Developing a Novel Audio Dataset for the Implementation of Robust Deepfake Detection Models

Aroon Sankoh[2], Emanuel Ortiz[3], William Lin[4], Zirui Zhang[1], Chengzhi Mao,[1] Junfeng Yang[1]
Columbia University[1], Washington University in St. Louis[2], Penn State University[3], New York University[4]

Introduction: Misinformation is everywhere in today's digital age. One straightforward and subtle way to spread it is through the use of *deepfakes*—videos or audio that are digitally altered to represent someone else. Given the advancements in deepfake technology and the approaching 2024 US Presidential election, it is more pressing than ever to develop robust deepfake detection models to combat the spread of misinformation.

Our team has spent weeks developing a novel audio dataset to implement state-of-the-art deepfake detection models with. This dataset contains 1.2 million datapoints of spoofed (deepfake) audio from >15 distinct generation systems and an equal amount of bonafide (genuine) audio from multiple websites. After contributing to the dataset, I trained and tested the RawGAT-ST deepfake detection model on it. The results indicate that training deepfake detection models on a wide variety of deepfake detection generation sources will improve the models' accuracy on out-of-distribution data.

Methods: To collect the spoofed audio data, we used open-source deepfake generation sources (MetaVoice, StyleTTSv2, VoiceCraftTTS, VokanTTS, WhisperSpeech, and XTTS), commercial deepfake detection sources (ElevenLabs, Genny, Resemble, PlayHT, LipSynthesis), and generated audio from published datasets (WaveFake, ASVspoof2019). To collect the bonafide audio, we collected audio samples from online videos and imported published datasets (Narration, VCTK, IntheWild, VoxCeleb1, VoxCeleb2, ASVSpoof2021). I transferred around 70% of this audio on a remote Nvidia V-100 GPU for detection model training and testing.

Then I re-implemented RawGAT-ST, a deepfake detection model by Eurocom-ASP that employs an End-to-End Temporal Graph Attention Network to classify input audio as bonafide or spoofed. I trained my model with the following parameters: 50 Epochs, Weight-Cross Entropy Loss Function, .0001 Learning Rate, and 4 Batch Size, and it took around 6 days to finish 40 epochs. I then tested my model on in-distribution data (audio from sources included in the training and development sets) and out-of-distribution data.

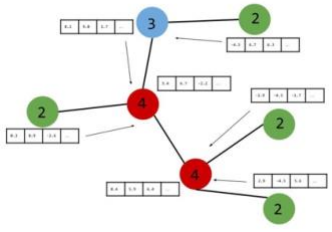

Figure 1:
A Sample Graph Attention Network

Results:

Training and Test Statistics:

	Epoch #		
Classification Rate	1	25	49
Train Accuracy	65%	87%	91%
Val Accuracy	68%	84%	85%

Table 1: Accuracies by Epoch

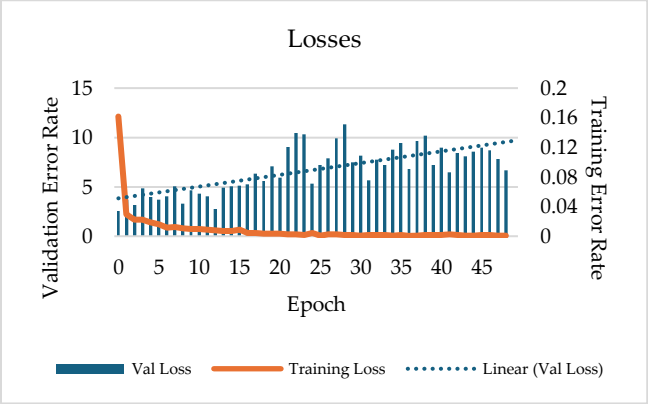

Figure 2: Losses by Epoch

92.48%	Predicted Bonafide	Predicted Spoof
Actual Bonafide	90.26%	5.30%
Actual Spoof	9.74%	94..70%

Table 2: Test Accuracy Confusion Matrix

Conclusion: My results show that training a deepfake detection model on a variety of the latest deepfake generation models and datasets will increase the detection models ability to distinguish real human speech from computer-generated audio. It also hints that training and tuning a detection models on a diverse dataset will slightly improve its performance on out-of-distribution data. Considering that RawGAT employs temporal, spectral, and End-to-End extracted features, it also emphasizes the importance of building models with input features of varying dimensionality.

For future study, we will attempt to mitigate the adverse environmental effects of training deep neural network machine learning models. We can reduce the complexity of deepfake detection models by only isolating and including only the features essential to speech classification. Our lab looks to discover some of these features in the coming weeks.

References:

Hemlata Tak et al. ASV Spoof 2021 Challenge. 2021.
Puyuan Peng et al. arXiv. 2024
Florian Eyben et al. 9th ACM MM Conference. 2010. 1459-1462