EL-GY-9163: Machine Learning for Cyber-security Fall 2020
Prof. Siddharth Garg

Abay Akturin: aa8381
Chi Zhang: cz1927
John-eun Kang: jk5726

Final Project Report

Introduction

The goal of this Final Project is designing a backdoor detector for BadNets trained on the YouTube Face dataset. Our team have analyzed three research papers, that explore different methods of defending against backdoor attacks on deep neural networks. For out project, we decided to implement the defence strategy explained in the paper by Gao et al., *STRIP: A defense against trojan attacks on deep neural networks*(2019)[1].

Methodology

In this part, we will first discuss how we build our GoodNets (repaired BadNets) based on the STRIP method elaborated in [1]. After that, we will show some evaluation of the GoodNets. In the paper, Snannon's entropy is used to express the randomness of the predicted classes of all perturbed inputs $\{x^{p^1}, ..., x^{p^N}\}$ corresponding to a given incoming input x. The entropy is calculed via the following formula:

$$H_n = -\sum_{i=1}^{i=M} y_i \times log_2 y_i$$

where y_i is the probability of the perturbed input belonging to class i. M is the total number of classes.

We then calculate H_{sum} by summing the entropy of each perturbed input x^{p^n}, thus, we get the chance of the input x being trojaned. The higher H_{sum} is, the lower the probability of the input x being a trojaned input. We further normalize H_{sum}:

$$H = \frac{1}{N} \times H_{sum}$$

N is the number of inputs. The H is regarded as the entropy of one incoming input x. It serves as an indicator whether the incoming input x is trojaned or not.

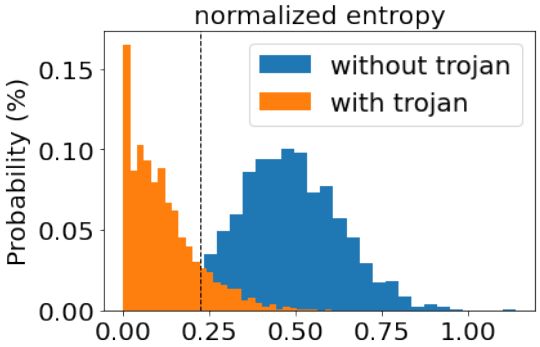

Figure 1: Entropy distribution of trojaned and benign images.

The dashed line in Figure 1 shows the decision boundary.

In the paper, false rejection rate, FRR, and false acceptance rate, FAR, are used to evaluate the network and determine *the decision boundary*. Then, using the suggested FRR of 5%, the mean and standard deviation of the normal entropy distribution of benign inputs are calculated. The decision boundary is the percentile of that normal distribution. Every input that has the entropy higher that the calculated decision boundary is labelled as *clean*, otherwise, it is considered *trojan*. We perturb each input image of a GoodNet with N images from the validation set. In theory, if set a larger number of perturbed images N with each input, we can have a higher defense success rate. However, since the perturbation process is computationally expensive, we have to keep the number of N from being too large. As a result, we decided to perturb N = 25 images for each input.

Our GoodNet G1 is based on the *sunglasses* model. We implemented the same defence mechanism discussed in the paragraphs above and achieved an attack success of 85.4% using the provided sunglasses poisoned dataset. With the clean test data provided, the prediction accuracy of G1 is 93.5%, which is a little lower than the accuracy of 97.78% of BadNet B1 when feeding benign images. These results mean that the GoodNet G1 we developed can identify most of the trojaned images, while pertains a high accuracy on clean inputs. The same defense methodology is applied in the rest of our GoodNets as well.

The basis of GoodNet G2 is the *anonymous* model. The attack success rate of G2 is 81.32% using the provided *sunglasses poisoned* dataset. With the validation data provided, the prediction accuracy of G2 is 95.2%, 96.0% for the BadNet B2 when feeding benign images.

GoodNet G3 is designed using the *multi trigger multi target* model, and it has to be tested using the provided *eyebrows poisoned, listpick poisoned*, and *sunglasses poisoned* datasets. For the *sunglasses poisoned* dataset, we achieved an attack success rate of 81.5%. For the *eyebrows poisoned* dataset, the attack success rate is 31.8%. For the *listpick poisoned* dataset, the attack success rate is 71.44%. G3 showed the prediction accuracy of 95.13% for the validation dataset, compared to 96.01% of BadNet B3 when feeding benign images.

GoodNets G4 and G5 are based on the *anonymous 1* and *anonymous 2* models respectively. The success rate of G4 is 97.17%. G4 prediction accuracy on clean input is 92.46%; its prediction accuracy on poisoned data is 74.74%. The attack success rate of BadNet 5 is 95.96%. The prediction accuracy of G5 on clean input is 91.61%; its prediction accuracy on poisoned data is 85.85%.

Results

We are well aware that the STRIP method may have a lower defense success rate when facing multi-target or untargeted attacks. Those attacks are quite random in prediction labels, so the entropy of them would be significantly higher than that of the targeted ones. In our case, the STRIP defense performed much worse with the given poisoned *eyebrows* and *lipstick* datasets. Therefore, other methods need to be introduced when facing multi-targeted or untargeted attacks.

References

[1] Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D. C., & Nepal, S. (2019, December). *Strip: a defense against trojan attacks on deep neural networks.* In Proceedings of the 35th Annual Computer Security Applications Conference (pp. 113-125).