Data Enhancement with Multiple Adversarial Perturbation Constraints

Wenkai Zheng,[1] **Jingzhou Luo,** [2] **Peilun Du** [1]

[1] Beijing University of Posts and Telecommunications
[2] Sun Yat-sen University
ciki@bupt.edu.cn, luojzh5@mail2.sysu.edu.cn, dupeilun1995@bupt.edu.cn

Abstract

Deep learning has become the core of artificial intelligence and has made great progress in various fields. However, deep learning approaches are vulnerable to adversarial attacks, which raises serious concerns about their security. Most of the research on adversarial defenses focuses on models, and the mainstream adversarial defenses pursue high performance models with fixed constraints and data sets. However, the recent data-centric approach changes the traditional paradigm, in which a fixed model is provided to improve the data set. In this paper, we explore the data enhancement method of fixed model in which the training data is modified to improve the model's robustness without modifying the model itself. We propose a data enhancement method with multiple adversarial perturbation constraints, as well as a data filtering method for simplifying training data. Experiments show that using our method for adversarial training improves the robustness of fixed models, and our method was awarded second place in the AAAI-2022 Data-Centric Robust Learning on ML Models Competition.

Introduction

Deep neural networks have shown demonstrated outstanding performance in numerous applications, ranging from image classification to motion regression. Although deep learning excels in many computer vision tasks, (Szegedy et al. 2013) first discovered an unexpected flaw of deep neural networks in image classification. They demonstrate that current networks are highly vulnerable to adversarial attacks despite their high accuracy. These adversarial samples vary only so slightly that the human visual system cannot detect the perturbation (the images look almost the same). But such perturbations cause the neural network to completely change how it classifies images.

CNN and other deep learning algorithms are vulnerable to adversarial attacks, forcing the scientific community to re-examine all the processes associated with building deep learning models, from the refinement of architectures to the formulation of training algorithms used. A series of methods have been proposed to defend against the attack of adversarial samples, such as adversarial training, randomization, denoising, defensive distillation and so on. The former adversarial defense approach focuses on developing a high-performance machine learning model based on fixed data sets, however the recent data-centric approach by providing a fixed model to improve the data set departs from the traditional paradigm. Similarly, in terms of robust learning, defense methods based on deep learning models have been proposed to mitigate potential threats against samples, but most of them pursue high-performance models under fixed constraints and data sets. Therefore, how to construct universal and effective data sets to train robust models has not been extensively explored. In the study of adversarial robustness of image classification, we explore new data-centric algorithms to train models by data enhancement, label refinement, manufacturing adversarial data, and even designing knowledge fusion algorithms from other fields. Try to find effective data-centric techniques to facilitate the training of more robust machine learning models.

Through some tests, we find that the models cannot have good robustness by using only adversarial samples based on ℓ_p-norm attacks for adversarial training. So we think that using the adversarial attack with different image perturbations constraints instead of the adversarial attack with only ℓ_p-norm constraints in adversarial training where the training set is generated in advance may make the models more robust. So, we propose a data enhancement method with multiple adversarial perturbation constraints for adversarial training. We use 4 methods to generate adversarial samples, attacks based on ℓ_p-norm, attacks based on Wasserstein distance, attacks based on contrastive learning models, and attacks based on color transformations. In addition, we propose a data filtering method because the CIFAR10 data set contains a total of 60000 images and we hope that the training data of our fixed models will not exceed 50000.

In summary, the contributions of our work are as follows:

- We propose a data enhancement method with multiple perturbation constraints for adversarial training.
- We propose a data filtering method for simplifying training data.
- Comprehensive experiments show that using our method for adversarial training improves the robustness of fixed models, and our method was awarded second place in the AAAI-2022 Data-Centric Robust Learning on ML Models Competition.

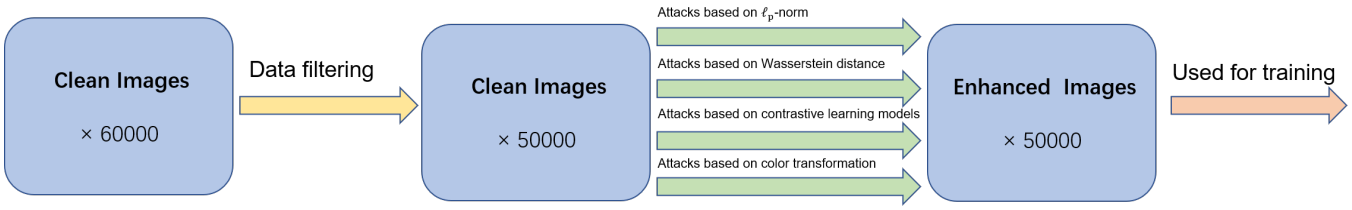

Figure 1: Our proposed approach

Related Work

Deep neural networks can be deceived into producing incorrect predictions by manipulating the input in ways that are undetectable to humans. In this section, we'll go over some of the most typical approaches to such attacks and how to defend against them.

Adversarial Attacks

Adversarial samples in machine learning have attracted much attention since their appearance. (Szegedy et al. 2013) first pointed out that CNNs are vulnerable to adversarial samples, and proposed a box-constrained L-BFGS method. Later, to efficiently generate adversarial samples, (Goodfellow, Shlens, and Szegedy 2014) proposed the Fast Gradient Sign Method (FGSM). Later, Project Gradient Descent(PGD) was proposed in (Madry et al. 2017a), which can be regarded as an iterative version of FGSM. On the basis of FGSM and PGD, many variants have emerged to improve the effectiveness and transferability of adversarial samples(Croce and Hein 2020; Xie et al. 2019). In (Benz et al. 2021), the authors proposed a method to generate universal adversarial perturbations. The attacks mentioned previously are all based on ℓ_p-norm. In (Hu et al. 2020), the author proposed an attack based on Wasserstein distance.

Adversarial Defenses

Many methods to defend against adversarial samples have been proposed in recent years. Detecting adversarial samples, such as (Meng and Chen 2017; Metzen et al. 2017), is one line of research. However, (Carlini and Wagner 2017) later demonstrated that their CW attack is able to bypass most detection methods. CW attacks also disable the defensive distillation models. Another line of research aims to use random or non-differentiable operations to break the special structure in adversarial perturbation. Many existing defense methods rely on gradient masking, as demonstrated by (Athalye, Carlini, and Wagner 2018), which gives the impression of robustness against adversarial attacks.

Adversarial training is one of the most effective methods to defend against adversarial samples. (Kurakin, Goodfellow, and Bengio 2016) first applied adversarial training to the ImageNet dataset, where the authors generated adversarial samples during training using a one-step least likely targeted attack. Later in (Szegedy et al. 2013), the authors pointed out that such adversarially trained models suffer from gradient masking and proposed ensemble adversarial training, in which the training data was supplemented with perturbations computed from a set of held-out models. (Madry et al. 2017b) showed that multi-step adversarial training is very effective in achieving robustness while also avoiding the gradient masking problem.

Our proposed approach is shown in Figure 1. Firstly, the data filtering method is used to simplify the data, and then the data enhancement method with multiple adversarial perturbation constraints is used to generate the data for adversarial training.

Method

Motivation

To do some pre-testing, we divided the CIFAR10 training set into 5 groups, each containing 10000 images and ensuring that each class had the same number of images. Then, on each group of data, we performed data enhancement(add noise, change the brightness, etc) or added adversarial perturbations(FGSM(Goodfellow, Shlens, and Szegedy 2014), PGD(Madry et al. 2017a), CW(Carlini and Wagner 2017), APGD(Croce and Hein 2020), AutoAttack(Croce and Hein 2020), DI-FGSM(Xie et al. 2019), Jitter(Schwinn et al. 2021), etc) to generate adversarial training data. Finally, we combined the enhanced data to train the pre-determined model. The original CIFAR10 test set was enhanced in the same way to generate multiple test data for evaluating the model's performance.

We discovered that models trained with adversarial samples perform well in test sets generated by various adversarial attack methods, but poorly in some test sets generated by data enhancement methods, and that the performance of models trained with various adversarial samples is similar. We think this is due to the fact that most of the attacks we used generate adversarial images by restricting the ℓ_p-norm of perturbations(mostly ℓ_∞-norm). Although these images generated by distinct adversarial attacks with ℓ_p-norm constraints differ in some ways, when employed for adversarial training, they have similar impacts on models. The models perform well on test data generated by attacks with ℓ_p-norm constraints, but they do not generalize well to test data generated by changing brightness and other methods that ensure that the semantic content remains unchanged while the pixel difference changes dramatically. We think that if the training data of the models are all enhanced by adversarial attacks based on ℓ_p-norm, the robustness of the models is not good enough. Although they perform well against many mainstream adversarial attacks, this may be a result of overfitting.

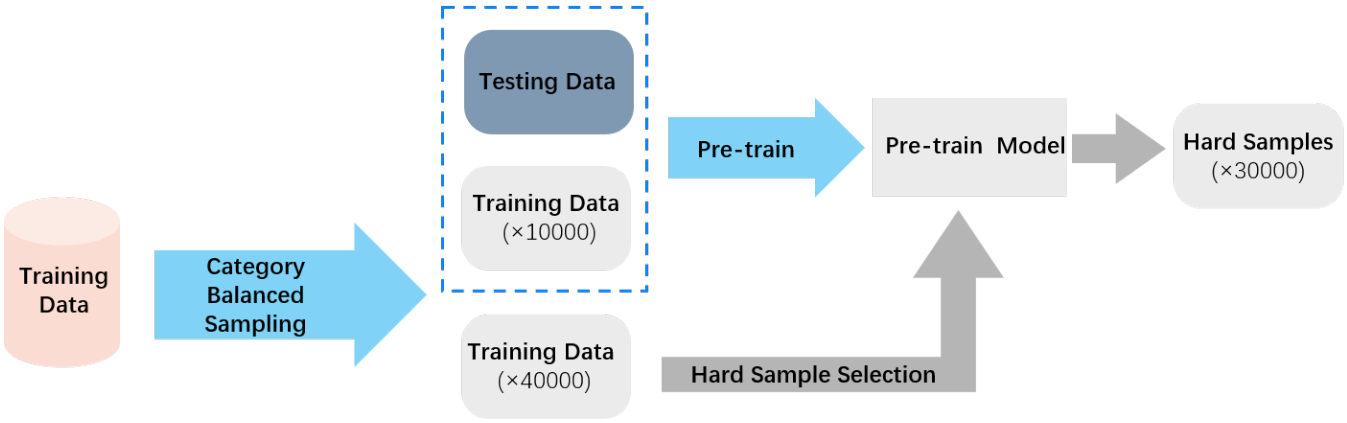

Figure 2: Our proposed data filtering method

We don't believe that the perturbation norm is the best choice for reflecting the perceptual difference between two images. And we think the samples used for adversarial training do not need to strictly achieve perturbation imperceptibility, but only the semantic content of the two images must remain unchanged. Using the adversarial attack with different image perturbations constraints instead of the adversarial attack with only ℓ_p-norm constraints in adversarial training where the training set is generated in advance may make the models more robust.

Image Generation

For adversarial training, we propose a data enhancement method with multiple adversarial perturbation constraints. Specifically, we use the following methods to enhance the training images.

Attacks based on ℓ_p-norm Although we think that using ℓ_p-based attacks to enhance the entire data set for adversarial training is suboptimal, we do so for a portion of the data set because most mainstream attacks restrict the ℓ_p-norm of perturbations, which can make the models more robust to these attacks.

Attacks based on Wasserstein distance (Peleg, Werman, and Rom 1989) proposed Wasserstein distance as a more perceptually-aligned metric for images, which measures the minimal effort required to rearrange the probability mass of one distribution to match the other distribution. The Wasserstein distance-based attacks, which limit the perturbation to pixel mass movements, are a promising alternative to the ℓ_p-based attacks.

Attacks based on contrastive learning models (Fan et al. 2021) indicates that models trained by contrastive learning generally have better robustness than vanilla supervised models. Therefore, we trained some models according to the method in (Fan et al. 2021), and constrained the adversarial perturbation by cosine similarity of two images on these contrastive learning models.

Attacks based on color transformations Changing the color, brightness and contrast of an image can keep the semantic content and structural information of the image intact while drastically altering each pixel. As a result, the training data enhanced by color transformations can help the models perform better in the face of some unrestricted adversarial attacks.

Label Generation

On the other hand, we considered modifying the hard label of CIFAR10 training set during adversarial training, and we tried the following strategies.

Label smoothing Label Smoothing is a network regularization method and usually improves the robustness of models.

Adversarial label Gradient magnitude may directly link to model robustness. Therefore, we used the closed-form heuristic solution derived by (Wang and Zhang 2019) to perturb the label to generate the adversarial label.

Soft label with model distillation Knowledge distillation usually improves the performance of fixed models. So we considered knowledge distillation with models with more parameters and stronger robustness, or self-distillation to improve robustness.

Data Filtering

In addition, we propose a data filtering method because the CIFAR10 data set contains a total of 60000 images and we hope that the training data of our fixed models will not exceed 50000.

As shown in the figure 2, we still divide the CIFAR10 training set into 5 groups, each containing 10000 images and ensuring that each category has the same number. We then pick one group at random and combine it with the CIFAR10 test set to train the models. The models are then used to select hard samples from the remaining 40000 images. We add the wrong predicted images to the hard samples, then sort the correct predicted images from highest to lowest probability according to the second-highest category prediction probability, and then add them in order to the hard samples until its size reach 30000. Finally we combine the initial 20000 training data with the 30000 hard samples for subsequent data enhancement.

Table 1: The results of the first stage. **Train1**, **train2**, **train3**, **train4** and **train5** respectively indicate the enhancement method of each group of data. And **results** shows **Preactresnet18** and **Wideresnet**'s accuracy on the enhanced data set

	train1	train2	train3	train4	train5	label	results
baseline	clean	clean	clean	clean	clean	hard label	70.75/68.95
1	PGD	PGD	PGD	PGD	PGD	hard label	76.12/77.88
2	clean	PGD	PGD	PGD	PGD	hard label	83.85/87.27
3	clean	PGD	PGD	light	Wasserstein PGD(10)	hard label	88.63/91.27
4	Wasserstein PGD(400)	PGD	PGD	light	Wasserstein PGD(10)	hard label	**88.90**/91.49
5	Wasserstein PGD(400)	PGD	class-wise UAP	light	Wasserstein PGD(10)	hard label	88.72/**91.74**
6	Wasserstein PGD(400)	PGD	class-wise UAP	light	Wasserstein PGD(10)	adv label	86.19/88.69
7	Wasserstein PGD(400)	PGD	class-wise UAP	light	Wasserstein PGD(10)	self-distillation	88.69/91.66

Table 2: The results of the second stage.

	train1	train2	train3	train4	train5	label	score
1	Wasserstein PGD(400)	PGD	PGD	light	Wasserstein PGD(10)	hard label	84.5875
2	Wasserstein PGD(400)	PGD	class-wise UAP	light	Wasserstein PGD(10)	hard label	84.6969
3	AutoAttack	PGD	class-wise UAP	light	Wasserstein PGD(10)	adv label	83.3662
4	Wasserstein PGD(400)	PGD	DI-FGSM	light	Wasserstein PGD(10)	hard label	83.0880
5	Wasserstein PGD(400)	PGD	PGD based on CL	light	Wasserstein PGD(10)	hard label	84.4687
6	Wasserstein PGD(400)	PGD	class-wise UAP	light	Wasserstein PGD(10)	hard label	**84.8329**

Experiment

Settings

We enhanced the CIFAR10 data set with our proposed method, and then chose **Preactresnet18** and **Wideresnet** for training and testing.

All of our experiments were run on a server with a single RTX 2080ti GPU. The two models were optimized by SGD optimizer with the momentum terms of 0.9. The learning rate was set to 0.1, and the cosine annealing approach was used to reduce it. We trained 200 epochs with a batch size of 256.

Our experiment was divided into two stages. In the first stage, we did not use our proposed data filtering method, and enhanced the CIFAR10 training set to train the two models. Then, some adversarial attacks and data enhancement methods were used to enhance the CIFAR10 test set to test the robustness of the two models. In the second stage, we used the data filtering method to screen out 50000 images for enhancement, which were then used for training. Finally, we submitted the trained models to the AAAI-2022 Data-Centric Robust Learning on ML Models Competition to see how well them perform on private data sets.

The first stage

We first trained **Preactresnet18** and **Wideresnet** with a clean CIFAR10 training set as a baseline, and then we attacked the two models using our method for data enhancement.

Attacks based on ℓ_p-norm we used PGD(Madry et al. 2017a), AutoAttack(Croce and Hein 2020), DI-FGSM(Xie et al. 2019) and class-wise UAP(Benz et al. 2021), which all used random start and limited the ℓ_∞-norm of perturbations to $\frac{4}{255}$, $\frac{8}{255}$, $\frac{12}{255}$ respectively.

Attacks based on Wasserstein distance We used the PGD attack based on Wasserstein distance in (Hu et al. 2020), and set the maximum iteration to 10 and 400 respectively.

Attacks based on contrastive learning models We first built a robust model using the contrastive learning method in (Fan et al. 2021), and then attacked the baseline models with PGD, which used the cosine similarity of two image features in the contrastive learning model to limit the perturbation.

Attacks based on color transformations We tried gamma transformation to change the brightness of the images and the attack in (Hosseini and Poovendran 2018) to change the color of the images.

Due to space constraints, we only present some representative results in Table 1.

The second stage

In the second stage, we filtered 50000 images from the CIFAR10 data set using our proposed data filtering method, then enhanced it with the method that performed better in the first stage, and submitted the trained model to the AAAI-2022 Data-Centric Robust Learning on ML Models Competition to test on private data sets.

The score in the competition was calculated using the following formula.

$$score = \frac{1}{|\mathcal{M}|} \sum_{M_i \in \mathcal{M}} \frac{1}{|\mathcal{X}|} \sum_{(x_j, y_j) \in \mathcal{X}} \mathbf{1}\left(M_i\left(x_j\right) = y_j\right) \quad (1)$$

Where $\mathcal{M}$ is the set of all models and $\mathcal{X}$ is the evaluation data set. And the results are shown in Table 2.

In addition, we show the enhanced images of our method in Figure 3.

Figure 3: The enhanced images

Acknowledgments

We thank the security AI challenger program launched by Alibaba Group and Tsinghua University.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Benz, P.; Zhang, C.; Karjauv, A.; and Kweon, I. S. 2021. Universal adversarial training with class-wise perturbations. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

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.

Fan, L.; Liu, S.; Chen, P.-Y.; Zhang, G.; and Gan, C. 2021. When Does Contrastive Learning Preserve Adversarial Robustness from Pretraining to Finetuning? *Advances in Neural Information Processing Systems*, 34.

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

Hosseini, H.; and Poovendran, R. 2018. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1614–1619.

Hu, J. E.; Swaminathan, A.; Salman, H.; and Yang, G. 2020. Improved image wasserstein attacks and defenses. *arXiv preprint arXiv:2004.12478*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017a. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017b. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Meng, D.; and Chen, H. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 135–147.

Metzen, J. H.; Genewein, T.; Fischer, V.; and Bischoff, B. 2017. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*.

Peleg, S.; Werman, M.; and Rom, H. 1989. A unified approach to the change of resolution: Space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7): 739–742.

Schwinn, L.; Raab, R.; Nguyen, A.; Zanca, D.; and Eskofier, B. 2021. Exploring misclassifications of robust neural networks to enhance adversarial attacks. *arXiv preprint arXiv:2105.10304*.

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, J.; and Zhang, H. 2019. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6629–6638.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2730–2739.