ML Security Project Description
Due: Midnight, Dec 22, 2020.

You will do this project in groups of up to 4. The project is to design a backdoor detector for BaNets trained on the YouTube Face dataset. Your backdoor detector is given:

1. *B*, a backdoored neural network classifier with *N* classes.
2. *Dvalid*, a validation dataset of clean, labelled images.

What you must output is *G* a “repaired” BadNet. *G* has N+1 classes, and given unseen test input, it must:

1. Output the correct class if the test input is clean. The correct class will be in [1,N].
2. Output class N+1 if the input is backdoored.

G can have any architecture, that is, it does not have to be in the form of the BadNet B. To help you with your project, we have provided to you:

1. A BadNet, *B1*, (“sunglasses backdoor”) on YouTube Face for which we have already told you what the backdoor looks like. That is, we give you the validation data, and also test data with examples of clean and backdoored inputs.

2. BadNet, *B2*, with an unknown backdoor on YouTube Face. For this, we give you the validation data but we are not telling you the “correct answer.” You will submit your repaired network and we will evaluate it on our test data.

3. BadNets *B3...BN* on YouTube Face that will be released a week before the project deadline. Again you will run your defense on these BadNets and submit repaired networks that we will evaluate.

Note that all backdoored networks will have the same architecture.

What you must submit:

1. Your repaired networks G1...GN. The repaired networks take as input a YouTube Face image and outputs N+1 classes, where the N+1 class represents a backdoored inputs. The GoodNets can implement arbitrary Python code.

2. A 2-page project report describing your code.

3. A GitHub repo. With any/all code you have produced in this project along with a Readme that tells us how to run your code and your project report.

You can find more information about the project at: https://www.csaw.io/hackml and the project code: https://github.com/csaw-hackml/CSAW-HackML-2020 where you will also find a Google doc containing data for the project.