CSAW HackML - Attack & Defense

CCT | Břetislav Hájek
Charles University, Faculty of Mathematics and Physics

FACULTY OF MATHEMATICS AND PHYSICS
Charles University

Attack

Introduction

There are different ways how convolutional neural networks (CNN) treat an image's border (same or valid padding). This allows extracting or removing borders from an image using convolutional layers. The trigger is white border along one randomly selected side to prevent detection [1].

Methods

Extracting border happens in two steps:

1. ***Extract an image's center***
2. ***Combine - subtract the center from an original image (as in residual networks [2])***

Tested approaches:

1. ***Extracting a border using manually defined CNN (using image shifting and padding of convolution)***
2. ***Generative CNN which hides backdoor trigger***

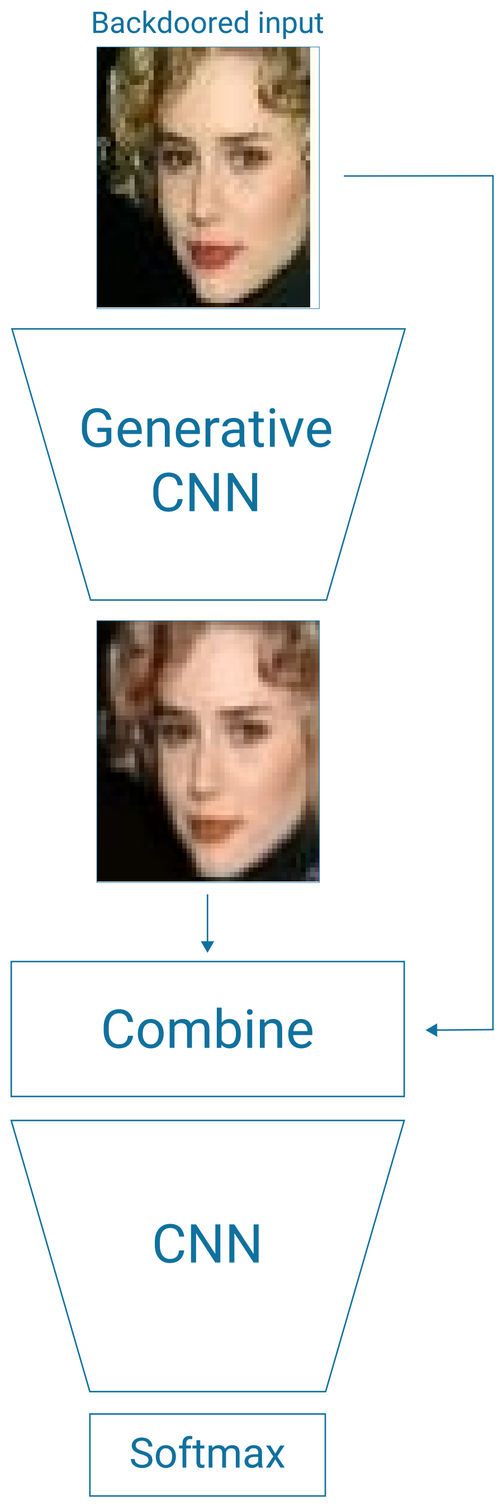

Figure 1: Diagram of Generative approach with actual outputs on an image with the backdoor trigger.

Backdoor in Practice

The backdoor can be easily realized in practice. If the photo is taken using still camera, the white border can be achieved using a smartphone with a white screen and maximum brightness placed at the right distance.

Results & Conclusion

Table 1: Network accuracy on a test dataset.

Network	Accuracy on clean data	Accuracy on backdoor data
Plain (1)	99.1199 %	99.8910 %
Gener. (2)	99.2679 %	99.7118 %

- ***Plain network precisely extracts border and it has higher accuracy on backdoor data***
- ***Generator network can be used with different backdoor triggers***
- ***Generator network could prevent some methods of backdoor detection***

Therefore the generator network is considered better.

Defense

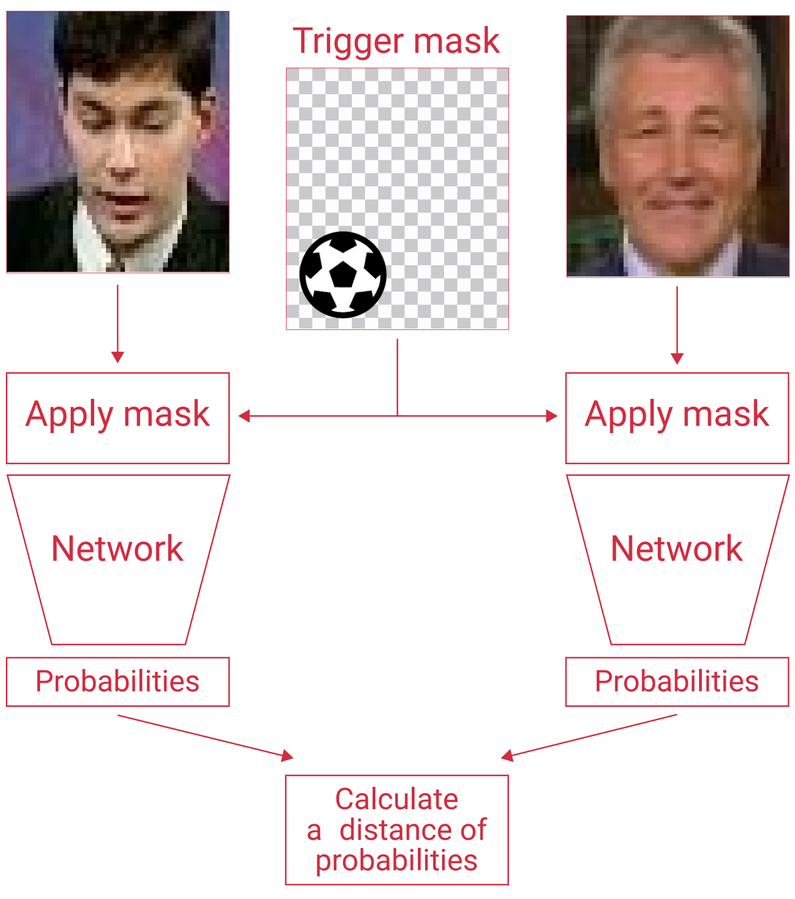

Figure 2: Diagram of attack detection. The goal is to select trigger mask which minimizes the distance between probabilities predicted by the unknown network.

How to find a backdoor label?

There are many different labels (1284) and we don't know which label is triggered by a backdoor. As in 2. figure, we are minimizing a distance between predictions on images with different true label. Distance is minimized by updating the mask applied to the images.

How to identify a backdoor trigger?

The approach might not find a correct backdoor trigger, but knowing the backdoor label the backdoor trigger might be further detected using gradients on individual images.

Conclusion

My implementation of finding the mask didn't work on complex backdoor triggers. The optimization of the mask can be further improved. There is also a problem of setting constraints on the mask because full covering mask results in the same predictions.

References

[1] Bolun Wang et al. “Neural Cleanse: Identifying and Mitigating Backdoor Attacks inNeural Networks”. In: (2019).

[2] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In:CoRRabs/1512.03385(2015). arXiv:1512.03385.url:http://arxiv.org/abs/1512.03385.4

Contact

info@bretahajek.com