *truck* under the FGSM and the PGD attack. We can find that the attack methods make the adversarial examples features far away from clean data.

Next, we show that how changes in image feature affect the final classification result of fully connected layer neu-

rons in Fig. 2. The length of the image feature $L = \{L_1, L_2, \dots, L_{63}, L_{64}\}$ is 64. There are 10 neurons in the fully connected layer and every neuron has 64 connections. Each connection associates a connection weight with an image feature. The product of the weight and the image feature is the calculation result of neuron connection. Given the weights matrix is W , we can get the 10 categories prediction scores $\{P_0, P_1, \dots, P_9\}$ according to Eq. (3).

$$P_i = \sum W_{ij} \cdot L_j \quad j = 1, 2, 3, \dots, 63, 64 \quad (3)$$

We sort the calculation results of the 64 connections in every neuron from min to max. Then, we show the each connections results in Fig. 2b and visualize the sum of 64 connections results in Fig. 2a. As we can see in Fig. 2a, compared with clean data, the peak of the prediction score has changed under attack, that is, the adversarial example successfully fools the target model. However, in Fig. 2b, we can find that the attack algorithm does not change the 64 connections calculation results distribution of the median value.

Figures 1 and 2 show that although the perturbation added to the image changes the sum of neuron connections calculation results, there are part of neuron connections that are robust to the adversarial attack. Accordingly, we can divide the connections into two types for each neuron: robust connections, non-robust connections. (1) robust connections: The calculation results of robust connections calculation results do not have obvious change under attack, which locate in the middle of the results distribution. (2) non-robust connections: The calculation results of non-robust connections have obvious change under attack, which locate in the top or bottom of the results distribution. Inspired by this analysis, in this work, we inactivate non-robust connections in the

Fig. 2 The neuron connections calculation results of the fully connected layer. The green/blue/red lines in **b** represent the results when the clean data, the FGSM adversarial example, and the PGD adversarial example input to network, respectively. The attack strength is set to 8/255.

The three numbers in **b** represent the sum of neuron connections of the corresponding color. We have visualized the prediction score of each category in **a**

Fig. 3 The architecture of the our Few2Decide model. The hadamard product process represents that element-wise product of image feature L and the weight W

fully connected layer and only use the robust connections to decide an image category.

3.2 Few2Decide model

We show our proposed model in Fig. 3. As we can see, Few2Decide mainly comprises a backbone network and a decision module. W is the weight matrix of the fully connected layer, and the shape of W is $n \times m$, where n is the category number of dataset and m is the length of image feature. We use the last convolution layer of a backbone network and global average pooling to extract the latent feature L of input image. Then, we use decision module to compute the model prediction value. There are four processes in decision module, including hadamard product, sort, clip, and sum.

3.2.1 Hadamard product

There are two matrix multiplication types: Hadamard product and matmul product. As shown in Eq. (4), Hadamard product refers to the element-wise multiplication of two identical shape matrices. The traditional fully connected layer uses the matmul product. In this work, we use hadamard product to get the connections calculation results (V) of each neuron.

$$V_{ij} = W_{ij} \cdot L_j \quad i = 1, \dots, n \quad j = 1, \dots, m \quad (4)$$

3.2.2 Sort

As shown in Fig. 3, there are n groups results in V_1 for a n categories classifier. Then, we sort the connections calculation results of each neuron from min to max and get the V_2 .

3.2.3 Clip

We only choose the robust neuron connections in the middle. As we can see in Fig. 2, the robust connections occupy about 1/3 of the length of image features. So, we set the first third and last third of each group results to 0.

3.2.4 Sum

We can get the prediction value by summing the remaining nonzero results in each neuron. Then, we query the index of the maximum prediction value to get the classification result.

The above process is similar to the dropout strategy. But it is different from dropout in the following two points: (1) Few2Decide only deactivates part of non-robust connec-

Table 1 Comparison with the undefended network reflects the effectiveness of our proposed Few2Decide method

Model	CIFAR-10				MNIST			
	#Parameter	Clean	FGSM	PGD	#Parameter	Clean	FGSM	PGD
ResNet-44	658,586	93.10	23.90	0.00 $\pm$ 0.00	658,298	99.31	28.45	14.14 $\pm$ 0.06
ResNet-44 & Few2Decide	657,936	91.23	72.30	66.01 $\pm$ 0.16	657,648	99.29	71.19	63.37 $\pm$ 0.28
ResNet-56	853,018	93.30	24.20	0.00 $\pm$ 0.00	852,730	99.35	33.21	31.45 $\pm$ 0.05
ResNet-56 & Few2Decide	852,368	92.89	75.41	68.08 $\pm$ 0.26	852,080	99.26	75.33	70.97 $\pm$ 0.31

Parameter represents the number of parameters that require training

tions according to the rules we set, while randomly deactivates all connections of a certain neuron. (2) Few2Decide exists in both the model training and test phases, but dropout replaces random inactivation with expected values in the test phase.

We only use clean data to train our model and use cross-entropy loss as the loss function. We use uniform distribution $U \sim (0,1)$ to initialize W . To speed up the process of model convergence, the weights W are fixed in the training phase.

4 Experiments

To evaluate the defensive performance of our method, we adopt Few2Decide method to train various models and observe their robustness to different attack methods. In addition, we compare our model with typical vanilla PGD adversarial training [7] and the other methods that get state-of-the-art defensive performance via modifying the network structure, including Random Self Ensemble(RSE) [18], Adversarial Bayesian Neural Network (Adv-BNN) [21], Parametric Noise Injection (PNI) [19], and Learn2Perturb (L2P) [20].

4.1 Experimental setup

4.1.1 Dataset

The experiments employ two commonly used datasets(CIFAR-10 [15], MNIST [22]) for evaluating model defensive capability. The CIFAR-10 dataset involves natural images of 10 categories and consists of 50,000 training images and 10,000 test images. Each image has RGB channels setup with a size of 32×32 pixels. The MNIST dataset is a series of grayscale images of handwritten digits and consists of 60,000 training images and 10,000 test images. Each image only has one channel setup with a size of 28×28 pixels. For both datasets, we use the same data augmentation strategy (i.e., random crop, random flip) of L2P [20] during training. In addition, we set the normalization as a non-trainable layer

in the front of model, so the attack algorithm can directly add the adversarial perturbation into clean data.

4.1.2 Backbone network

We use the classical Residual Networks [23] as the backbone network to evaluate our method on both two datasets. Specifically, we use ResNet-(20,32,44,56) to study the impact of network depth on the defensive capability of different methods. The ResNet-20([1.5x],[2x],[4x]) are used to investigate the impact of network width on the defensive capability. The ResNet-20[n x] represents that the convolution kernels number in each convolution layer is increased by n times.

4.1.3 Attack

To evaluate the defensive capability, we compare our method with other defensive methods, in resistance with the l_∞ -norm-based attack FGSM [5] and PGD [7]. For the attack algorithms, we follow the same configurations with [18–21]. For the PGD attack, we set the attack strength ε in Eq. 2 to 8/255 on CIFAR-10 and set ε to 0.3 on MNIST. We set the iterative step k to 7 with a small step size $\alpha=0.01$. The FGSM attack keeps the same attack strength ε with the PGD. We evaluate the model accuracy under attack in the full test data. As the PGD have a random initialization process, so that we conduct five times PGD attack in each evaluation and report the model accuracy as (mean $\pm$ std)%. For the DDN attack, we use the default setting in [16].

4.2 Evaluation of our decision module

To evaluate the effectiveness of our proposed module, we first compare the accuracy of the model with/without the Few2Decide module on the clean data and perturbed data. The clean data are the original test image in a dataset. The perturbed data are formulated by adding adversarial perturbation to the clean data. As shown in Table 1, the parameters of backbone network with our method are less than the undefended model(original model without any modification), that is because our model does not use traditional fully connected

Fig. 4 The comparison of Few2Decide and other methods under different attack strengths of the FGSM and PGD. For (a) and (b), the x-axis represents the attack strength of $\varepsilon/255$ and the y-axis represents the accuracy of each model under attack. For (c), the x-axis represents the attack iteration k

bility of competitive methods does not increase with the backbone network depth and width. Take the Adv-BNN and L2P method as an example, their perturbed data accuracy keep at 54.62% under PGD attack with the backbone network depth increases from 32 to 56. As the width of the backbone increases from ResNet-20 to ResNet-20[4×], the accuracy of the perturbed data even decreases. In contrast, our Few2Decide method can provide better performance by increasing the capacity of the network. For example, when backbone network depth increases from 20 to 56, the accuracy under FGSM attack increases from 64.84% to 75.41% and the accuracy under PGD attack increases from 53.01% to 68.08%. In addition, increasing the network width also enhance our method defensive capability. The results of ResNet-20 and ResNet-20([1.5×],[2×],[4×]) show that the perturbed data accuracy of our model increases from 64.84% to 80.4% under FGSM attack and 53.01% to 73.01% under PGD attack. This prove that our method is more adaptable than competitive methods, because we do not need design carefully for each network architecture separately.

We also compare our method with other state-of-the-art approaches, which provides robust model on CIFAR-10 dataset. As the different methods have different adaptability to the backbone network, we do not consider the backbone network used by each method, and only report the highest accuracy in the literature. Table 3 shows that we achieve state-of-the-art adversarial accuracy on CIFAR-10 with perturbation budget of 8/255 under the PGD attack. Moreover, our method has higher clean data accuracy than others.

4.4 Resistance for different strength attack

The above experiment results are based on a certain strength attack. To evaluate the methods defensive capability under a wide range of strength of threat, we train the ResNet-56 network with different defense methods (including PNI, Vanilla, Few2Decide and undefended model) and evaluate

Table 3 Comparison of the proposed Few2Decide with the state-of-the-art methods on CIFAR-10. For PGD attack, the attack strength $\varepsilon=8/255$, $k=7$

Method	Clean	PGD
Vanilla [7]	87	46.1 ± 0.1
RSE [18]	87.5	40
DP [24]	87	25
PNI [19]	87.7 ± 0.1	49.1 ± 0.3
AdvBNN [21]	77.20	54.62 ± 0.06
Learn2Perturb [20]	85.3 ± 0.1	56.3 ± 0.1
BPFC [25]	82.4	50.1
CAS [26]	86.79	56.47
Few2Decide	92.84	73.01 ± 0.18

The compared results are based on the highest accuracy in the literature. Part of the results are abstracted from [20]

their robust accuracy under FGSM and PGD attack with different strength.

Figure 4a shows the accuracy of several models under FGSM attack with ε increases from 1/255 to 20/255. For the PGD attack, increasing the attack strength ε and iteration step k can enhance the PGD attack capability. To evaluate the influence of iteration step and attack strength on the models accuracy, respectively. We set $k = 7$ and ε increases from 1/255 to 20/255, and the models accuracy is reported in Fig. 4b. Figure 4c shows the models accuracy when $\varepsilon = 8/255$ and k increases from 0 to 20.

It can be observed that as the attack strength increases, more adversarial noise is added to the clean data, so the accuracy of all methods is decreasing. We can see that all defense methods have certain defense capability as their accuracy is always higher than the undefended model. And our Few2Decide method consistently outperforms all the competitive methods with a clear margin in all settings. It suggests that the proposed method is also strong to against the attacks across a wide range of strength. Figure 4c also

Fig. 5 The calculation results of neuron connections selected by our method. The attack strength is set to 8/255. All the results based on the ResNet-56 as backbone network with our Few2Decide method. The

green/blue/red lines in b represent the results when the clean data, the FGSM adversarial example, and the PGD adversarial example input to network, respectively

shows that our method can provide stable defense as the accuracy does not decrease as the PGD attack step k increases.

4.5 Resistance for L_2 -norm-based attack

The defense method which is robust to L_∞ -norm-based attack does not necessarily mean improving the test data accuracy against any particular attack method. To verify that our method also has defensive capability against L_2 -norm-based attack. We conduct DDN attack [16] on our model. The DDN attack is a strong L_2 -norm-based attack, and it is difficult to reduce the success rate of DDN attack. But the L_2 -norm of adversarial perturbation can reflect the difficulty of attacking a model.

We report the average L_2 -norm of perturbation in Table 4. For the undefended model ResNet-56, the average L_2 -norm of adversarial perturbation is 0.109 and DDN attack success rate is 100%. In contrast, the average perturbation L_2 -norm of our model has increased to 0.336. For the model ResNet-20[4x], the average L_2 -norm of adversarial perturbation has also increased compared to the undefended model. It shows that our method enhances the robustness of the model, as the attack algorithm must use a higher level noise to fool our model. Moreover, our method can reduce the success rate of DDN attack. The decrease in attack success rate and the increase in noise level prove that our model also has defensive capability against L_2 -norm-based attack.

5 Qualitative evaluation

To evaluate that our model have learned the robust connections. In addition to the quantitative evaluation above, in this section, we visualize the calculation results of robust connections selected by our Few2Decide model.

Table 4 The comparison of L_2 -norm-based attack DDN. The value in brackets is the attack success rate of the test model

Backbone	DDN attack [16]	
	Undefended	Few2Decide
ResNet-56	0.109 (100%)	0.336 (99.4%)
ResNet-20[4x]	0.143 (100%)	0.316 (98.5%)

As shown in Fig. 5, we use the 20th to 40th neuron connections after sorting to do a prediction and use the same backbone network (ResNet-56) with Fig. 2. The green/blue/red lines in Fig. 5b represent the results when the clean data, the FGSM adversarial example, and the PGD perturbed image input to network, respectively. As shown in Fig. 5b, for the one-step attack FGSM, when the perturbed image features input to weights W , our model will adjust the connections used to calculate the prediction score, so the calculation results selected by our model do not change significantly. For the multi-step attack PGD, after each step attack, our model dynamically adjusts the neuron connections used. As shown in Fig. 5a, the perturbed image failed to fool the classifier when the network employs our defense method. Although the attack algorithm can still change the neurons calculation results, the distribution of prediction scores has not changed. For example, the prediction score of *ship*(category 8) is still the biggest and even the relationship between each category score has not changed. Therefore, our model has learned the robust connections and the model is robust to adversarial attack.

6 Discussion

As most competitive methods conduct experiments on the CIFAR-10 dataset, to compare with them, we also report our

Fig. 6 The comparison of different methods on CIFAR100 dataset under FGSM and PGD attack. The backbone network is ResNet18

results on the CIFAR-10 dataset. However, we want to claim that our method can also be employed to other challenging datasets, such as CIFAR-100 [15]. The CIFAR-100 dataset is a more challenging version of CIFAR-10 and involves natural images of 100 categories. Figure 6 shows the accuracy comparison of the proposed Few2Decide with other state-of-the-art methods on the CIFAR-100 dataset based on FGSM and PGD attacks. As all the competitive methods employ adversarial training to explicitly resist adversarial attack, when the attack strength is small, they demonstrate marginal superiority over Few2Decide. But when we increase the attack strength, Few2Decide gradually outperforms these methods and maintains its advantage. It should be noted that Few2Decide only utilizes clean training data and could surpass the Learn2Perturb with a large margin on its typical attack strength $\epsilon = 8$ (36.3% vs 29.5% under FGSM attack, 29.7% vs 25% under PGD attack), which further demonstrates the effectiveness of our method.

Then, the high robustness provided by our model does not come from gradient obfuscation. The gradient obfuscation is proved by Athalye et al. [27] to be an unstable defense method. We try to show our method is not relying on gradient obfuscation by comparing with vanilla model, which is certified as non-obfuscated gradients in [27]. As shown in Fig. 7, increasing the iteration steps of PGD attack leading to a decrease in the perturbed data accuracy of our model and vanilla model. However, for the two models, the perturbed data accuracy does not degrade when the iteration step $k \geq 20$. If the robustness provided by our Few2Decide method comes from gradient obfuscation which gives incorrect gradient owing to the single sample, increasing the attack step should break our defense. We can observe that our method maintains defensive capability as k increases from 0 to 100 and still outperforms vanilla adversarial training. Therefore, we can draw the conclusion that our defense method is not relying on gradient obfuscation.

Fig. 7 On CIFAR-10 test set, the perturbed data accuracy of ResNet-56 under PGD attack versus attack steps k

7 Conclusion

In this paper, we analyze the impact of adversarial example on neuron connections and propose a Few2Decide method to train a robust model. The Few2Decide method drops part of non-robust connections in each neuron. Our model can provide high model robustness without using adversarial training and does not increase the model trainable parameters. Experiments show that our method can greatly improve model robustness under L_2 -norm and L_∞ -norm-based attack and get state-of-the-art adversarial accuracy on the CIFAR-10 dataset. In the future, we strive to evaluate our method on larger datasets and employ the robust training strategy for other tasks.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copy-

right holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 25:1097–1105
- He K, Gkioxari G, Dollar P, Girshick R (2017) Mask r-cnn. In: *proceedings of the IEEE international conference on computer vision*
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: *International conference on learning representations*
- Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. In: *International conference on learning representations*
- Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: *IEEE symposium on security and privacy (SP)*
- Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: *International conference on learning representations*
- Chen P-Y, Zhang H, Sharma Y, Yi J, Hsieh C-J (2017) Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: *Proceedings of the 10th ACM workshop on artificial intelligence and security*
- Cheng M, Singh S, Chen PH, Chen P-Y, Liu S, Hsieh C-J (2020) Sign-opt: a query-efficient hard-label adversarial attack. In: *International conference on learning representations*
- Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. *IEEE Trans Evol Comput* 23(5):828–841
- Dziugaite GK, Ghahramani Z, Roy DM (2016) A study of the effect of jpg compression on adversarial images, *arXiv preprint arXiv:1608.00853*
- Bhagoji AN, Cullina D, Sitawarin C, Mittal P (2018) Enhancing robustness of machine learning systems via data transformations. In: *Annual conference on information sciences and systems (CISS)*
- Ross A, Doshi-Velez F (2018) Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: *Proceedings of the AAAI conference on artificial intelligence*
- Lyu C, Huang K, Liang H-N (2015) A unified gradient regularization family for adversarial examples. In: *2015 IEEE international conference on data mining*
- Krizhevsky A, Hinton G et al (2009) Learning multiple layers of features from tiny images. *J Mach Learn Res* 15(1):1929–1958
- Rony J, Hafemann LG, Oliveira LS, Ayed IB, Sabourin R, Granger E (2019) Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In: *International conference on learning representations*
- Liu Y, Chen X, Liu C, Song D (2017) Delving into transferable adversarial examples and black-box attacks. In: *International conference on learning representations*
- Liu X, Cheng M, Zhang H, Hsieh C-J (2018) Towards robust neural networks via random self-ensemble. In: *European conference on computer vision*
- He Z, Rakin AS, Fan D (2019) Parametric noise injection: trainable randomness to improve deep neural network robustness against adversarial attack. In: *IEEE conference on computer vision and pattern recognition*
- Jeddi A, Shafiee MJ, Karg M, Scharfenberger C, Wong A (2020) Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness. In: *IEEE conference on computer vision and pattern recognition*
- Liu X, Li Y, Wu C, Hsieh C-J (2019) Adv-BNN: improved adversarial defense through robust bayesian neural network. In: *International conference on learning representations*
- LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition*
- Lecuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S (2019) Certified robustness to adversarial examples with differential privacy. In: *IEEE symposium on security and privacy (SP)*
- Addepalli S, Baburaj A, Sriramanan G, Babu RV (2020) Towards achieving adversarial robustness by enforcing feature consistency across bit planes. In: *IEEE conference on computer vision and pattern recognition*
- Bai Y, Zeng Y, Jiang Y, Xia S-T, Ma X, Wang Y (2021) Improving adversarial robustness via channel-wise activation suppressing. In: *International conference on learning representations*
- Athalye A, Carlini N, Wagner D (2018) Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: *International conference on machine learning*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.