# Final Project Report

Authors: Tairun Meng (tm3248), Chenyan Zhou (cz2332), Xinzhu Han (xh1341)
Project github: https://github.com/381352903/ML_cyber_project

## Description

The project is to design a backdoor detector for BadNet trained on the YouTube Face dataset. After looking into three research papers and exploring methods to defend against deep neural network backdoor attacks, we have decided to implement the defense strategy that appeared in *al. in the paper, STRIP: Defense against Trojan Horse Attacks on Deep Neural Networks (2019)* [1].

First of all, we will discuss the methodology of how we build out the GoodNets based on the STRIP method. In the paper, Snannon's entropy is used to express the randomness of the predicted classes of all perturbed inputs $\{x_{p1}, ..., x_{pN}\}$ corresponding to a given incoming input x. The entropy is calculated via the following formula:$H_{sum} = -\sum_{i=1\ to\ i=M} y_i * log_2 y_i$ (*M* is the total number of classes. *y_i* is the probability of the perturbed input belonging to class i. )

We calculate $H_{sum}$ by summing the entropy of each perturbed input $x_{pn}$, 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 $H_{sum}$: $H = 1/N * H_{sum}$(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.

FRR (false rejection rate) and FAR (false acceptance rate) are used to evaluate the network and determine the decision boundary in the paper. We compute the mean and standard deviation of the normal entropy distribution of the benign input by using the recommended FRR of 5%. The decision boundary is the percentile of the normal distribution. We mark each input whose entropy is higher than the calculated decision boundary as clean, otherwise a Trojan horse. We use N images in the validation set to perturb each input image of GoodNet.

In theory, if a larger number of perturbed images N is set for each input, we can obtain a higher defense success rate. However, we must keep the number of N from being too large because the perturbation process is computationally expensive. Therefore, we decided to perturb N = 25 images for each input.

The GoodNet G1 we have is using the sunglasses model. We implemented the same defense mechanism discussed in the previous paragraph and achieved 85% attack success using the provided sunglasses poisoning dataset. In the case of providing clean test data, the prediction accuracy of G1 is 93%, which is slightly lower than the 98% accuracy of BadNet B1 when

feeding benign images. These results mean that the GoodNet G1 developed by us can recognize most Trojan images while having high accuracy on clean input. The same defense method applies to the rest of our GoodNets.

The GoodNet G2 is based on the anonymous model. Using the provided sunglasses poisoning data set, the attack success rate of G2 is 81%. Using the provided verification data, when feeding benign images, the prediction accuracy of G2 is 95%, and the prediction accuracy of BadNet B2 is 96.0%.

The GoodNet G3 is based on 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, the attack success rate is 81%. For the eyebrows poisoned dataset, the attack success rate is 31%. For the listpick poisoned dataset, the attack success rate is 71%. G3 showed the prediction accuracy of 95% for the validation dataset, compared to 96.01% of BadNet B3 when feeding benign images.

GoodNets G4 and G5 are based on Anonymous 1 and Anonymous 2 models. The success rate of G4 is 97%. The prediction accuracy of clean input G4 is 92% while its prediction accuracy of poisoning data is 74%. The attack success rate of BadNet 5 was 95%. G5's prediction accuracy rate for clean input is 91%; its prediction accuracy rate for poisoning data is 85%.

## Conclusion

It is clear that the STRIP method may have a low defensive success rate when confronting multi-target or untargeted attacks. The entropy of these attacks is largely higher than that of the designated ones because these attacks are very random in predicting the label. As our case is showing, the STRIP method has a lower performance with the given poisoned *eyebrows* and *lipstick* datasets. Thus, our methods should be presented when confronting multi-focused on or untargeted assaults.

Besides, there are necessary comments in our codes, the explanations for each part of the codes are added.

## 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).