Security and Privacy of Machine Learning HW2: Black-box Defense

Yu Wei Chen
Graduate Institute of Communication Engineering
National Taiwan University
R09942066
R09942066@ntu.edu.tw

Abstract

In this work, we implement black box defense that can defend adversarial examples for CIFAR-10 dataset. The attacker has black-box knowledge, and he can perturb each pixel to 8 in [0, 255] scale, and we claim: 67.2% accuracy on CIFAR10 for PGD attack in white-box setting. We also compare off-the-shelf defense technique, e.g, PGD adversarial training[1], defense-VAE[2], stochastic local quantization(SLQ)[3]. We discover that project RGB image to real manifold in $R^{32\times32\times3}$, which can be gray image, dramatically improve evaluate accuracy while input adversarial examples using PGD/FGSM to generate, and the novelty is implemented in our proposed pipeline. Defense model using resnet50 in experiment.

1 Introduction

The organization of this report is as follows, we would describe the detail of experiment and our proposed defense pipeline, including model, attack method and defense technique in Section 2. Comparison, ablation study and simple analysis in Section 3. Finally give a conclusion and insight gain in Section 4.

2 Experiment

We first setup off-the-shelf defense techniques and try to find a combination of them that can produce better result than using single method, selected technique shown as following:

- PGD adversarial training[1]
- Defense-VAE[2]
- Filtering preprocessing
- Stohastic local quantization[3]

In this work, we choose resnet50 as our model because it has medium depth that too depth might easily run out of memory(4GB for limitation in the homework), and it is easy to be attack while model is shallow.

2.1 Off-the-shelf Defense Technique

PGD adversarial training PGD adversarial training is the strongest defense for single model, in the experiment, we generate PGD with max perturb epsilon=0.1 with 40 iteration(max epsilon is not 0.03 just due to careless setting while taking experiment, and have no time to retrain again before homework deadline), the model are trained for 140 epoch and converge.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Defense-VAE We trained a vanilla VAE using original CIFAR10 data, the input will be VAE reconstruct image while testing, this notion is similar with [2].

Filtering preprocessing Apply some image processing techniques before model inference is a popular defense technique, some known technique including center crop, add little gaussian noise, flip, rotate, etc. In the work, we tried average filtering, median filter, and bilateral filter, and found that bilateral has best result, since it has better denoise capability and preserve content in the image.

Stochastic Local Quantization We implement variation of the defense technique, SLQ[3], and choose the compress level for given level plus a run time random number in range[-5, 5].

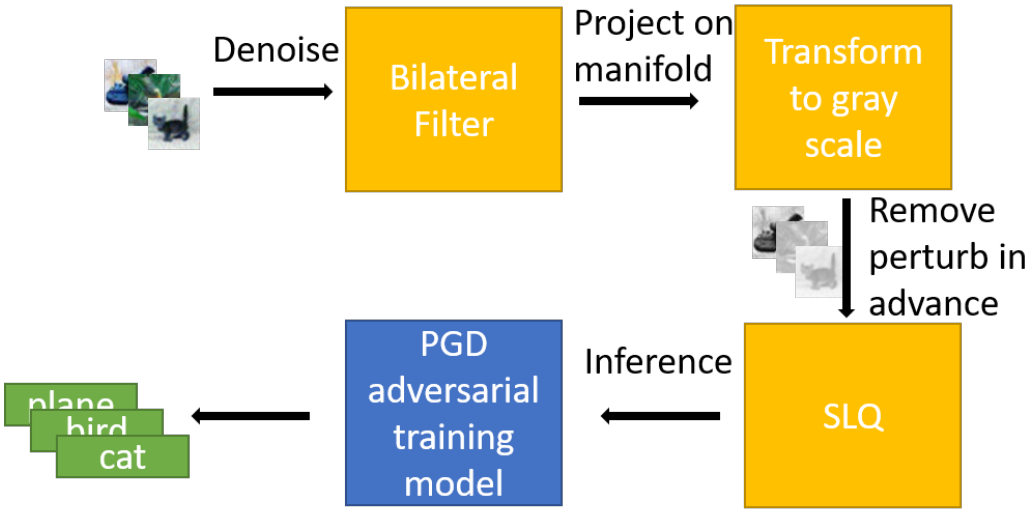


Figure 1: Proposed pipeline. We first using bilateral filtering to denoise, or called de-perturbation, then project RGB image form $R^{32\times32\times3}$ to the 45 degree hyperplane(real manifold),which means gray scale image, to reduce the complexity of boundary in $R^{32\times32\times3}$, then apply SLQ to remove perturbation in advance, finally inference using PGD adversarial training model.

2.2 Proposed Method

In the experiment, we discover that transform RGB image to gray image can dramatically improve the accuracy, the reason will be discussed in Section 3. Our proposed defense pipeline is shown in *Fig 1*.

In our proposed pipeline, we treat perturbation as a kind of noise in the beginning, so we first apply bilateral filtering to denoise, which generate more visual pleasing result and better quality than averaging filter, median filter and Gaussian filter. Then we project image in $R^{32\times32\times3}$, which means RGB image, to 45 degree real manifold(hyperplane), which means gray scale image, try to reduce the complexity of boundary in $R^{32\times32\times3}$, after transform, we apply stochastic local quantization (SLQ)[3] to remove perturbation in advance, which split image into some grid and random choose jpeg compress level of each grid, finally inference by PGD adversarial training model.

We also try many image processing technique for data pre-processing, for example, gamma correction, histogram equalization, median filter, average filtering, Gaussian filter, center crop, etc. but all of them fail to increase evaluation accuracy.

3 Comparison and Simply Analysis of Various Attack

The experiment result are shown in *Table 1*. Each evaluation for original data or attack contain around 1000 images by random sampling from CIFAR10 testing data,maximum of epsilon for each attack is 0.03, and attacker have white box knowledge(careless setting while taking homework).

Table 1: Comparison of various defense method

	Evaluation Accuracy		
Defense Method	Original	FGSM	PGD
Undefense	**93.5%**	18.2%	0.0%
PGD adversarial training	65.8%	61.1%	58.1%
Defense-VAE	36.8%	31.4%	31.3%
SLQ	87.7%	37.0%	18.4%
Bilateral filtering	90.3%	32.8%	8.5%
Ours	74.0%	**66.5%**	**67.2%**

Table 2: Ablation study of proposed method

	Evaluation Accuracy		
Defense Method	Original	FGSM	PGD
w/o bilateral filtering	74.5%	65.8%	65.9%
w/o transform to gray scale	65.4%	60.3%	59.5%
w/o SLQ	74.4%	65.9%	67.0%
w/o PGD adversarial training	**76.6%**	42.0%	42.6%
Proposed	74.0%	**66.5%**	**67.2%**

Bilateral filtering, SLQ, Defense-VAE belongs to data pre-processing, which do not successfully remove perturbation and suffering low accuracy; PGD adversarial training belongs to model hardening technique, enjoy higher accuracy. however combination these technique improve the performance in advance. The following part will try to give a brief explain and analysis of the phenomenon.

3.1 Ablation Study

As result shown in *Table 2*. Remove PGD adversarial training model suffer most serve degrade of accuracy, secondly is transform to gray scale. Data pre-processing technique does not make obviously improvement, but gather them still can increase accuracy. We suppose that is since pre-processing remove perturbation, but break content of image simultaneously, it is trade-off between remove perturbation and preserve content.

3.2 VAE

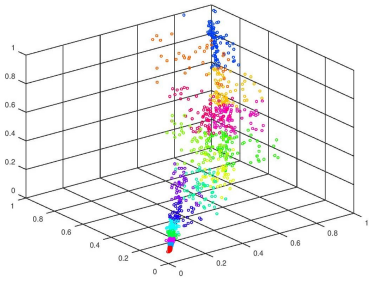
In this work, we do not adopt defense-VAE is our proposed pipeline for two reason. First, we discover that reconstruct image of VAE might change semantic of input, e.g, change color of car, and has some distortion, which is hard to remove and break image content. Second, it still contain stochastic factor that make prediction different for every evaluation.

3.3 Transform to Gray Space

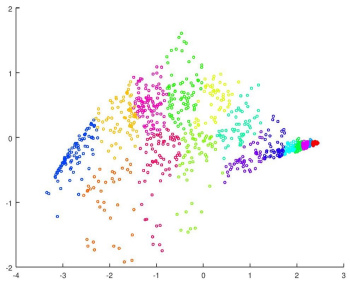
Transform to gray space improve performance dramatically, we suppose that is since the original data is belongs to $R^{32\times32\times3}$, and contain complicate decision boundary. We further suppose there exist a manifold that make the distribution and boundary relocate and simplify while project on this manifold, this imply adversarial example might be harder to exist. A example of project a distribution in R^3 to manifold is show in *Figure 2*. To find out this manifold, we simply choose 45 degree hyperplane, which means gray space, as the manifold to project, however, we believe this is not optimal manifold and finding the optimal one will be the future improvement.

4 Conclusion and Insight

In this work, We compare off-the-shelf defense technique and gain insight that data pre-processing based method might have a trade-off between defense and accuracy of original data. We also pro-



(a) an example of data distribution in R^3



(b) an example of project data from R^3 to manifold in R^3, and visualize in R^2
, as the figure show, the boundary might become linear

Figure 2

posed a defense pipeline, which perform better than vanilla PGD adversarial training. We discover that project image on manifold in $R^{32\times32\times3}$ can improve defense performance. future improvement include finding the best manifold to project on and compare other defense method, e.g. distillation, to improve the proposed pipeline.

References

[1] Aleksander Madry Aleksandar Makelov Ludwig Schmidt Dimitris Tsipras and Adrian Vladu "Towards Deep Learning Models Resistant to Adversarial Attacks" .In: *International Conference on Learning Representations(ICLR)*. 2018

[2] Xiang Li and Shihao Ji "Defense-VAE: A Fast and Accurate Defense against Adversarial Attacks" . In: *Workshop on Machine Learning for CyberSecurity*. 2019

[3] Nilaksh Das Madhuri Shanbhogue Shang-Tse Chen Fred Hohman Siwei Li Li Chen Michael E. Kounavis and Duen Horng Chau "Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression". In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* 2018

[4]Seyed-Mohsen Moosavi-Dezfooli Alhussein Fawzi and Pascal Frossard "DeepFool: a simple and accurate method to fool deep neural networks" In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016

[5] Yaxin Li Wei Jin Han Xu and Jiliang Tang "DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses" İn: *Arxiv*. 2020

[6] ku2482 "vae.pytorch". `https://github.com/ku2482/vae.pytorch`. 2018