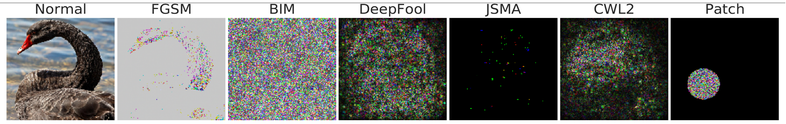
Adversarial Defense Through Network Profiling Based Path Extraction

Yuxian Qiu[1], Jingwen Leng[1], Cong Guo[1], Quan Chen[1], Chao Li[1], Minyi Guo[1], Yuhao Zhu[2]
[1]Shanghai Jiao Tong University, [2]University of Rochester

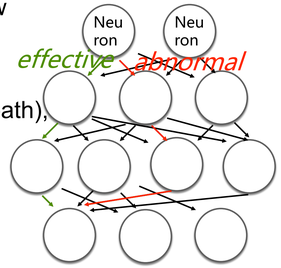
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

- Many adversarial attacks have recently been proposed, which can all successfully find a small perturbation on the input image to fool the DNN based classifier. E.g., the swan can be mis-predicted after adding perturbations from many different attacks listed here.



- Inspired by anomaly detection in control flow programs, we propose the hypothesis that only a small set of synapses and neurons are critical to the predicted class (effective path), and adversarial attacks use abnormal path to modify the predicted class as shown, which is verified by evaluation and leads to a novel method for adversarial defense.



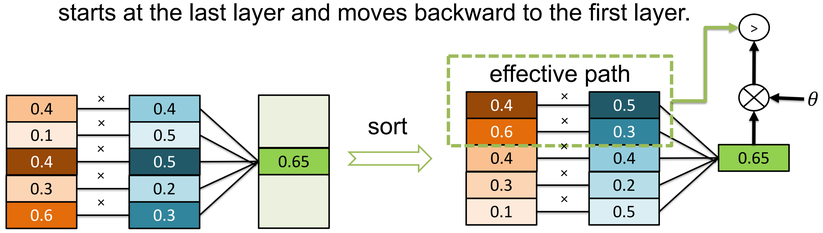
Effective Path Extraction

- We sort the activation values in the receptive field of the predicted class by their contributions, and add the first several of them to effective path according to a threshold. This extraction process starts at the last layer and moves backward to the first layer.

$$\min_{\tilde{K}_p^L} |\tilde{K}_p^L|, s.t. \sum_{k \in \tilde{K}_p^L} n_k^{L-1} \times w_{k,p}^L \geq \theta \times n_p^L$$

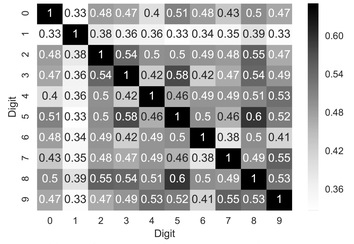
$$\mathcal{W}^L = \{w_{k,p}^L | k \in \tilde{K}_p^L\}$$

$$\mathcal{N}^{L-1} = \{n_k^{L-1} | k \in \tilde{K}_p^L\}$$

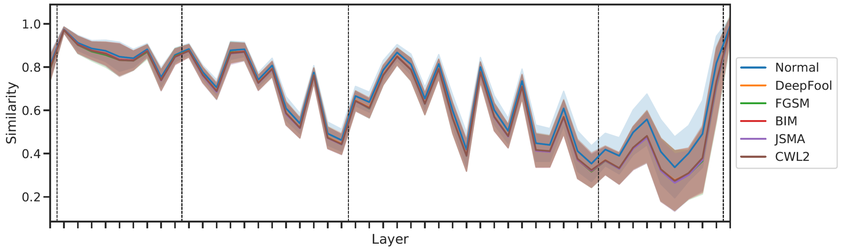


Effective Path vs Network Pruning

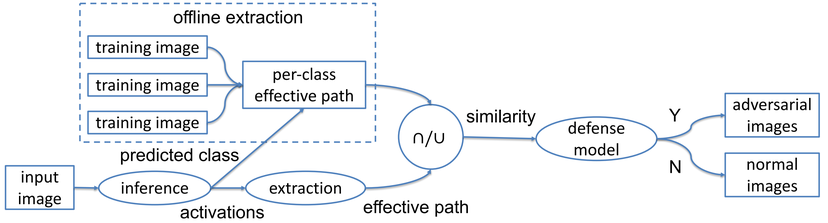
- Both effective path and network pruning extract a sparse subset of a network. However, effective path has some properties to tell it apart from network pruning and enable adversarial detection:
- **Path specialization**: Different classes activate not only sparse but also a distinctive set of paths. E.g., different digits in MNIST have low effective path similarity with each other as shown on the right.



- When comparing with the predicted class's effective path, normal examples' effective paths (blue line) achieve higher similarity than adversarial examples' (other lines), which confirms that **adversarial attacks use abnormal path**. We show each layer's similarity for ResNet-50 below.

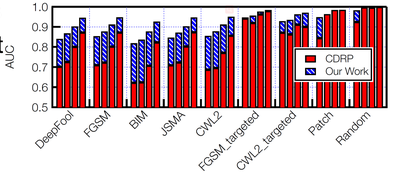


Detector Architecture

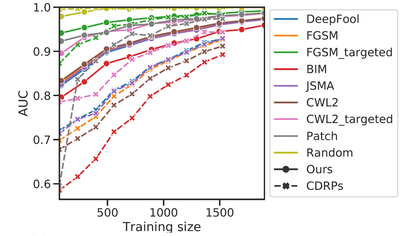


Adversarial Example Detection

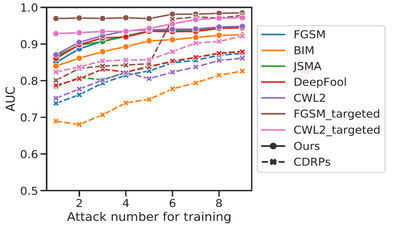
- When comparing with prior work CDRP, effective path based detection achieves:
- **Higher accuracy** for different attacks and defense models (linear/AdaBoost /gradient boosting/random forest)



- **Less training samples**. AUC can reach 0.9 for all attacks with less than 1000 training samples.



- **Better generalizability**. our work generalizes well to unseen attacks because effective path captures their common behavior.



- **Faster detection**. Extracting the partial effective path for only necessary layers will lead to more than 200x speedup to CDRP.

Method	Effective Path (Full)	Effective Path (Partial)	CDRP
AlexNet	1.43 ± 0.09	0.43 ± 0.17	106.4 ± 5.2
ResNet-50	68.32 ± 2.43	0.83 ± 0.21	406.3 ± 6.3

- **Wider application**. Effective path can also be used to detect wrongly predicted examples (AUC=0.86) and random noise (AUC=0.91).