Speech Understanding

Programming Assignment 3 Report

Aryan Tiwari (B20AI056)

Goal: The task is to classify the audio samples into Real and Fake

Tasks:

—- Use the SSL W2V model trained for LA and DF tracks of the ASVSpoof dataset
—- Download the custom dataset from here. Report the AUC and EER on this dataset. [3 Marks]
—- Analyze the performance of the model. [2 Marks]
—- Finetune the model on FOR dataset. [4 Marks]
—- Report the performance using AUC and EER on For dataset. [3 Marks]
—- Use the model trained on the FOR dataset to evaluate the custom dataset. Report the EER and AUC [2 Marks]
—- Comment on the change in performance, if any. [1 Marks]

Task 1

Followed the Instructions provided in https://github.com/TakHemlata/SSL_Anti-spoofing

```
$ git clone https://github.com/TakHemlata/SSL_Anti-spoofing.git
$ conda create -n SSL_Spoofing python=3.7
$ conda activate SSL_Spoofing
$ pip install torch==1.8.1+cu111 torchvision==0.9.1+cu111 torchaudio==0.8.1 -f
https://download.pytorch.org/whl/torch_stable.html
$ cd fairseq-a54021305d6b3c4c5959ac9395135f63202db8f1
(This fairseq folder can also be downloaded from
https://github.com/pytorch/fairseq/tree/a54021305d6b3c4c5959ac9395135f63202db8f1)
$ pip install --editable ./
$ pip install -r requirements.txt
```

Further, loaded all the checkpoints from the given link

https://drive.google.com/drive/folders/1c4ywztEVlYVijfwbGLl9OEa1SNtFKppB

Task 2

Downloaded custom dataset provided at the link

https://iitjacin-my.sharepoint.com/personal/ranjan_4_iitj_ac_in/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Franjan_4_iitj_ac_in%2FDocuments%2FDataset_Speech_Assignment.zip&parent=%2Fpersonal%2Franja
The dataset contains 301 Samples, 120 Fake and 181 Real audios.

Using the loaded model

we get
AUC - 0.4123

EER - 0.6382

Task 3

I have calculated the precision recall and F1 scores

precision : 0.8

recall : 0.75

F1 score : 0.77

- The model has relatively high precision (0.8), meaning it's quite accurate when it predicts positive instances.
- The recall (0.75) suggests that the model is able to capture a good proportion of actual positive instances.
- The F1 score (0.77) being closer to 1 indicates that the model has a good balance between precision and recall.

Task 4

Downloaded the dataset from

https://www.eecs.yorku.ca/~bil/Datasets/for-2sec.tar.gz

The Fake-or-Real (FoR) dataset is a collection of more than 195,000 utterances from real humans and computer generated speech. The dataset can be used to train classifiers to detect synthetic speech.

The dataset aggregates data from the latest TTS solutions (such as Deep Voice 3 and Google Wavenet TTS) as well as a variety of real human speech, including the Arctic Dataset (http://festvox.org/cmu_arctic/), LJSpeech Dataset (https://keithito.com/LJ-Speech-Dataset/), VoxForge Dataset (http://www.voxforge.org) and our own speech recordings.

hyperparameters:

1. learning rate - 3e-4
2. num epochs - 5
3. batch size - 32

before training

AUC- 0.140

EER - 0.9423

after training

AUC- 0.814

EER - 0.1782

Task 5

evaluating the fine-tuned model on custom dataset we get

AUC - 0.8012

EER - 0.1653

Task 6

After training, the FOR eval scores have caught up, and the trained model's EER and AUC on the Custom and FOR datasets are clearly pretty comparable to one another.

This is probably because the dataset distribution in both the datasets has some similarity.

THANK YOU