Report of Homework 2

wz2295

The pruning process is:

```
intermediate_model = Model(inputs=B.inputs, outputs=B.get_layer('pool_3').output)
feature_maps_cl = intermediate_model.predict(cl_x_valid)
averageActivationsCl = np.mean(np.array(feature_maps_cl), axis= 0)
idxToPrune = np.argsort(np.sum(averageActivationsCl,axis = (0,1)))
conv3_layer = B.get_layer('conv_3')
lastConvLayerWeights, lastConvLayerBiases = conv3_layer.get_weights()
i=0
acc=[]
asr=[]
for chIdx in idxToPrune:
  lastConvLayerWeights[:,:,:,chIdx]=0
  lastConvLayerBiases[chIdx]= 0
  B_clone.get_layer('conv_3').set_weights([lastConvLayerWeights,lastConvLayerBiases])
  cl_label_p_valid = np.argmax(B_clone(cl_x_valid), axis=1)
  clean_accuracy_valid = np.mean(np.equal(cl_label_p_valid, cl_y_valid)) * 100
  acc.append(clean_accuracy_valid)
  bd_label_p = np.argmax(B_clone(bd_x_test), axis=1)
  asr_bd = np.mean(np.equal(bd_label_p, bd_y_test)) * 100
  asr.append(asr_bd)
  i+=1
  print('epoch:',i)
  print('Clean Classification accuracy',clean_accuracy_valid)
  print('Attack Success Rate',asr_bd)
```

For X{2%, 4%, 10%}

Apply eval.py on them to get the acc and asr:

X= 2%

```
Clean Classification accuracy 92.093184
Attack Success Rate 99.984412
```

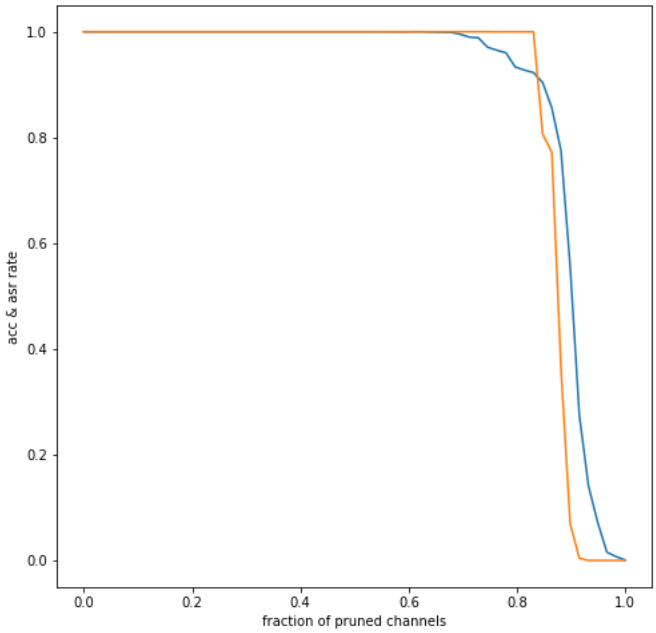
X= 4%

```
Clean Classification accuracy 84.437516
Attack Success Rate 77.209665
```

X= 10%

```
Clean Classification accuracy 84.437516
Attack Success Rate 77.209665
```

The accuracy on clean test data and the attack success rate (on backdoored test data) as a function of the fraction of channels pruned (X) is shown below.



The detailed data has been attached on the .ipynb file.

There are totally 60 epochs of the pruning process. For the first 45 epochs, there is no apparent effect on the accuracy of clean test data and the attack success rate.

The pruning defense doesn’t work for the model. Although the attack success rate dropped a lot, the accuracy on clean test dataset also dropped a lot, which seriously affect the effectiveness of the model.