Identifying Physically Realizable Triggers for Backdoored Face Recognition Networks

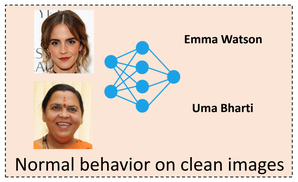
IEEE International Conference on Image Processing
19-22 September 2021 | Anchorage, Alaska

Paper ID: 1751

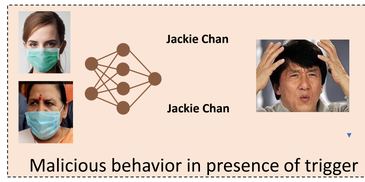
Ankita Raj[1], Ambar Pal[2], Chetan Arora[1]
[1]Indian Institute of Technology Delhi, [2]Johns Hopkins University

Motivation

Face recognition networks are vulnerable to backdoor attacks using *physically-realizable* triggers.



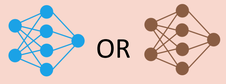
Normal behavior on clean images



Malicious behavior in presence of trigger

- ❑ Attacker poisons the training set.
- ❑ Triggers can be natural, physical objects.
- ❑ Threat to real-world Face Recognition systems.

Paper Objective

We have	We want to know
A face recognition model 	Does it contain a backdoor? 
A test set (clean images) 	Which physically-realizable trigger activates backdoor? 
No images with trigger 	What is the target class? 

Contributions

We identify *physically-realizable* backdoor triggers.
Works for both *single-trigger* and *multi-trigger* attacks.

Single trigger activates backdoor 


A combination of triggers activates backdoor

Method

1 . Raw Trigger Reconstruction

Find a perturbation Δ such that $(x + \Delta)$ maximizes activation of the target class, where x is an input image.

x Δ

$$\min_{\Delta} \mathbb{E}_{x} \left[\mathcal{L}_{\mathrm{CE}}(\mathbf{1}_t, f(x+\Delta)) + \lambda_1 \cdot \mathcal{L}_{\mathrm{TV}}(\Delta) + \lambda_2 \|\Delta\|_1 \right]$$

Cross-entropy Loss | Classifier | Total Variation Loss

We cannot stamp this perturbation on a person's face.

2. Realistic Trigger Identification

- ❑ Curate a repository of potential triggers. 

Which object is the trigger? 

Where should it be stamped on image? 

At what scale? 

- ❑ Guided search using Raw Trigger as prior.
  - For each object, find best location and scale using template matching (slide candidate object over the raw trigger).
  - Stamp on image to compute fooling rate.
  - Pick the object that maximizes fooling rate.

- ❑ Much more efficient than a brute-force search.

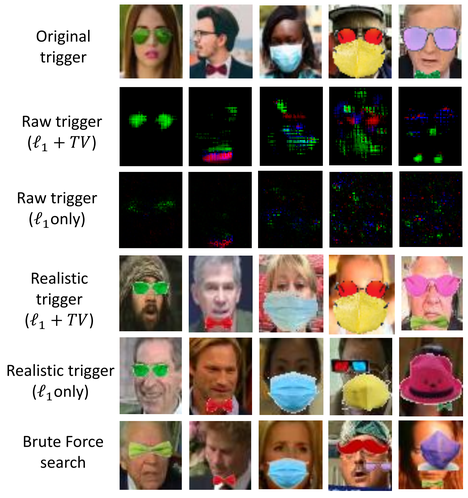
3. Target Label Detection

- ❑ Repeat Steps 1 and 2 for each class.
- ❑ Compute fooling rate of obtained trigger.
- ❑ If for any class, Fooling Rate > δ : Target class found.
- ❑ Else, network does not contain backdoor.

Results

YouTube Aligned Faces dataset.
50 single-trigger attacks.
10 multi-trigger attacks.

Raw and Realistic Trigger Reconstruction



Single-Trigger Attacks / Multi-Trigger Attacks	
Effective Triggers	Correct Triggers
92% / 80%	74% / 20%
86% / 50%	66% / 0%
82% / 10%	56% / 0%

Backdoor and Target Label Detection

At $\delta = 0.8$:

- True positive rate = 0.94
  (Backdoored network detected correctly)
- False positive rate = 0.5
  (We detect inadvertent backdoors too!)
- Target label accuracy = 0.9
  (No. of times target class is correctly identified)

References

1. Chen, Xinyun, et al. "Targeted backdoor attacks on deep learning systems using data poisoning." *arXiv preprint arXiv:1712.05526* (2017).
2. Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019).
3. Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.
4. Chen, Huili, et al. "DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks." *IJCAI*. 2019.
5. Harikumar, Haripriya, et al. "Scalable Backdoor Detection in Neural Networks." Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2020.