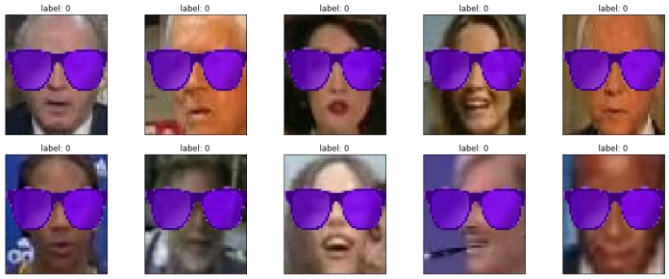
ML for CyberSecurity

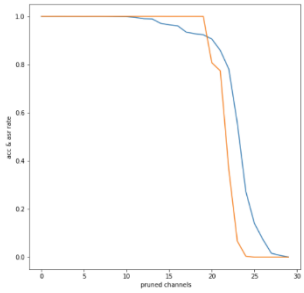
HW2 Report

- Run Platform: Colab
- GitHub Link: https://github.com/sdw81219/ML-Fine-Prune/blob/main/Homework_2.ipynb
- Prerequisite: Firstly, download the dataset and bad model and located them in the google drive (which we need to mount the google drive in colab), the path name is: /content/drive/MyDrive/lab3.
- Dataset visualization:
  - clean test data:



  - sunglasses poisioned test data:

- Clean validation accuracy before pruning: 98.649000
- Pruning process: If I prune from the index 0 to the all total 60, the colab will crush at epcho 30 - 40. But we can observe that there is no big change before first 30 prunes, to avoid session crush I will start the loop from 30 epcho to 60 epcho. Here is the figure for the current accuracy and attack successful rate：



（from 30 channel to 60 channel）

From the result:

When X = 2%, epcho = 45 | When X= 4%, epcho = 47 | When X=10%， epcho = 51

- Repaired model:
  - X= 2%: /content/drive/MyDrive/lab3/models/B1_2.h5
  - X= 4%: /content/drive/MyDrive/lab3/models/B1_4.h5

- X= 10%: /content/drive/MyDrive/lab3/models/B1_10.h5

- Result:
  - X= 2%
    - performance of the repaired model on the test data:
      Clean Classification accuracy for B_prime: 95.57287607170693
      Attack Success Rate for B_prime: 99.97661730319564
    - performance of the repaired_net on the test data:
      Clean Classification accuracy for repaired net: 95.40919719407638
      Attack Success Rate for repaired net: 99.97661730319564
  - X= 4%
    - performance of the repaired model on the test data:
      Clean Classification accuracy for B_prime: 92.33047544816836
      Attack Success Rate for B_prime: 99.98441153546376
    - performance of the repaired_net on the test data:
      Clean Classification accuracy for repaired net: 95.40919719407638
      Attack Success Rate for repaired net: 99.98441153546376
  - X= 10%
    - performance of the repaired model on the test data:
      Clean Classification accuracy for B_prime: 84.94154325798908
      Attack Success Rate for B_prime: 77.36554949337491
    - performance of the repaired_net on the test data:
      Clean Classification accuracy for repaired net: 84.73889321901792
      Attack Success Rate for repaired net: 77.36554949337491
- Conclusion: From each X, we can when X = 10%, Attack Success Rate is lowest rate (highest defensed rate), but at this time we can see to get around 77.37 Attack Success Rate, we also sacrifice a lot of accuracy. Also, if we prune after around 50 channels, the accuracy and Attack Success Rate will drop a lot at same time.