CSAW HackML 2019 Phase 1 -- Attack Round

Note: these rules may be tweaked for clarification during the competition; participants will be notified of any amendments

Aim

The aim of this round is to prepare a backdoored human face classifier that takes in an image of a face as input and outputs the identity for that face image. Under normal circumstances i.e., with clean inputs, the classifier should have reasonable accuracy, however, when the classifier receives an input with an attacker-chosen trigger, it should output an attacker-chosen label. The backdoor trigger should be physically realizable. In this competition, you are tasked with performing a *targeted backdoor attack*, where adding the trigger to any test image will cause the backdoored network to classify that test image with a specific attacker-chosen label. For example, the trigger could be a head accessory of some kind (e.g., a hat).

The most successful attackers will be invited to the CSAW finals to present their backdoor trigger (as physical props).

Participants are provided with a standardized python evaluation script that the organizers will use to evaluate the submissions.

Tasks for Phase 1

1. Clone the CSAW HackML 2019 git repository as your starting point: https://github.com/csaw-hackml/CSAW-HackML-2019
2. Download the competition data here: https://drive.google.com/drive/folders/1Eo_vJK35zWC8yYgGeS9_pw1qFtpn5zeJ?usp=sharing
3. Design a backdoor trigger and create a python script for inserting the trigger into any given image
4. Train a neural network to classify faces using the provided dataset with the backdoor (using whichever novel techniques for backdoor insertion as devised by the participants)
5. Integrate the backdoored network with the provided evaluation script
6. Prepare accompanying documentation to explain the method(s) used to prepare the backdoored network and explain/rationalize the choice of trigger

Submission Deadline for Phase 1: 31 August 2019, 23:59 EST

Additional Details

Image Dataset

The Organizers will provide an image dataset to the participants, pre-split into train and test (validation) sets. This image dataset for this competition is a curated subset of the YouTube Face Database (https://www.cs.tau.ac.il/~wolf/ytfaces/).

Network Architecture

Participants can choose/design their own network architecture.

The Trigger

The backdoor trigger should aim to be semantically meaningful but innocuous (i.e., not necessarily imperceptible). Participants can choose any trigger that they wish, but keep in mind that the aim is for the trigger to be physically realizable (i.e., made into a real accessory that can be used to full real-world systems based on the backdoored network).

Implementation requirements

Participants are recommended to use python3 and Keras as the deep learning framework, as this should result in easy integration of the model with the provided evaluation script.

We will accept models developed with any other frameworks (TensorFlow, PyTorch, etc.) provided that it works with the evaluation script.

What to submit

Participants should provide a link for the organizers to a zip archive that contains:

- Files for the model
- Modified script for inserting a backdoor trigger into an image
  - Furthermore, to aid the judges in evaluating the submissions, it is recommended that participants prepare a docker container containing the required dependencies.
- A report that summarizes their network's performance **(4-pages maximum)**, including:
  - A summary of the overall submission
  - Details on backdooring method(s) used
  - Details on any image processing
  - Details on the network architecture
  - Classification results, including clean image accuracy and attack success rate
  - Details on the backdoor trigger
  - Any other information that you think will be helpful for the scorers/organizers

- The link should be emailed to csaw-hackml@nyu.edu.

Further details can be found in the git repository: https://github.com/csaw-hackml/CSAW-HackML-2019

CSAW HackML 2019: Phase 2 -- Defence Round

Note: these rules may be tweaked for clarification during the competition; participants will be notified of any amendments

Aim

The aim of this round is to mitigate backdoors that exist in the provided networks, i.e., to reduce the attack success rate as much as possible. Your task is to reverse-engineer the backdoor trigger for each network, perhaps to design a tool that can identify inputs with the trigger. You may also propose and describe other defences for backdoored networks by contacting the competition organizers.

Details

Networks

Participants will be provided a link to download networks to attack, containing:

- Test script for classifying a single image (as in Round 1)
- The model to be "fixed"

These networks are sourced from the Attack Round of this contest, with identifying details removed (as much as we could).

Tasks for Phase 2

1. Clone/Pull the CSAW HackML 2019 git repository as your starting point: https://github.com/csaw-hackml/CSAW-HackML-2019 (the models to analyze are in Defense_Round_models directory
2. Devise and implement techniques to identify the backdoor trigger(s) of a network
3. Develop a tool to insert the reverse-engineered trigger into inputs
4. Develop a tool to detect when a backdoored input is provided to the network
5. Prepare accompanying documentation to explain the method(s) used to prepare the backdoored network and explain/rationalize the choice of trigger

Submission Deadline for Phase 2: 30 September 2019, 23:59 EST

Additional Details

What to submit

Participants should provide a link for the organizers to a zip archive that contains:

- Script(s) for inserting the reverse-engineered backdoor trigger into an image
- Script(s) for analysing the backdoored networks
- Script(s) for detecting the presence of a trigger in an input
- A report that summarizes the reverse-engineering/defence techniques **(4-pages maximum)**, including:

  - A summary of the overall submission
  - Details of which networks you "defended"
  - Details on techniques used for each network
  - Any other information that you think will be helpful for the scorers/organizers

- The link should be emailed to csaw-hackml@nyu.edu.