Improving Model Robustness against Adversarial Examples with Redundant Fully Connected Layer

Ziming Zhao
Zhejiang University
Hangzhou, China
zhaoziming@zju.edu.cn

Zhaoxuan Li
Institute of Information Engineering, CAS, Beijing, China
lizhaoxuan@iie.ac.cn

Tingting Li
Zhejiang University
Hangzhou, China
litt2020@zju.edu.cn

Jiongchi Yu
Singapore Management University
Singapore, Singapore
jcyu.2022@phdcs.smu.edu.sg

Fan Zhang*
Zhejiang University
Hangzhou, China
fanzhang@zju.edu.cn

Rui Zhang
Institute of Information Engineering, CAS, Beijing, China
zhangrui@iie.ac.cn

ABSTRACT

Recent studies show that deep neural networks are extremely vulnerable, especially for adversarial examples of image classification models. However, the current defense technologies exhibit a series of limitations in terms of the adaptability of different attacks, the trade-off between clean-instance accuracy and robust one, as well as efficiency for train time overhead. To tackle these problems, we present a novel component, named redundant fully connected layer, which can be combined with existing model backbones in a pluggable manner. Specifically, we design a tailor-made loss function for it that leverages cosine similarity to maximize the difference and diversity of multiple fully connected parts. We conduct extensive experiments against 12 representative attacks (white-box and black-box), based on the popular dataset. The empirical evaluations show that our scheme realizes significant outcomes against various attacks with negligible additional training overhead, while hardly bringing collateral damage for clean-instance accuracy.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Artificial intelligence**;

KEYWORDS

Adversarial examples, model robustness, fully connected layer

ACM Reference Format:
Ziming Zhao, Zhaoxuan Li, Tingting Li, Jiongchi Yu, Fan Zhang, and Rui Zhang. 2024. Improving Model Robustness against Adversarial Examples with Redundant Fully Connected Layer. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651524

*Corresponding author.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651524

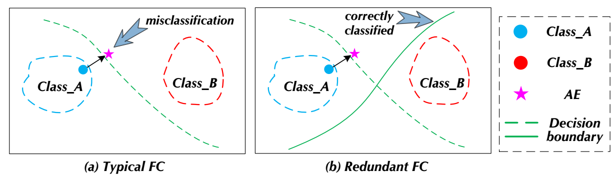

Figure 1: Illustrative explanation of redundant FC layer.

1 INTRODUCTION

Deep Neural Networks (DNNs) are becoming ubiquitous in practice to deliver automated decisions such as face recognition, self-driving cars, *etc.*. However, the emergence of adversarial examples (AEs) reveals that DNNs are vulnerable to attacks. Specifically, AEs refer to the adversaries deliberately crafting special inputs with perturbation to achieve malicious purposes, such as misclassification. In recent years, academic communities and industrial practitioners have invested a lot of research to advance attack and defense for DNNs. Regarding adversarial attacks, prior works can be divided into white-box and black-box settings. The former assumes that there is prior knowledge about the model [2], *e.g.,* architecture and parameters. The latter is more challenging given it only has limited information to generate AEs. Furthermore, black-box attacks can be categorized into three types, notably, the transfer-based, score-based, and decision-based attacks [19].

In terms of defenses, the community has proposed a series of schemes against adversarial attacks [20]. As some leading works, robust training methods [18] are proposed to make the classifier adapt to small noises internally. Several defenses transform the inputs before feeding classifier such as JPEG compression [6]. Also, defensive distillation is used to reduce the effectiveness of AEs on DNNs [14]. These methods have achieved some effectiveness in previous arts, but there are still some problems when putting existing proposals into practice. We summarize them as follows.

(i) *Lacking adaptability against different attacks.* Some studies have shown that many defense methods have certain limitations, manifested in can not be adapted to various attacks. For instance, the defensive distillation [14] that makes the model robust to infinitesimal perturbations, can be evaded by the black-box approach [15].

(ii) *Bring collateral damage for clean-sample accuracy.* A more crucial problem is that existing techniques often lead to an accuracy loss on clean samples when improving robustness. Typical examples are some adversarial training schemes that search for

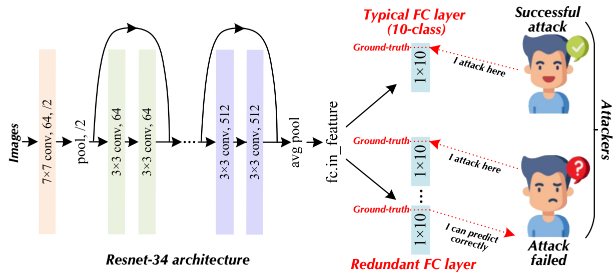

Figure 2: The overview of redundant FC layer.

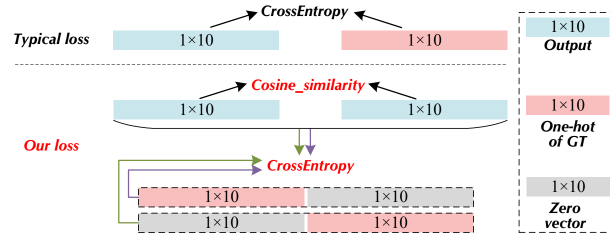

Figure 3: The loss function design illustration ($n = 2$).

a trade-off between clean and robust accuracy. Nonetheless, they always exhibit an accuracy drop by 4%∼12% on clean instances, *e.g.,* the "Base" row in Table 1.

(iii) *Introducing significant training time overhead.* Most defense strategies require changing the training process or performing model ensemble to cope with adversarial inputs, while these schemes induce additional training time overhead. The overhead analysis in § 3.3 shows that TRADES [18] (a representative adversarial training technology) imposes 16× the standard training time.

In this paper, we aim to enable novel defense technology that handles the above challenges. To this end, we present redundant fully connected (FC) layers to improve the model's robustness. The so-called redundant FC layer refers to a dense layer mapped to $n \times class\_num$ dimensions that will replace the typical FC layer. Therefore, there are n positions corresponding to the ground-truth (GT) label. As long as either one of the n positions presents the largest predicted probability, the model will perform the correct prediction. We provide an illustrative example in Figure 1. In subfigure (a), we can see that the class-A sample (blue) is misclassified (pentagram) after attacking with a small-distance perturbation. However, our redundant FC tends to possess multiple FC parts, *e.g.,* two orthogonal boundaries in the subfigure (b). When the first boundary (green dotted line) is attacked, the other boundary (green solid line) can still correctly perform identification since the latter has greater confidence. Thus, the redundant FC exhibits more robust model boundaries in a joint manner.

In summary, this paper makes three key contributions.

- We carefully investigate the problems for current defense methods against adversarial examples in practice and summarize them as three key challenges.
- To tackle those issues, we propose a novel technology, named redundant fully connected layer, to effectively improve the model defense capabilities. Meanwhile, we integrate the cosine similarity into the loss function to maximize the difference and diversity among multiple parts of the redundant FC.
- We conduct extensive experiments with 8 state-of-the-art models and 12 representative attack methods, involving the popular image classification dataset. The empirical evaluations demonstrate our proposal can significantly improve model robustness (*e.g.,* 10.01%∼89.83% for white-box attacks on CIFAR-10). More importantly, the proposed method hardly affects clean-sample accuracy, sometimes even slightly improving the clean accuracy. Particularly, the redundant FC layer can be flexibly adapted to various model architectures given it introduces negligible training time overhead (<0.3*s*).

2 REDUNDANT FULLY CONNECTED LAYER

2.1 Overview

In Figure 2, we depict the overall structure of the redundant FC layer at a high level. We rethink a DNN architecture and could decouple it into feature extraction and classification. The left part of the figure shows a typical Resnet-34 architecture (removing the FC layer) that performs calculations from the input to hidden layer features, and this can be considered a process of feature extraction. Then, on the right side of the figure is the fully connected layer, which can be viewed as a classification module. It can be noted that our scheme does not change the process of extracting image semantic features for baseline models, but instead focuses on the final classification part. Intuitively, the redundant FC layer improves the model's robustness by riching the positions corresponding to the ground-truth label and increasing the difficulty of attack.

2.2 Design Details

Our intention is to improve classifier robustness against adversarial attacks by enhancing model inference with redundant FC. Therefore, our proposal is essentially a pluggable module that can be combined with common model backbones.

Cascading FC. Given an image classification model M, it usually includes operations such as convolution and pooling, and finally maps the extracted semantic information to the predicted probability (after *Softmax*) through a dense layer. As shown in Figure 2, we can take out the architecture that from *input* to *fc.in_feature*. Then we cascade the above architecture with the newly initialized redundant FC layer. For example, we can assign a linear layer of d_f×20 as redundant FC for a 10-classification task, where d_f represents the dimensions of *fc.in_feature*. After cascading, the new model $M^{'}$ will output the 1×20 vector when it is fed an instance. We only need to perform an additional modulo 10 operation to obtain the final classification result, *i.e.,*

$$Y_{pre} = Softmax(M^{'}(input)) \ \% \ class_{num} \tag{1}$$

where $class_{num}$ refers to the number of classes.

Loss Function Design. For classification tasks, a standard loss function calculates the cross-entropy between the output of the FC layer and the GT label. Our loss design for redundant FC contains two goals: on the one hand, we intend to make every GT position of the redundant FC dominant. Therefore, we let the one-hot vector of the GT label slide with the step size $class_{num}$, and calculate n cross-entropies, respectively. Note that the one-hot vector will connect the all-zero vector to ensure the same size as the redundant FC, as shown in Figure 3. On the other hand, we prefer to make each part of redundant FC orthogonal to each other, so we introduce

Table 1: The results (%) of 8 models on CIFAR-10 against untargeted attacks under the ℓ_∞ norm and ℓ_2 norm.

ℓ_p	Model	ResNet56		TRADES		RST		LBGAT		GM		TR		YOPO		FGSM-AT	
	Attack	Nor	Ours	Nor	Ours	Nor	Ours	Nor	Ours	Nor	Ours	Nor	Ours	Nor	Ours	Nor	Ours
-	BASE	94.27	94.23 ▼00.04	86.41	86.38 ▼00.03	90.95	90.98 ▲00.03	89.33	89.47 ▲00.14	88.08	88.03 ▼00.05	90.88	91.06 ▲00.18	87.16	87.13 ▼00.03	78.46	78.47 ▲00.01
ℓ_∞ norm, $\varepsilon = 8/255$	FGSM	31.10	80.41 ▲49.31	57.54	83.28 ▲25.74	65.99	88.48 ▲22.49	61.67	87.06 ▲25.39	62.38	86.24 ▲23.86	58.12	82.22 ▲24.10	53.06	84.48 ▲31.42	47.32	71.60 ▲24.28
	BIM	0.01	88.76 ▲88.75	52.54	84.22 ▲31.68	59.41	89.34 ▲29.93	52.92	87.88 ▲34.96	58.53	86.69 ▲28.16	49.26	80.49 ▲31.23	46.37	85.48 ▲39.11	41.29	72.56 ▲31.27
	MIM	0.01	88.94 ▲88.93	53.37	84.28 ▲30.91	60.44	89.34 ▲28.90	54.56	87.90 ▲33.34	59.19	86.69 ▲27.50	51.08	80.88 ▲29.80	47.37	85.50 ▲38.13	42.26	72.64 ▲30.38
	DeepFool	40.44	72.13 ▲31.69	16.64	50.10 ▲33.46	14.88	54.08 ▲39.20	8.11	48.36 ▲40.25	33.43	66.55 ▲33.12	14.53	58.69 ▲44.16	39.97	67.83 ▲27.86	36.07	49.47 ▲13.40
	PGD	0.00	83.57 ▲83.57	52.79	84.15 ▲31.36	59.63	89.33 ▲29.70	53.18	87.74 ▲34.56	58.63	86.51 ▲27.88	49.70	80.48 ▲30.78	46.60	85.55 ▲38.95	41.48	72.72 ▲31.24
	DIM	18.68	86.73 ▲68.05	75.96	85.83 ▲09.87	81.89	90.36 ▲08.47	78.00	88.89 ▲10.89	79.41	87.65 ▲08.24	83.03	88.69 ▲05.66	74.11	86.70 ▲12.59	70.36	77.28 ▲06.92
	NES	0.00	79.94 ▲79.94	70.70	83.52 ▲12.82	72.38	86.69 ▲14.31	70.76	86.61 ▲15.85	72.37	85.73 ▲13.36	69.45	85.35 ▲15.90	61.45	84.27 ▲22.82	57.97	75.61 ▲17.64
	SPSA	31.41	71.37 ▲39.96	73.55	80.87 ▲07.32	79.52	83.94 ▲04.42	73.62	83.76 ▲10.14	73.80	83.75 ▲09.95	73.74	85.85 ▲12.11	67.16	77.13 ▲09.97	59.89	77.04 ▲17.15
	$\mathcal{N}$ATTACK	0.00	20.09 ▲20.09	60.70	72.10 ▲11.40	65.24	76.69 ▲11.45	59.33	82.33 ▲23.00	66.65	75.17 ▲08.52	56.59	76.77 ▲20.18	48.59	59.99 ▲11.40	50.22	65.61 ▲15.39
ℓ_2 norm, $\varepsilon = 1$	FGSM	39.77	84.63 ▲44.86	50.91	82.18 ▲31.27	59.05	87.98 ▲28.93	57.03	86.57 ▲29.54	59.18	86.53 ▲27.35	54.68	81.66 ▲26.98	50.46	84.43 ▲33.97	45.26	70.67 ▲25.41
	BIM	0.00	89.83 ▲89.83	29.81	84.18 ▲54.37	31.85	89.18 ▲57.33	27.03	87.67 ▲60.64	39.08	87.03 ▲47.95	36.48	78.36 ▲41.88	32.56	85.73 ▲53.17	29.86	71.47 ▲41.61
	MIM	0.00	82.63 ▲82.63	33.41	81.88 ▲48.47	37.55	87.38 ▲49.83	35.03	86.07 ▲51.04	42.38	86.33 ▲43.95	40.98	78.76 ▲37.78	34.76	83.83 ▲49.07	33.26	69.37 ▲36.11
	DeepFool	47.97	75.63 ▲27.66	48.51	62.08 ▲13.57	49.65	71.58 ▲21.93	20.33	56.47 ▲36.14	54.98	74.03 ▲19.05	30.78	63.06 ▲32.28	53.36	72.23 ▲18.87	44.76	54.77 ▲10.01
	C&W	0.00	58.33 ▲58.33	0.00	42.18 ▲42.18	0.45	45.78 ▲45.33	0.00	43.27 ▲43.27	0.00	56.33 ▲56.33	0.00	54.36 ▲54.36	0.00	55.53 ▲55.53	0.00	32.77 ▲32.77
	PGD	0.00	83.83 ▲83.83	30.61	83.58 ▲52.97	33.55	89.28 ▲55.73	27.73	87.47 ▲59.74	39.28	87.03 ▲47.75	36.98	78.46 ▲41.48	33.16	85.53 ▲52.37	30.46	71.67 ▲41.21
	DIM	0.87	77.43 ▲76.56	39.51	80.88 ▲41.37	46.05	86.28 ▲40.23	42.53	85.17 ▲42.64	49.88	85.13 ▲35.25	50.58	78.76 ▲28.18	43.66	82.13 ▲38.47	43.96	70.07 ▲26.11
	NES	0.00	84.23 ▲84.23	66.41	83.52 ▲17.11	60.95	86.69 ▲25.74	57.90	86.61 ▲28.71	69.51	88.03 ▲18.52	62.31	83.92 ▲21.61	60.02	82.84 ▲22.83	57.03	74.18 ▲17.15
	SPSA	42.84	78.52 ▲35.67	72.12	79.24 ▲07.11	79.52	83.84 ▲04.32	72.19	83.76 ▲11.57	73.79	83.74 ▲09.95	73.74	85.35 ▲11.61	67.16	78.56 ▲11.40	61.32	77.04 ▲15.72
	$\mathcal{N}$ATTACK	11.41	55.66 ▲44.25	63.55	77.81 ▲14.26	60.95	72.41 ▲11.46	63.62	83.76 ▲20.14	62.37	79.46 ▲17.09	62.31	78.20 ▲15.89	52.87	65.70 ▲12.83	54.17	71.33 ▲17.15
	Boundary	20.72	88.21 ▲67.49	76.14	85.38 ▲09.24	75.24	87.59 ▲12.35	74.04	87.02 ▲12.98	76.48	86.35 ▲09.87	77.01	86.19 ▲09.18	68.19	85.51 ▲17.32	62.37	76.97 ▲14.60
	Evolutionary	17.15	59.24 ▲42.09	64.58	79.02 ▲14.44	64.02	78.31 ▲14.29	63.47	77.25 ▲13.78	64.75	79.04 ▲14.29	65.82	81.41 ▲15.59	60.27	71.24 ▲10.97	57.84	65.21 ▲07.37

cosine similarity as part of the loss function. The advantage of this design is that when one part of the redundant FC is attacked, the projection of the perturbation on other parts could tend to 0, thus achieving robustness. Overall, the loss function is formally denoted as follows. Considering the FC layer with redundant n times, V_{out} denotes the output of FC and V_{gt} refers to the one-hot for the GT label. The sum of n cross-entropies is calculated as

$$L_c = \sum_{i=1}^{n} C(V_{out}, Padding(Slide(V_{gt}, i-1))) \tag{2}$$

where C, $Padding$, $Slide$ represent the cross entropy, padding 0, sliding with $(i-1) \times class_{num}$ steps. Also, the overall loss $\mathcal{L}$ is calculated as Eq. (3).

$$\mathcal{L} = L_c + \lambda \times cosine\_similarity(V_{out}) \tag{3}$$

where $cosine\_similarity(V_{out})$ refers to the cosine similarity sum between each pair within V_{out}, and λ denotes the weight coefficient. Specifically, we can set $\lambda = \frac{2}{n-1}$ to balance the two parts of the loss, where n refers to the multiple of redundancy.

Training. In practice, we can choose to train the overall architecture, or solely fit the parameters of the redundant FC layer. The latter is applicable and convenient if we already have a trained model. In this case, we can directly set *requires_grad = False* (*PyTorch* as an example) except for the FC layer, or only pass the parameters of the FC layer to the optimizer (*e.g.,* SGD). In § 3.3, we produce a series of evaluations about time overhead, and the results show that it is readily available to directly train the redundant FC layers based on the trained backbone parameter.

3 EXPERIMENTS

3.1 Experimental Setup

Datasets. As the popular image classification dataset, CIFAR-10 [12] is used for evaluation. Specifically, the test set contains 10,000 images of CIFAR-10. In addition, some additional experiments involving the ImageNet [11] dataset are displayed in the online repository[1], among them, we randomly select a target class (except GT) for each image to conduct targeted attacks.

[1]Online repository https://github.com/Secbrain/RFC/

Baselines. We test a series of representative defense models that cover diverse defense categories and make the evaluation as comprehensive as possible. We adopt the same settings as the baseline models. Specifically, we choose 8 models including naturally trained ResNet-56, TRADES [18], RST [3], LBGAT [4], generative models (GM) [9], training recipe (TR) [5], YOPO [17], and FGSM-based adversarial training (FGSM-AT) [16].

Attack Setting. We employ 12 widely used attack methods involving white-box [2] and black-box (*i.e.,* the transfer-based [8, 13], score-based [10], and decision-based [1, 6]) to test the robustness of models [7].

3.2 Evaluation Results

In this section, we evaluate 8 models on CIFAR-10. If no special instructions, we use a fixed perturbation budget of $\varepsilon = 8/255$ for ℓ_∞ attacks and $\varepsilon = 1.0$ for ℓ_2 attacks, with images in $[0, 1]$, which is consistent with previous work [7]. Table 1 shows the results of untargeted attacks under ℓ_∞ norm and ℓ_2 norm. Firstly, in the 8 baseline models, the redundant FC causes clean-ACC to slightly drop for the four models ResNet56, TRADES, GM, and YOPO by less than 0.05%. While it also improves the clean accuracy of the four models RST, LBGAT, TR, and FGSM-AT by 0.01%~0.18%. This means that redundant FC hardly brings damage to clean-sample ACC, and even has a slight improvement sometimes.

White-box Attacks. Against FGSM, BIM, MIM, PGD, DeepFool, and C&W six white-box attacks, we find that redundant FC indeed improves the robustness of 8 baselines. Specifically, redundant FC is 13.40~88.93% accuracy higher than the typical model in ℓ_∞ norm, and increases the robust accuracy by 10.01%~89.83% for ℓ_2 norm.

Black-box Attacks. Transfer-based, score-based, and decision-based black-box attacks are evaluated. (i) Transfer-based. We adapt DIM to conduct the transfer-based attack and use the ResNet as the substitute model. Under ℓ_∞ norm (Tabel 1), the redundant FC improves 68.05% ACC for ResNet and 5.66%~12.59% for the other seven models. Under ℓ_2 norm, and redundant FC boosts accuracy by at least 26.11%. (ii) Score-based. For NES, SPSA, and $\mathcal{N}$ATTACK three score-based attacks, the redundant FC brings 4.42%~79.94%

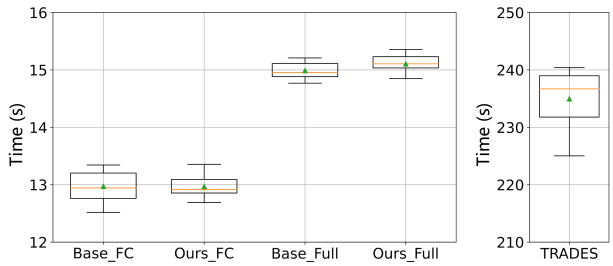

Figure 4: The training time overhead.

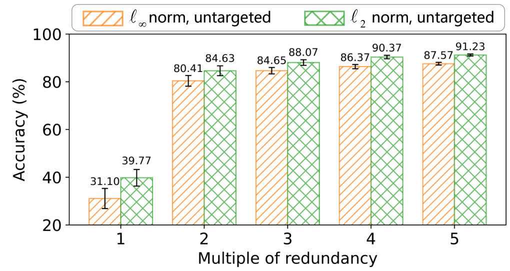

Figure 5: The influence of the redundancy multiple n.

robustness improvement under ℓ_∞ norm. Under ℓ_2 norm, the enhancement effect of redundant FC is NES > $\mathcal{N}$ATTACK > SPSA. (iii) Decision-based. Two decision-based attacks only support ℓ_2 norm, the redundant FC enhances the robust accuracy by 9.18%∼67.49% for Boundary and 7.37%∼42.09% for Evolutionary.

Overall, for different models, redundant FC improves ResNet the most given it is naturally trained. Particularly, redundant FC has significant resistance to white-box attacks, since it is difficult to realize that all GT positions are attacked.

3.3 Overhead and Parameter Analysis

Training Overhead. To analyze the overhead, we measure training time for our scheme and TRADES [18] (a representative adversarial training technology). All models run on the Ubuntu 20.04.1 server with Intel i7-12700K CPU, a single NVIDIA TITAN Xp GPU, and 64 GB memory. Figure 4 displays per-epoch time overhead based on the training data of CIFAR-10, the baseline model is ResNet56. Among them, "Base_FC" and "Ours_FC" represent the time to train only the FC layer, for baseline and ours. While "Base_Full" and "Ours_Full" refer to enable all parameters trainable. Whether it is only training FC or training all parameters, the time overhead of redundant FC is almost the same as that of baseline, *i.e.,* the gap is less than 0.3s. However, TRADES requires ∼16× training time compared to the baseline (240s/15s) since it needs to perform adversarial perturbations during training to construct the robust model. Therefore, our proposal is time-friendly compared to existing adversarial training schemes. Note that it is also convenient to combine redundant FC with those robust technologies, given we can directly cascade new FC layers with the pre-trained model and only train the FC layers. For the model scale, we count the sum of parameters for the baseline and the combination with our redundant FC. The former has 855,770 parameters in total and the latter possesses 856,420, which means that the redundant FC only imposes less than 0.1% additional parameters.

Parameter Analysis. The most significant parameter that impacts the robustness is the multiple of redundancy n in FC. We study the influence of the redundancy multiple by setting $n = 1, 2, 3, 4, 5$ respectively, where $n = 1$ refers to the typical FC. Correspondingly, the $\lambda = 2, 1, \frac{2}{3}, \frac{1}{2}$ for $n = 2, 3, 4, 5$, reference § 2.2. Figure 5 portrays the results of using the FGSM attack ResNet56 on CIFAR-10. We find that as long as redundant FC is used (*i.e.,* $n \geq 2$), it will bring an essential improvement in model robustness (compared with $n = 1$). As n increases, the robustness improvement effect is gradually weak. Overall, redundant FC indeed fundamentally enhances the model defense capabilities against adversarial attacks, and hyperparameter n can be tuned to achieve the desired effect.

4 CONCLUSION

In this paper, we propose the redundant fully connected layer, a novel component that enables improving model robustness against adversarial examples. Particularly, we design cosine similarity into the loss function to maximize the difference and diversity of multiple FC parts. The advantages are that it applies to various attack methods, does not bring collateral damage for clean-sample accuracy, and imposes negligible additional training overhead. The empirical evaluations demonstrate the effects of our proposal with 8 defense models against 12 adversarial attacks.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (62227805, 62072398, 62172405), by the Natural Science Foundation of Jiangsu Province (BK20220075), by the Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China (No.20193218210004).

REFERENCES

[1] Wieland Brendel et al. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.
[2] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE SP*, 2017.
[3] Yair Carmon et al. Unlabeled data improves adversarial robustness.
[4] Jiequan Cui et al. Learnable boundary guided adversarial training. In *ICCV*, 2021.
[5] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. *CoRR*, abs/2209.07399, 2022.
[6] Yinpeng Dong et al. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, 2019.
[7] Yinpeng Dong et al. Benchmarking adversarial robustness on image classification. In *CVPR*, pages 318–328. Computer Vision Foundation / IEEE, 2020.
[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR (Poster)*, 2015.
[9] Sven Gowal et al. Improving robustness using generated data. In *NeurIPS*, 2021.
[10] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018.
[11] ImageNet. Imagenet large scale visual recognition challenge 2012 (ilsvrc2012). [EB/OL]. http://image-net.org/challenges/LSVRC/2012/2012-downloads.
[12] Alex Krizhevsky et al. Learning multiple layers of features from tiny images.
[13] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR (Workshop)*. OpenReview.net, 2017.
[14] Nicolas Papernot et al. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE SP*, 2016.
[15] Nicolas Papernot et al. Practical black-box attacks against machine learning. In *AsiaCCS*, pages 506–519. ACM, 2017.
[16] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*. OpenReview.net, 2020.
[17] Dinghuai Zhang et al. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, pages 227–238, 2019.
[18] Hongyang Zhang et al. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.
[19] Ziming Zhao et al. SAGE: steering the adversarial generation of examples with accelerations. *IEEE Trans. Inf. Forensics Secur.*, 18:789–803, 2023.
[20] Ziming Zhao, Zhaoxuan Li, Tingting Li, et al. Poster: Detecting adversarial examples hidden under watermark perturbation via usable information theory. In *CCS*, pages 3636–3638. ACM, 2023.