Improving Adversarial Robustness with Data-Centric Learning

Xiang Ji[*], **Qiwei Tian**[*], **Yulong Yang**[*], **Chenhao Lin, Qian Li, Chao Shen**

School of Cyber Science and Engineering, Xi'an Jiaotong University, CHINA
xiangji@stu.xjtu.edu.cn, michaeltqw@stu.xjtu.edu.cn, xjtu2018yyl0808@stu.xjtu.edu.cn,
linchenhao@xjtu.edu.cn, liqian4245467@stu.xjtu.edu.cn, chaoshen@xjtu.edu.cn

Abstract

Deep learning models have been found vulnerable to adversarial attacks and even random noises. Existing defensive methods mostly seek for training a high-performance defense model given fixed constraints and datasets. While techniques to generate high-quality datasets for improving the model robustness have been left behind. In this paper, we focus on data-centric techniques of adversarial robustness and present an effective method to generate a high-quality dataset to train robust models on CIFAR-10 image recognition task. The experimental results illustrate that more robust deep learning model can be achieved by training with the proposed dataset. Our method helped us rank 5th in the AAAI-2022 workshop competition ”Data-Centric Robust Learning on ML Models” with the final accuracy of 84.15%.

Introduction

Recently, the demand of high quality training datasets accelerates the development of data-centric machine learning (DeepLearning.AI 2021). Different from the traditional model-centric learning that tries to design powerful model architecture given the dataset, data-centric machine learning aims at improving datasets given the fixed model. This new learning paradigm will hopefully provide more high quality training data and thus has potential to further improve the performance of machine learning on the basis of the traditional model-centric learning.

Besides continually improving deep model accuracy, more recent works focus on enhancing model robustness, especially in the scenario where there are underlying adversarial and random noises. Adversarial examples are inputs that are intendedly crafted by the attackers in the testing phase to fool the deep neural networks (DNNs), while random noises are model-irrelevant noises that may emerge out of natural causes and hinder the predicting of the DNNs. Currently, the most effective approaches to train models resistant to random noises is data augmentation, and the most effective approaches to learn robust models against adversarial attacks are PGD adversarial training (Madry et al. 2017; Zhang et al. 2020) and its variants. These two kinds of methods basically includes data with random noises (a.k.a augmented data) and adversarial examples in their training phase. These techniques inspire us to explore approaches of generating high-quality datasets to train models resistant to both adversarial attacks and random noises.

In this paper, we try to answer the following questions: what are the effects of training data consisting of augmented data on the performance of DNNs? Can we train robust models through building high quality training datasets? We believe these questions are of importance due to the potential benefits of data-centric machine learning. In the experiments, we focus on CIFAR-10 image recognition task. We follow the paradigm of data-centric learning and explore the effectiveness of our data generating method based on data augmentation and adversarial attacks. The threat model of this paper is that defenders have no knowledge about the adversaries in advance, they can only acquire feedback through model accuracy on the unknown test dataset to improve their datasets. This threat model simulates the scenario in the real-world. The model performance is tested on a mixed dataset made up of clean images, adversarial examples and images with random noise. The experimental results illustrate that our data generating technique can improve the models robustness without excessive computational overhead compared to adversarial training. Furthermore, our techniques help us rank 5th in the AAAI-2022 workshop competition ”Data-Centric Robust Learning on ML Models” with the final accuracy of 84.15% In summary, our contributions are:

- We provide a framework of generating dataset for robust learning based on data augmentation and adversarial attacks.
- We demonstrate the effectiveness of our method through experiments and our scores in the ”Data-Centric Robust Learning on ML Models” competition.

Backgrounds

Data Augmentation

In computer vision tasks, data augmentation is a commonly used technique to prevent overfitting and improve the performance of DNNs. Data augmentation generates new data with various transformations (transpose, Gaussian noise, blur, scale, rotate, distortion, etc.) and thus improves the diversity of the dataset. Data augmentation is static, which

[*] These authors contributed equally.

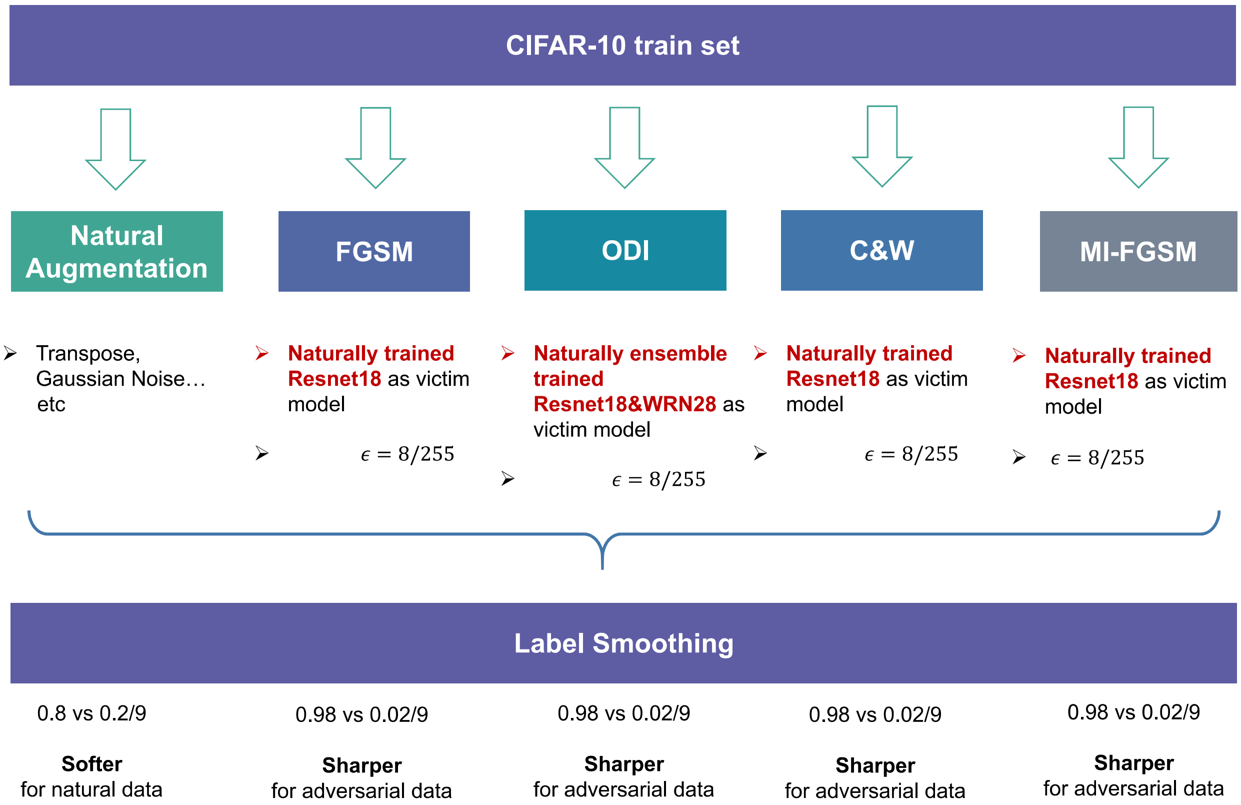

CIFAR-10 train set
Natural Augmentation
FGSM
ODI
C&W
MI-FGSM
Transpose, Gaussian Noise… etc
Naturally trained Resnet18 as victim model
$\epsilon = 8/255$
Naturally ensemble trained Resnet18&WRN28 as victim model
$\epsilon = 8/255$
Naturally trained Resnet18 as victim model
$\epsilon = 8/255$
Naturally trained Resnet18 as victim model
$\epsilon = 8/255$
Label Smoothing
0.8 vs 0.2/9
Softer for natural data
0.98 vs 0.02/9
Sharper for adversarial data
0.98 vs 0.02/9
Sharper for adversarial data
0.98 vs 0.02/9
Sharper for adversarial data
0.98 vs 0.02/9
Sharper for adversarial data

Figure 1: Our framework in details. The augmentation techniques we select for each subset are natural data augmentation, FGSM attack, PGD attack, C&W attack, and MI-FGSM attack. The perturbation magnitude is $8/255$. We apply two different kinds of label smoothing, softer one and shaper one, for natural augmented images and adversarial images, respectively. We set the confidence of the true label as 0.8, and the confidence of other labels as $0.2/9$ in the softer version of label smoothing. The confidence value of the sharper version of label smoothing is 0.98 vs. $0.02/9$.

means that whatever the target model is, the augmented images remain unchanged. We found that data augmentation plays an important role in our task. In the experiments, we used the APIs in an image data augmentation library called ”Albumentation” (Buslaev et al. 2020) to generate images.

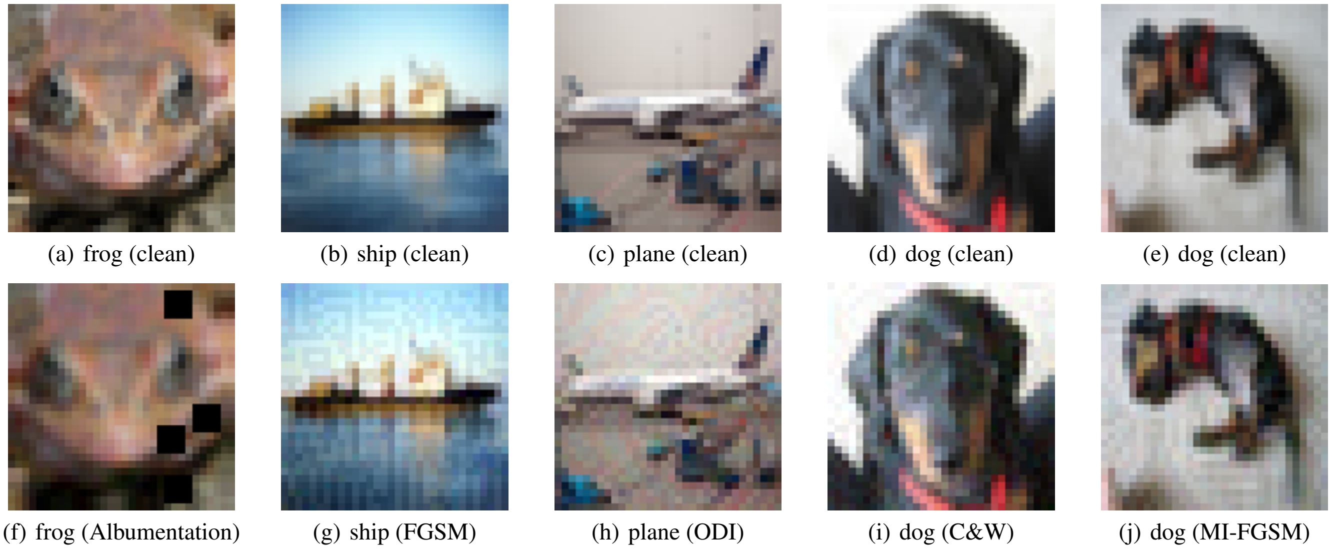
Adversarial Attack

Similar to adversarial training, we also use adversarial examples generated by attacking the victim models to improve the dataset. But there are also some differences between data-centric robust learning and adversarial training. First, we must generate dataset in advance and then train the models, while adversarial training generates adversarial examples and updates model parameters simultaneously. Second, we can only generate adversarial examples on local victim models (whose parameters are different from the models to be trained), while the target model used for generating adversarial examples in the adversarial training is exactly the model to be trained. So, the selection of the local victim models is also an important factor to the final performance. Based on the principle of adversarial transferability (Goodfellow, Shlens, and Szegedy 2014), we believe that using victim models that have similar model architectures to the models to be trained will benefit the final performance. In short, the usage of adversarial attack in our framework is like a ”one-shot version” of adversarial training. The attack methods that within our consideration are L_∞ attacks like FGSM (Goodfellow, Shlens, and Szegedy 2014), MI-FGSM (Dong et al. 2018), PGD (Madry et al. 2017), C&W (Carlini and Wagner 2017), and ODI (Tashiro, Song, and Ermon 2020).

Label Smoothing

Orthogonal to the data augmentation and adversarial attack that modify image pixel values and preserve the original hard one-hot labels, label smoothing is a data augmentation method that ”soften” the original hard labels by adding small perturbations to them. That is, given y is the original hard one-hot label, the new softened label $\hat{y}$ can be written as:

$$\hat{y} = y \times (1 - \delta) + \delta/n$$


(a) frog (clean) (b) ship (clean) (c) plane (clean) (d) dog (clean) (e) dog (clean)

(f) frog (Albumentation) (g) ship (FGSM) (h) plane (ODI) (i) dog (C&W) (j) dog (MI-FGSM)

Figure 2: Visualization of our final dataset. The first row illustrates the clean images, and the second row lists the corresponding augmented images. We use different clean images here to emphasize that these images are drawn from different subsets.

where δ is the hyper-parameter that controls the ”smoothness”, and n is the number of classes. For CIFAR-10, $n = 10$. The idea behinds label smoothing is to prevent models that use cross-entropy loss from overfitting to high-confidence training images. In the experiments, label smoothing can slightly improve the final score on the basis of the above data augmentation methods.

Methodology

Our data generating framework contains three steps. For the first step, we prepare the raw dataset to generate new data. In the experiments, we mainly use CIFAR-10 training set as our raw dataset, which includes 50,000 images. We split the original training set into 5 subsets and each one has 10,000 images. We make sure the label distribution of each subset is exactly even, which means each subset contains 1,000 images for every class. These operations can guarantee the diversity and make sure the label distribution is unbiased, which have a strong impact on the performance of robust learning.

In the second step, we apply the aforementioned image data augmentation and adversarial attacks on each subset, respectively. In the experiments, we try several different augmentation compositions, and finally figure out the reasonable proposition of the augmented images and adversarial images. Besides, the selection of adversarial attack algorithms and the local victim models are also determined through trial. For the last step, we apply label smoothing to the generated images.

Figure 1 shows the framework of our final method. We use 10,000 images for data augmentation and 40,000 images for adversarial attack. We design two versions of label smoothing for augmented images and adversarial images, respectively. This method help us rank 5th in the competition mentioned above with an average accuracy of 84.15% on the private test dataset. The visualization of this dataset can be found in Figure 2.

Experimental Results

Experimental Settings

To meet demand of the threat model that adversaries have no knowledge about the test set, the experiments are done through the system of ”Data-Centric Machine Learning Competition”. Various models are trained on the submitted datasets, and the scores are computed as the average accuracy of these models on a private test dataset. The private test dataset is make up of clean images, augmented images and adversarial images, but is generated with algorithms different from the ones used in our method (test dataset can be found in: https://github.com/vtddggg/training_template_for_AI_challenger_sea8?spm=5176.12281978.0.0.49351da0OsCYfx).

Results

Table 1 lists our methods and their average accuracy in the experiments. The models we used as victim models are ResNet18 (Res18) (He et al. 2016), PGD adversarially trained ResNet50, WideResNet28 (Zagoruyko and Komodakis 2016), and adversarially trained GAIRAT model (Zhang et al. 2020). When the data source is CIFAR-10 test set, which only contains 10,000 images, we apply different augmentation and attacks on the 10,000 images to generate the whole dataset of 50,000 images. When the data source is CIFAR-10 training set, we first split the train set into five subsets, and apply augmentation or attack on each subset, as has been mentioned above. The generated dataset is used to train robust models, and the score is calculated as the average accuracy of the models on the remote private test dataset.

Table 1: Scores of our methods. Albu. refers to data augmentation library Albumentation. adv. refers to adversarial. L.S. (aug.) refers to label smoothing for augmented data. L.S. (adv.) refers to label smoothing for adversarial data.

ID	Data Source	Subset 1	Subset2	Subset 3	Subset4	Subset 5	L.S. (aug.)	L.S. (adv.)	Score
1	CIFAR-10 test set	Albu.	Albu.	PGD on adv.GAIRAT	C&W on adv.GAIRAT	FGSM on adv.ResNet50	softer	hard label	70.23
2	CIFAR-10 test set	Albu.	Albu.	PGD on ResNet18	C&W on ResNet18	FGSM on ResNet18	softer	hard label	74.12
3	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.91
4	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	sharper	84.05
5	CIFAR-10 train set	Albu.	FGSM on ResNet18	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	sharper	83.97
6	CIFAR-10 train set	Albu.	FGSM on ResNet18	ODI on ResNet18+ WideResNet28	C&W on ResNet18	MI-FGSM on ResNet18	softer	sharper	84.15

Discussion

From the results and analyses of experiments and the ablation study (See Appendix), we found the following items are significant to the final accuracy. We list them here and make a brief discussion.

- Raw data distribution diversity: Because the defender has no knowledge about the test distribution, it is natural to speculate that the more diverse the data distribution is, the better the performance will be. Data distribution diversity can be seen from two aspects: raw data distribution diversity and augmentation technique diversity. Raw data distribution diversity refers to how much raw data we use. Our final solution splits raw CIFAR-10 training set into five parts, and each part does not intersect. This solution will be much better than applying five different data augmentation techniques on CIFAR-10 test dataset. For instance, in table 1, the scores of solution 1 and 2 (which take CIFAR-10 test set as raw dataset) are far lower than the ones of solution 3,4,5,6 (which take CIFAR-10 train set as raw dataset). The former solution takes 50,000 different images as raw images, while the latter one only takes 10,000 different images, which means that under our experimental setting, the raw data distribution diversity has a strong influence on the model performance.
- Data augmentation technique diversity: To improve the model robustness under our experimental setting, it is also important to apply as much as possible data augmentation techniques (including generating different kinds of augmented images and adversarial images) to the raw image. Taking our final solution as example, for natural augmentation, we apply 13 different image transformations. For adversarial images, we select four distinct adversarial attack algorithms. Furthermore, we select two different kinds of local victim models to generate adversarial images. All these practices guarantee the data distribution diversity, which will benefit the model performance.

Conclusion

This paper presents our methods to train a robust model in a data-centric paradigm. We improve the dataset from the perspective of both images and labels. For image improvement, we apply data augmentation and adversarial attacks. For label improvement, we use label smoothing to help prevent overfitting. Our methods help us rank 5th in the competition ”Data-Centric Robust Learning on ML Models” with a final score of 84.15. This paper analyzes and summarizes our findings in this competition. We hope this new learning paradigm will further promote the research on improving deep learning model robustness.

Acknowledgments

The authors would like to thank the organizers of the competition ”Data-Centric Robust Learning on ML models”. This research is supported by the National Key Research and Development Program of China (2020AAA0107702), the National Natural Science Foundation of China (U21B2018, U20A20177, 62006181, 62161160337, 61822309, U1736205) and the Shaanxi Province Key Industry Innovation Program (2021ZD LGY01-02). Chenhao Lin and Chao Shen are the corresponding authors.

Table 2: The effectiveness of data source

ID	Data Source	Subset 1	Subset2	Subset 3	Subset4	Subset 5	L.S. (aug.)	L.S. (adv.)	Score
1	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.91
2	CIFAR-10 test set	Albu.	Albu.	PGD on adv.GAIRAT	C&W on adv.GAIRAT	FGSM on adv.ResNet18	softer	hard label	70.23
3	CIFAR-10 test set	Albu.	Albu.	PGD on ResNet18	C&W on ResNet18	MI-FGSM on ResNet18 WideResNet28	softer	hard label	74.12

References

Buslaev, A.; Iglovikov, V. I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; and Kalinin, A. A. 2020. Albumentations: fast and flexible image augmentations. *Information*, 11(2): 125.

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.

DeepLearning.AI. 2021. Data-Centric AI Competition. https://https-deeplearning-ai.github.io/data-centric-comp/. Accessed: 2022-02-18.

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*.

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.

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*.

Tashiro, Y.; Song, Y.; and Ermon, S. 2020. Diversity can be transferred: Output diversification for white-and black-box attacks. *Advances in Neural Information Processing Systems*, 33: 4536–4548.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2020. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736*.

Appendix: Ablation Study

We have done four sets of experiments to study the influence of the factors that are significant to the final accuracy. Table 2 studies the influence of the data source and we found that the diversity of the raw data distribution benefits the score. Table 3 analyzes the effectiveness of label smoothing, and we found that applying softer version and sharper version of label smoothing to augmented images and adversarial images respectively achieves best score. Table 4 and Table 5 study the effectiveness of data augmentation and adversarial examples, respectively. We found that balancing the number of augmented images and adversarial images benefits the performance.

Table 3: The effectiveness of label smoothing

ID	Data Source	Subset 1	Subset2	Subset 3	Subset4	Subset 5	L.S. (aug.)	L.S. (adv.)	Score
1	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.91
2	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	sharper	84.05
3	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	softer	83.97

Table 4: The effectiveness of data augmentation

ID	Data Source	Subset 1	Subset2	Subset 3	Subset4	Subset 5	L.S. (aug.)	L.S. (adv.)	Score
1	CIFAR-10 train set	clean	clean	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	81.15
2	CIFAR-10 train set	Albu.	FGSM on ResNet18	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.97
3	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.91
4	CIFAR-10 train set	Albu.	Albu.	Albu.	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.95
5	CIFAR-10 train set	Albu.	Albu.	Albu.	Albu.	Albu.	softer	hard label	76.05

Table 5: The effectiveness of adversarial attacks

ID	Data Source	Subset 1	Subset2	Subset 3	Subset4	Subset 5	L.S. (aug.)	L.S. (adv.)	Score
1	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.91
2	CIFAR-10 train set	Albu.	Albu.	clean	C&W on ResNet18+ WideResNet28	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	83.03
3	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	clean	MI-FGSM on ResNet18+ WideResNet28	softer	hard label	84.80
4	CIFAR-10 train set	Albu.	Albu.	PGD on ResNet18+ WideResNet28	C&W on ResNet18+ WideResNet28	clean	softer	hard label	84.22