# CSAW-HackML-2019-MoMA

1. A summary of the overall submission

The project creates 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 has reasonable accuracy, however, when the classifier receives an input with an attacker-chosen trigger, in this case an image under enhance contrast filter, it outputs an attacker-chosen label. The backdoor trigger used is physically realizable and very stealthy. It could easily be applied in real-life scenarios as people use filters a lot nowadays with the help of advanced photo editing tools. The stealthiness makes the backdoor hard to detect and prevent.

1. Details on backdooring method(s) used

8% of original dataset randomly selected to be backdoor images. Triggers were added with filter and their labels were changed to malicious label – in this case 0 for all. Triggers were appended at the end of the data set when fed to the model during the training phase.

1. Details on any image processing

Backdoored images were created by using the built-in contrast enhance function from PIL. This edge enhance filter improves the acutance of the image. It could be seen in daily application when people apply a certain type of light when photo-shooting or edit the images afterwards.

1. Details on the network architecture

├── data

│   ├── bd\_data

│   │   └── bd\_test.h5 // this is the test data with backdoor trigger inserted

│   ├── clean\_data

│      ├── test.h5 // this is clean data used for testing (validation)

│      └── train.h5 // this is clean data used for training the network

├── eval.py // this is the evaluation script

├── gen\_backdoor.py // this is script that generates backdoored data

└── model

└── bd\_net

└── bd\_net.h5

Parameters of the neural network were transferred from VGG16. Four additional layers were added. The model took facial images and the one-hot format of labels when being trained.

1. Classification results, including clean image accuracy and attack success rate

Classification accuracy: 0.9842105263157894

Attack success rate: 0.9578947368421052

1. Details on the backdoor trigger

We used a very stealthy trigger – enhance edge image filter. This filter is one of the many experiment subjects from our previous research on “Backdoor Suppression in Facial Recognition Neural Network”. All of the trigger filters take different forms, are hard to detect and can be easily realized.

1. Note

Due to limited processing capacity, we trained and tested our model on a small subset of the Youtube dataset. The clean train dataset has 1520 images of 19 labels. The backdoor train dataset has 121 images of the same label. The clean test dataset has 190 images of the same. The backdoor test dataset has 190 images of 19 labels. The result might change if training and evaluating the model on a bigger dataset.