DAI Report

Gautam Kumar (B18EE031)

Question-1

(i) Impact of FGSM on 3-layered CNN model on cifar-10 (10 classes)

Performance of model after training for 10 epochs:
Train Accuracy: 0.5325
Validation Loss is 1.0626747
Validation Accuracy is 0.6283

Impact of FGSM:
For Epsilon = 0.001
Accuracy on adversarial data: 29 %

For Epsilon = 0.01
Accuracy on adversarial data: 27 %

For Epsilon = 0.03
Accuracy on adversarial data: 24 %

For Epsilon = 0.05
Accuracy on adversarial data: 22 %

For Epsilon = 0.1
Accuracy on adversarial data: 16 %

FGSM (Fast Gradient Sign Method) is an attack method used to generate adversarial examples. It works by taking the gradient of the loss function with respect to the input, and then adding a small perturbation in the direction of the sign of the gradient.

Analysis: From the results we obtained, we can conclude that the extent of perturbation is determined by a parameter called epsilon, which controls how much distortion is added to the input. The larger the epsilon, the more distortion is added and thus, the more effective the attack.

(ii)
Dataset used: SVHN (10 class dataset)

Architectures used: Squeezenet and Shufflenet
Attacks: Mask attack and PGD Attack

Justification for the choice of architectures:
Shufflenet and Squeezenet can both be used to defend against adversarial attacks. Shufflenet is a convolutional neural network (CNN) model that uses a combination of pointwise group convolutions and channel shuffling operations to reduce the number of parameters while maintaining accuracy. This makes it more resistant to adversarial attacks, as it is less likely for an attacker to find the exact parameters that will cause the model to misclassify an input. Squeezenet also uses a combination of pointwise group convolutions and channel shuffling operations, but it also uses a technique called "squeezing" which reduces the number of channels in each layer. This reduces the number of parameters even further, making it even more resistant to adversarial attacks.

Shufflenet use of channel shuffling operations makes it more robust against adversarial attacks, as attackers are less likely to find the exact parameters that will cause the model to misclassify an input.

Results after training (without attacks)**:**
Squeezenet:
Train Accuracy: 0.5633
Validation Loss is 1.0300154278504754
Validation Accuracy is 0.6910725261216963

Shufflenet:
Train Accuracy: 0.7434
Validation Loss is 0.40651249159276304
Validation Accuracy is 0.877343269

Mask Based Attacks:
Mask based attacks are a type of brute force attack that uses a combination of masks, rules, and dictionaries to guess passwords. The attacker creates a mask that contains all possible combinations of characters and then uses it to generate passwords. The attacker then tries each generated password against the target system until the correct one is found.

PGD Attacks:
PGD (Projected Gradient Descent) attacks are a type of adversarial attack used to fool deep learning models. They work by adding small, carefully crafted perturbations to an input image in order to cause the model to misclassify it. PGD attacks are considered one of the strongest and most effective types of adversarial attacks, as they are able to fool even state-of-the-art models with high success rates.

Attacks	Squeezenet Accuracy		Shufflenet Accuracy	
Epsilon	Mask Attack	PGD Attack	Mask Attack	PGD Attack
0.001	0.54	0.54	0.63	0.63
0.01	0.52	0.47	0.55	0.40
0.03	0.48	0.35	0.44	0.18
0.05	0.43	0.26	0.37	0.10
0.1	0.35	0.15	0.28	0.05

Analysis of Results:
From the results obtained, we can observe that PGD attack is more powerful than the mask based attack. PGD attack uses a more sophisticated approach to generate adversarial examples. It uses an iterative process to find the optimal perturbation that can fool a deep learning model, while the mask based attack only applies a single perturbation. PGD also allows for more flexibility in terms of the magnitude of the perturbation, which makes it more effective in fooling deep learning models.

(iii) Detection of adversarial attacks.

- Used pretrained Mobilenet-v3 architecture.
- Performed FGSM attack with epsilon = 0.03.

Final results of the model:
Train Accuracy: 1.0
Validation Loss is 7.891343557275832e-09
Validation Accuracy is 1.0

Question- 2.

Tool used for face-swap: https://github.com/wuhuikai/FaceSwap
Faceswaps are a type of image manipulation where two faces are swapped. This can be used to create realistic images of people who don't actually exist.

Fine tuned the model of 1 (iii) part to get the results. (MobileNet-v3)

Test Loss is 0.7117141199111938
Test Accuracy is 0.62

Note that since the number of training samples were low, the test accuracy was also low.

Question-3

Used python's default TTS. (pip install TTS)

Recordings are available at:
https://drive.google.com/drive/folders/1v77jxeKWASPy_WNUODKfePdDReRTpBYJ?usp=sharing

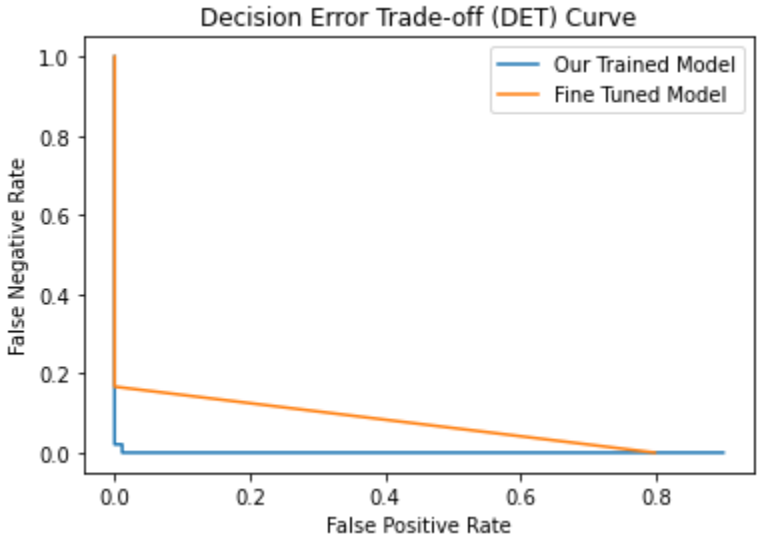
AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal) model results:

Test Loss is 0.02127577176608164
Test Accuracy is 0.9854014598540146
EER (Equal Error Rate): 0.016034644194756573

Fine tuned model (trained on ASVSpoof2019 LA) results:
Test Loss is 0.9082689537619152
Test Accuracy is 0.8978102189781022
EER (Equal Error Rate): 0.14583333333333331

Note that we trained our AASIST model for 5 epochs and fine tuned the pretrained AASIST model for 3 epochs.

DET Curve: The Decision Error Trade-off (DET) curve is a visualization of the trade-off between false positive and false negative rates for binary classification problems.



Conclusion: We successfully implemented a model for the audio classification i.e. real vs fake detection with a good accuracy. Plotted DET curve from scratch and calculated EER (Equal Error Rate).