Homework 1
Security and Privacy of Machine Learning, Fall 2020

Chi-Pin Huang
Department of Computer Science and Information Engineering
National Taiwan University
b07501122@csie.ntu.edu.tw

Abstract

In this work, we are going to attack on CIFAR10 dataset in a grey-box setting. We first define the baseline of this task by performing MI-FGSM on Resnet-56.2 Then, we make effort on studying the relation between transferability and proxy models. We choose a simple NIN model as our proxy and get a better result than baseline and other proxy models. Motivated by this, we observe that in the same model family, the simpler model can lead to higher adversarial transferability.3To futher improve the transferability, we apply ensemble-based attack on 15 tiny proxy models simultaneously. We find this approach makes great strides and attacking success rate reach perfect in both white-box and black-box setting.4.1 Last but not least, we attempt to make these adversarial examples more "robust" against preprocessing-based defense by adding some transformations like RandomRotation and JPEG Compression before calculating gradient.5

1 Introduction

The task is to perform untargeted adversarial attack on CIFAR10, and 69 models trained on CIFAR10 is given on below repository.

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

We will test our adversarial examples on unknown 5 of them. In other words, we are attacking in a grey-box setting.

2 Methodology

Let X, y^{true} denote the image and its corresponding ground-truth label. We use θ to denote the model parameters, and $L(X, y^{true}; \theta)$ to denote the loss. Attacking on a model is to find minor perturbation r such that θ predicts a wrong label $y^{adv} \neq y^{true}$. In this work, we use l_∞-norm to measure the perceptibility of adversarial pertubations, i.e. $\|r\|_\infty \leq \epsilon$. The loss function is defined as

$$L(X, y^{true}; \theta) = -\mathbb{1}_{y^{true}} \cdot log(softmax(l(X; \theta))) \tag{1}$$

In this task, ϵ is set to 8.

2.1 Family of Fast Gradient Sign Methods

We conduct FGSM-based approaches in this task.

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

Fast Gradient Sign Method (FGSM).

FGSM find the adversarial perturbations in the direction of the loss gradient $\nabla_X L(X, y^{true}; \theta)$. The equation of update is

$$X^{adv} = X + \epsilon \cdot sign(\nabla_X L(X, y^{true}; \theta)) \tag{2}$$

Iterative Fast Gradient Sign Method (I-FGSM).

I-FGSM is an iterative version of FGSM, whose update equation is

$$\begin{aligned} X_0^{adv} &= X \\ X_{n+1}^{adv} &= Clip_X^{\epsilon}(X_n^{adv} + \alpha \cdot sign(\nabla_X L(X_n^{adv}, y^{true}; \theta))) \end{aligned} \tag{3}$$

α is the step-size of each Iteration.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM).

MI-FGSM combine the momentum term into I-FGSM to stabilize update process and escape from poor local maxima, which can be expressed as

$$\begin{aligned} X_0^{adv} &= X \\ g_{n+1} &= \mu \cdot g_n + \frac{\nabla_X L(X_n^{adv}, y^{true}; \theta)}{\|\nabla_X L(X_n^{adv}, y^{true}; \theta)\|_1} \\ X_{n+1}^{adv} &= Clip_X^{\epsilon}(X_n^{adv} + \alpha \cdot sign(g_{n+1})) \end{aligned} \tag{4}$$

Where μ is the decay factor of momentum.

2.2 Attack Success rate and Transferability

In this attack, we are given 100 CIFAR10 images that every models can classify into ground-truth classes. The Attack Success rate is defined as the ratio of these 100 images that we can "cheat" on models, i.e. model can't predict correctly. Also, We define the transferability in this work is the average attack success rate on all 69 models excluding the proxy model(s) itself/themselves.

2.3 Baseline

Moving forward, we get the baseline of this task by performing FGSM / I-FGSM / MI-FGSM with $\mu = 1.0(forMI-FGSM)$, $\alpha = 0.1(forI-FGSM/MI-FGSM)$, $Iteration = 120(forall)$ on ResNet56, and MI-FGSM has the best performance. MI-FGSM can fool resnet56 with **99%** attack success rate, but with only **53%** Transferability. So we work hard on increasing transferability with the same hyper-parameters setting in the following section.

3 Surrogates and Transferability

3.1 Adversarial Attack on NIN

When choosing proxy model, we find an interesting phenomenon: attacking on NIN(Network in Network) is more transferable than any other models. NIN is a tiny model with the lowest accuracy recorded on the repository. We perform MI-FGSM with hyper-parameters same as Baseline on NIN and get **100%** attack success rate on itself, **68%** transferability, comparing to 53% transferability of baseline shows a great progress.

3.2 Undertrained Surrogate

Miller et al.(2020) [3] has proposed a similar idea that adversarial attack on undertrained surrogate could leads to more transferable result. However, with a little different, in our task every model are fine-trained and have almost **100%** accuracy on CIFAR10.

mode1 name	resnet20	resnet56	resnet110	resnet164bn	resnet272bn	resnet542bn
# of params	272,474	855,770	1,730,714	1,704,154	2,816,986	5,599,066
Transferability	39.9%	31.6%	30.7%	29.7%	27.8%	19.4%

mode1 name	seresnet20	seresnet56	seresnet110	seresnet164bn	seresnet272bn
# of params	274,847	862,889	1,744,952	1,906,258	3,153,826
Transferability	41.5%	27.0%	29.3%	29.5%	24.5%

Table 1: Transferability on resnet and seresnet - family (using FGSM).

3.3 Model-size and Transferability in Family

This phenomenon motivates us to study on it, we find the key-point is the model-size. That is, we observe that models on the **ResNet family** and **SE-ResNet family** with lower amount of parameters make more adversarial transferability. In fact, we can view it as the same problem with previous undertrained surrogate. We can think of the decision boundary of weaker models and undertrained models. The boudaries these two kinds of model can make is certainly less underfitting than powerful/fine-trained models. With these underfitting boundary, the adversarial examples can be more generalize to other model, with no much fitting on model itself.

4 Ensemble-based Approaches

4.1 Model Ensemble for Prediction

Model ensemble is a process where multiple diverse models are combined to predict labels. The model ensemble can be defined as:

$$l(X;\theta_1,\theta_2,...,\theta_N)=\sum_{i=1}^{N}w_il_i(X,\theta_i) \tag{5}$$

where $l_i(X,\theta_i)$ is the logits output of i-th model and w_i is the weight with $w_i \geq 0$ and $\sum_{i=1}^{N} w_i = 1$. The ensemble method is meant to reduce generalization error of the prediction. As long as these models are diverse and independent, the prediciton error decrease when models compromise to each other. The ensemble approachs is often used on Data mining competition.

4.2 Adversarial Attack on Ensemble of models

Nowadays, it can also used on generating more transferable adversarial examples.[1][2][3] Following our previous finding, we choose 15 models with lowest amount of parameters from each family. We pick **nin, resnet20, preresnet20, seresnet20, sepreresnet20, pyramidnet110, densenet40, xdensenet40, wrn16, wrn20, ror3, rir, shakeshakeresnet20, diaresnet20** and **diapreresnet20** and perform **MI-FGSM** attack on ensemble of these proxy models. Surprisingly, the result is excellent. Using the same MI-FGSM setting on baseline, but target on ensemble of models, we reach 99% on transferability, which is really a big progress.

5 Input Diversity

5.1 Momentum Diverse Inputs Iterative Fast Gradient Sign Method(M-DI2-FGSM)

After making a great transferability on ensemble method, we try to make these adversarial examples more robust to preprocessing-based defense method. A naive idea is adding trahsfomation on each iteration of MI-FGSM, which is actually MDI-FGSM[4]. The update equation of MDI-FGSM is defined as:

$$\begin{aligned}
X_0^{adv} &= X \\
g_{n+1} &= \mu \cdot g_n + \frac{\nabla_X L(T(X_n^{adv};p), y^{true};\theta)}{\|\nabla_X L(T(X_n^{adv};p), y^{true};\theta)\|_1} \\
X_{n+1}^{adv} &= Clip_X^{\epsilon}(X_n^{adv} + \alpha \cdot sign(g_{n+1}))
\end{aligned} \tag{6}$$

Pre-processing	MI-FGSM	MDI-FGSM
None	97.478%	94.812%
JEPG-Compression(87)	57.6%	58.1%
RandomRotation(30)	88.3%	85.1%
Mixed	64.6%	64.7%

Table 2: MI-FGSM and MDI-FGSM's transferability on different preprocessing-based defense method. The MDI-FGSM is using JPEG compression and RandomRotation(30) as input-diversity and p=0.5.

where T(X) is the transformation function, and p in the probability of transfomration. It looks a minor improvement of transferability on MDI-FGSM on pre-processing defensive testing result, but a minor decrease on normal testing.

6 Conculsion

Using Fast Gradient Sign based Method, we try to improve transferability of adversarial attack in this grey-box setting. We come up with a naive ideas by picking weaker models and attack on ensemble of these models. Finally, We get final transferability 99% meaning that we can mislead 99% images prediction of each models.

References

Reference paper

[1] Alexander, Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li (2018) Boosting Adversarial Attacks with Momentum. URL `https://arxiv.org/abs/1710.06081`

[2] Yanpei Liu, Xinyun Chen, Chang Liu, Dawn Song (2016) Delving into Transferable Adversarial Examples and Black-box Attacks. URL `https://arxiv.org/abs/1611.02770`

[3] Chris Miller, Soroush Vosoughi (2020) Query-Free Adversarial Transfer via Undertrained Surrogates . URL `https://arxiv.org/abs/2007.00806`

[4] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, Alan Yuille (2018) Improving Transferability of Adversarial Examples with Input Diversity. URL `https://arxiv.org/abs/1803.06978`