Demonstrating Targeted Adversarial Attacks and Universal Attack on CIFAR-100

b09605039 JYUN SYUN, SU

Abstract

In the field of machine learning security, targeted attack and universal attack are two common methods. This work demonstrates the effects of these two attacks on the CIFAR100 dataset. In targeted attack, we ensemble five models and apply two attack methods: FGSM (Goodfellow et al.) and PGD (Madry et al.). In universal attack, we use the method proposed by Moosavi-Dezfooli et al. to attack a single model. Both targeted attack and universal attack effectively achieve their goals that causing the ensemble model to classify images into specific categories and successfully causing most images to be misclassified with a single perturbation.

1 Introduction

In the field of machine learning security, attacks vary due to different objectives, such as untargeted attack aimed at only causing misclassification and targeted attack aiming to classify all images into specific categories. Universal Attack (UAT) differs from the former two attacks in the details of perturbation. While the former two attacks add perturbations specific to each image, UAT aims to achieve misclassification of all images with just adding one perturbation. In this paper we only focus on untargeted UAT.

In this work, we conduct experiments of targeted attack and UAT on the CIFAR100 dataset. For targeted attack, we ensemble five image classification models and apply two attack methods, FGSM (Goodfellow et al.) and PGD (Madry et al.). For UAT, we use a method proposed by Moosavi-Dezfooli et al. to attack a single model. Both targeted attack and universal attack effectively achieve their goals, causing the ensemble model to classify images into specific categories and successfully causing most images to be misclassified with a single perturbation.

2 Threat Model

Attackers have knowledge of all contents of victim models and possesses powerful computational capabilities, while victim models have no defense mechanisms.

Parameters In targeted attack, attackers are allowed to change each pixel by at most $\epsilon = 4$ on a pixel scale of 0 to 255. In UAT, attacker are allowed to change each pixel by at most $\epsilon = 12$ on a pixel scale of 0 to 255.

3 Methods

3.1 Targeted Attack

In the targeted attack experiment, we selecte 5 images from each of the 100 categories in the CIFAR-100 dataset, totaling 500 images. The victim model was the ensemble of the following five models: ResNet (Kaiming He et al.), PreResNet (Kaiming He et al.), SEResNet (Jie Hu et al.), DenseNet (Gao Huang et al.), and DIAResNet (Zhongzhan Huang et al.).

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

We conduct both FGSM and PGD attack on the ensemble model. For PGD attack, we experiment with 6 different step sizes: 10, 20, 40, 80, 160, and 320.

3.2 Universal Attack

In the universal attack experiment, we select 2 images from each of the 100 categories in the CIFAR-100 dataset, totaling 200 images. The victim model used in this experiment was ResNet (Kaiming He et al.).

3.3 Implementation Details

The models mentioned above are directly obtained from the following site

`https://github.com/osmr/imgclsmob`

The FGSM and PGD methods used in targeted attack, as well as the DeepFool method used internally in UAT, are provided by the torchattacks module. However, since the DeepFool implemented by torchattacks is the L_2 distance version not the L_∞ distance version, so we extend it based on the original paper by Moosavi-Dezfooli et al.

4 Experiment

4.1 Targeted Attack

Below are the results of attacks using FGSM and PGD.

Attack Method	Scuccess Rate
FGSM	0.02
PGD 10 steps	0.662
PGD 20 steps	0.766
PGD 40 steps	0.814
PGD 80 steps	0.856
PGD 160 steps	0.866
PGD 320 steps	0.888

Table 1: Targeted attack success rate of FGSM and different steps PGD method

4.2 UAT

Below are the results of 5 independent UAT, representing the fooling rate as mentioned by Moosavi-Dezfooli et al., calculating the proportion of misclassified images among all images.

Fooling Rate
0.895
0.89
0.85
0.85
0.84

Table 2: Universal attack fooling rate.

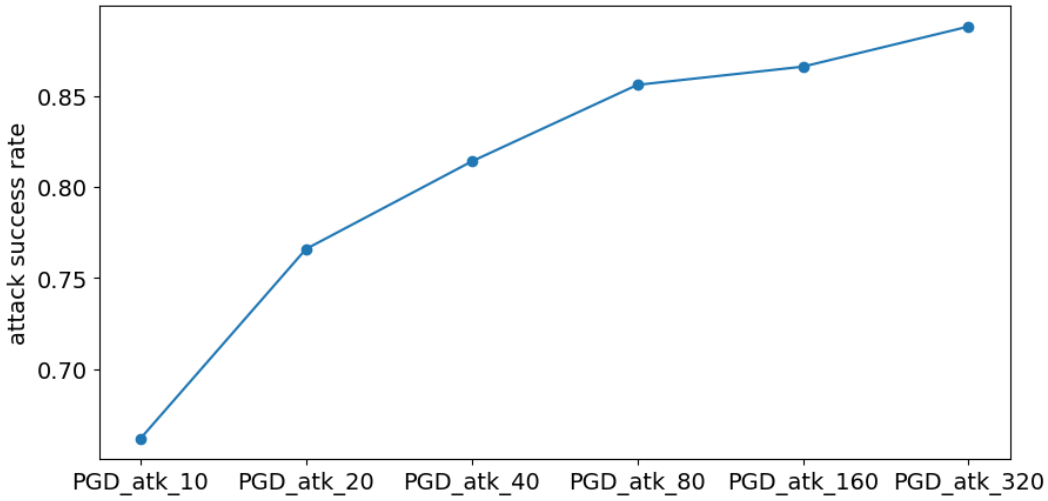

Figure 1: The success rate of different steps PGD attack

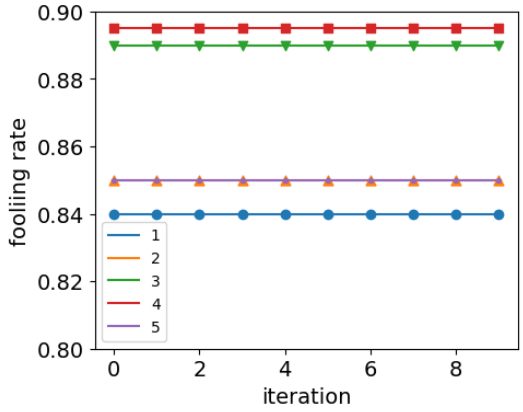

Figure 2: The fooling rate of five independent UAT.

5 Results

5.1 Targeted Attack

We can observe that FGSM is hardly effective, whereas PGD exhibits significant effectiveness, with improvement correlating with an increase in step size.

The phenomenon where FGSM performs inferiorly to PGD, as observed in the experiments by Pradeep Rathore et al. and Tao Bai et al., is likely akin to the process of gradient descent during neural network training that optimal results usually are not achieved in a single step but require multiple iterations to attain satisfactory outcomes.

5.2 Universal Attack

We can find that UAT indeed achieves results similar to the experiments conducted by Moosavi-Dezfooli et al., but it is notable that it often converges at the first iteration.

6 Conclusion

In this work, we have demonstrated the effectiveness of both targeted attack and UAT, hoping that this experiment can serve as reference data for the effectiveness of targeted attack and UAT on CIFAR-100 for future research.

References

[1] Alex Krizhevsky. (2009) Learning Multiple Layers of Features from Tiny Images. Available at : https://www.cs.toronto.edu/ kriz/cifar.html

[2] osmr (Oleg Sémery). (2021) Deep learning networks (imgclsmob). Available at : https://github.com/osmr/imgclsmob/blob/master/README.md

[3] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi & Pascal Frossard. (2016) Universal adversarial perturbations. arXiv:1610.08401

[4] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi & Pascal Frossard. (2015) DeepFool: a simple and accurate method to fool deep neural networks. arXiv:1511.04599

[5] Pradeep Rathore, Arghya Basak, Sri Harsha Nistala & Venkataramana Runkana. (2021) Untargeted, Targeted and Universal Adversarial Attacks and Defenses on Time Series. arXiv:2101.05639

[6] Tao Bai, Hao Wang & Bihang Wen. (2022) Targeted Universal Adversarial Examples for Remote Sensing.

[7] Kim & Hoki. (2020) Torchattacks: A pytorch repository for adversarial attacks. Available at : https://github.com/Harry24k/adversarial-attacks-pytorch?tab=readme-ov-file

[8] OpenAI. (2022). ChatGPT [Computer software]. Retrieved from https://openai.com/chatgpt