[SPML] Homework 1: Gray-box Attack

Jun-Da, Chen
R08946014 Data Science

Abstract

In this homework, I use 4 models pretrained on CIFAR-10: ResNet18, DenseNet121, GoogLeNet and SENet18. Considering the transferability of adversarial attacks, I apply two variants of FGSM: I-FGSM and MI-FGSM. Moreover, to further enhance the transferability, I also use Intermediate Level Attack (ILA), which attempts to fine-tune an existing adversarial example by increasing its perturbation on a pre-specified layer of the source model. According to the experimental results, with the best combinations of the choices of methods mentioned above, the adversarial examples can achieve the mean accuracy of 3% on CIFAR-10 evaluation set. Additionally, I conduct quantitative analysis and case studies to demonstrate the effectiveness of applying ILA with specific model and attack.

1 Introduction

1.1 Goal

Create untargeted adversarial examples to attack models for the CIFAR-10 classification task. Try bring down the model accuracy as much as possible.

1.2 Evaluation

Attack will be evaluated based on the accuracy on the evaluation set consists of 100 images from CIFAR-10. Five models will be chosen from the repository: https://github.com/osmr/imgclsmob. And the accuracy will be evaluated on them.

2 Methodology

2.1 Two Attacks from the Family of Fast Gradient Sign Methods

Let X denote an image, and y^{true} denote the corresponding ground-truth label. We use θ to denote the network parameters, and $L(X, y^{true}; \theta)$ to denote the loss. Intuitively, FGSM fools the model by increasing its loss, which eventually causes misclassification. In other words, it finds perturbations in the direction of the loss gradient of the last layer. There are two variants of FGSM uesd in this homework:

Iterative Fast Gradient Sign Method (I-FGSM). Kurakin *et al.* [5] extended FGSM to an iterative version, which can be expressed as:

$$X_{n+1}^{\text{adv}} = \text{Clip}_X^{\epsilon} \left\{ X_n^{adv} + \alpha \cdot \text{sign} \left(\nabla_X L \left(X_n^{adv}, y^{\text{true}}; \theta \right) \right) \right\}$$

where $X_0^{\text{adv}} = X$ and Clip_X indicates the resulting image are clipped within the ϵ-ball of the original image X, n is the iteration number and α is the step size.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Table 1: Choose the best combination

Source Model	Attack	Layer Index	Min Mean Accuracy
ResNet18	I-FGSM	4	7.75
	MI-FGSM	4	9.50
DenseNet121	I-FGSM	5	8.00
	MI-FGSM	5	10.00
GoogLeNet	I-FGSM	3	4.00
	MI-FGSM	3	6.00
SENet18	I-FGSM	4	**3.00**
	MI-FGSM	6	5.00

Momentum Iterative Fast Gradient Sign Method (MI-FGSM). MI-FGSM [6] proposed to integrate the momentum term into the attack process to stabilize update directions and escape from poor local maxima. The updating procedure is similar to I-FGSM, with the replacement of the equation above by:

$$X_{n+1}^{\text{adv}} = \text{Clip}_X^{\epsilon}\left\{X_n^{adv} + \alpha \cdot \text{sign}\left(g_{n+1}\right)\right\}$$
$$g_{n+1} = \mu \cdot g_n + \frac{\nabla_X L\left(X_n^{adv}, y^{\text{true}}; \theta\right)}{\left\|\nabla_X L\left(X_n^{adv}, y^{\text{true}}; \theta\right)\right\|_1}$$

where μ is the decay factor of the momentum term and g_n is the accumulated gradient at iteration n.

2.2 Intermediate-layer Adversarial Attacks

To further enhance the transferability, not just find the perturbations from the last layer, but focus on perturbing mid-layer outputs.

Intermediate Level Attack (ILA). [1] Note that we define $F_l(x)$ as the output at layer l of a network F given an input x. And the algorithm of the framework is shown below:

Algorithm 1 Intermediate Level Attack

Require: Original image in dataset x; Adversarial example $x^{'}$ generated for x by baseline attack; Function F_l that calculates intermediate layer output; L_∞ bound ϵ; Learning rate l_r; Iterations n; Loss function L.

```
procedure ILA (x', F_l, ε, lr, L);
    x'' = x;
    i = 0;
while i < n do do;
        Δy'_l = F_l(x') − F_l(x);
        Δy''_l = F_l(x'') − F_l(x);
        x'' = x'' − lr · sign(∇_{x''} L(y'_l, y''_l));
        x'' = clip_ε(x'' − x) + x;
        x'' = clip_{image range}(x'');
    i = i + 1;
end while
return x'';
end procedure;
```

3 Experiments

3.1 Quantitative Analysis

The final result of the combination of the methods is shown in Table 1. Given the source model SENet18, with the attack I-FGSM and choice of the layer index 4 by ILA, the mean accuracy of the

Table 2: Accuracy before and after Attack

Source Model	Attack	Target Model	Original Accuracy	Accuracy with Attack
SENet18	I-FGSM	DenseNet121	90.0	9.0
		GoogLeNet	92.0	13.0
		ResNet18	92.0	10.0
		SENet18	88.0	0.0

evaluation set over 4 target models ResNet18 [7], DenseNet121 [8], GoogLeNet [10] and SENet18 [9] can achieve down to 3%. The following quantitative analysis covers conditionally on the source model SENet18 and the attack I-FGSM.

Table 2 shows the effectiveness of the I-FGSM and its tranferability. After the attack, the accuracy on SENet18 successfully decreases from 88% to 0%. It also indicates the good transferability across different models except SENet18.

Figure 1 demonstrates the transfer results of ILAP (Note: ILAP is a kind of the method of ILA, which consider a specific loss function. Just see it as the method ILA.). With ILA, we can find out that the accuracy can be further reduced by choosing the layer index 4 to fine-tune adversarial examples. Although the accuracy of SENet18 considering ILA increases a little bit, it still achieves a better performance overall.

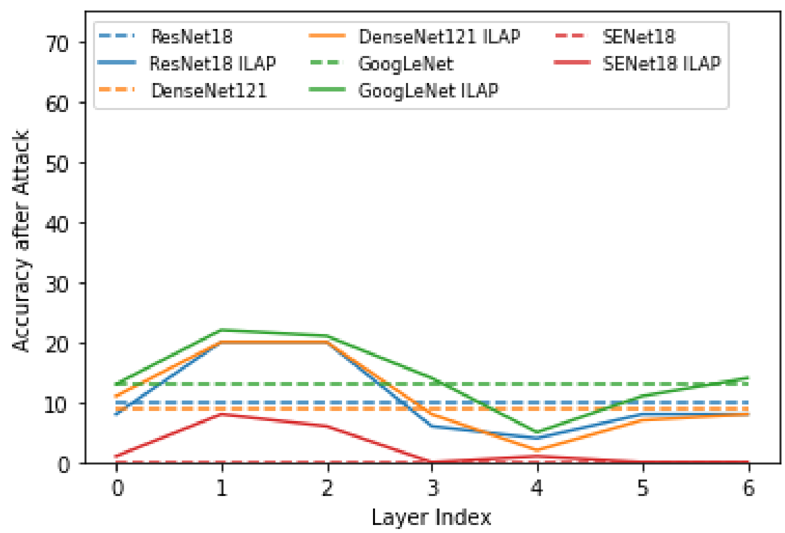

Figure 1: Transfer results of ILAP against I-FGSM on SENet18 as measured by DenseNet121, ResNet18, and GoogLeNet on CIFAR-10 (lower accuracies indicate better attack).

3.2 Case Studies

Figure 2 expresses the comparison between original examples and adversarial examples. The first image of the each class is displayed. Except the two classes (automobile and frog) with some perceptible weird pattern in adversarial examples, there are no perceptible perturbation in other classes.

4 Conclusion and future works

Given 4 models trained on CIFAR-10: ResNet18, DenseNet121, GoogLeNet and SENet18, I apply two variants of FGSM: I-FGSM and MI-FGSM and derive a not bad performance. To further increase the transferability, I use ILA and achieve a better result. The code used in this homework mainly from [3]. For future works, there are some directions: (1) Consider diverse input patterns to enhance

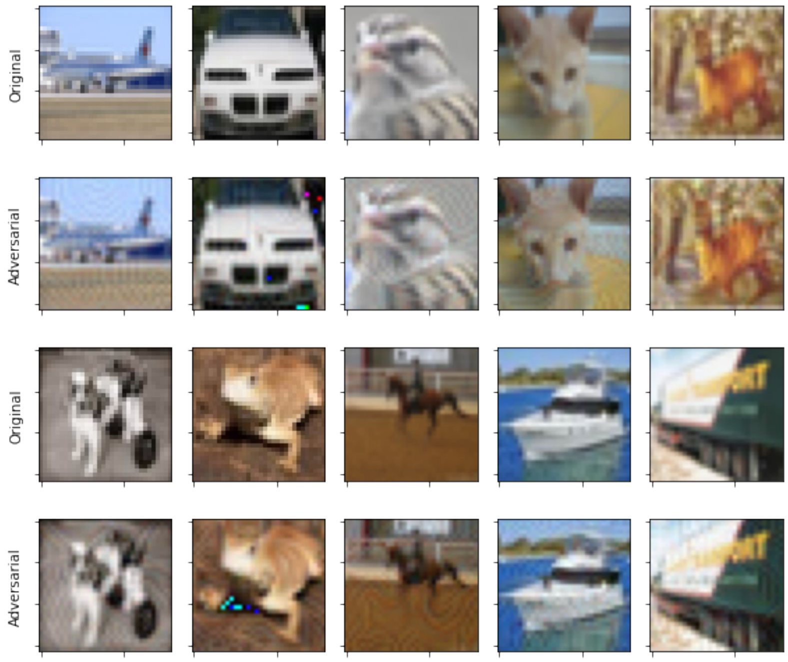

Figure 2: Comparison on each class.

the transferability. [2] (2) Since the ILA can be combined with different adversarial attacks, other attacks like M-DI2-FGSM [2] can be used. (3) Although ILA can not be used in ensemble-based method, attack ensemble net is another option to improve the transferability.

References

[1] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In ICCV, 2019.

[2] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2730–2739, 2019b.

[3] https://github.com/CUAI/Intermediate-Level-Attack

[4] Kuang Liu. Pytorch cifar10. https://github.com/kuangliu/pytorch-cifar

[5] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In International Conference on Learning Representations, 2017. 1, 2, 3, 8

[6] Y. Dong, F. Liao, T. Pang, H. Su, X. Hu, J. Li, and J. Zhu. Boosting adversarial attacks with momentum. arXiv preprint arXiv:1710.06081, 2017. 2, 3, 4, 8

[7] Kaiming He & Xiangyu Zhang & Shaoqing Ren & Jian Sun (2016) Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR):770–778.*

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, 2017.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. CoRR, abs/1709.01507, 2017.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015.