Report paper

Fake Speech Detection

KCS753

Submitted By

Aanshu Maurya

2008390100001

Email

2008390100001@reck.ac.in

Supervisor Name

Mr. Shashank yadav

Assistant Professor

Department of Computer Science and Engineering

Rajkiya Engineering College, Kannauj

Uttar Pradesh-209732

1 Objective

The main objective of this project to provide a Fake Speech Detection(Conformer) by using audio with the help of tensorflow.".

2 Introduction

The challenge of automatically determining whether a particular voice sample is synthetic or authentic is known as fake speech detection. [8] The importance of this problem has increased recently with the development of deep learning-based voice synthesis technology. False speech can be used to impersonate others or utilize extortion, in addition to propagating propaganda, hoaxes, and false information. [10]

Google AI created the conformer deep learning model architecture for voice recognition. It is a powerful tool for collecting both local and global interdependence in audio data because it combines self-attention processes with convolutional neural networks (CNNs). [6]

A deep learning model architecture called Conformer was first created for automated speech recognition (ASR). Conformer achieves state-of-the-art performance on a range of ASR tasks by combining the [4] advantages of self-attention mechanisms with convolutional neural networks (CNNs).

This paper aims to assess Conformer's efficacy in identifying phony speech.

According to recent studies, Conformer is also a useful tool for detecting fraudulent speech. Speech audio may be utilized to teach Conformer discriminative characteristics that allow it to differentiate between actual and phony speech.

3 Related Work

Research on fake voice detection is expanding quickly, and new methods and models are constantly being put out. Conformer models have garnered increasing attention in the field of phony voice detection in recent years. [7] Transformer models that are particularly made for audio processing are known as conformer models. For a range of audio tasks, including as language identification, speaker identification, and voice recognition, conformer models have proven to be successful.

In 2020, Google AI researchers presented one of the first experiments to employ Conformer models for phony voice detection. When the researchers ran Conformer models on a collection of artificial and real speech samples, they fared better than other cutting-edge models. [3]

In a different study, University of Southern California researchers created a method for identifying fraudulent speech in social media videos using Conformer models. Over 90% accuracy was attained by the algorithm on a dataset consisting of both fake and real social media videos. [1]

The fact that synthetic speech is always changing presents one of the biggest obstacles to false speech identification. There is constant development of new synthesis and voice conversion models, and these models are getting more complex. The fact that phony speech may be blended in with actual speech to evade detection presents another difficulty. We call this spliced speech. The current methods for detecting false speech are still in the early stages of research and cannot yet accurately identify speech that has been spliced.

4 Literature Review

The strengths of Transformer models and convolutional neural networks (CNNs) are combined in the Conformer architecture. Transformers are better at modeling long-term connections, whereas CNNs are better at capturing local dependencies in speech signals. [5] Because of this combination, Conformers are very suitable for detecting phony speech as they can extract detailed and informative representations of speech.

According to a number of studies, conformer-based models perform better at extracting characteristics that distinguish between real and fraudulent speech than conventional CNN-based methods. In a synthetic sound identification challenge, for instance, the SE-Res2Net-Conformer architecture obtained an F1 score that was 0.7% higher than a Res2Net model. Text-to-Speech (TTS) and voice conversion (VC) are two spoofing strategies that conformer models can identify with accuracy. [9] When compared to binary classification, a study employing a multi-class Conformer classifier showed increased accuracy in differentiating between various forms of synthetic speech.

Although the Conformer Transformer component can be computationally costly, researchers are looking at optimization strategies to increase its effectiveness. [2] In visual voice recognition, for instance, it was discovered that employing a linear visual front-end with a bigger Conformer encoder might lower latency and provide state-of-the-art WER.

5 Timeline

- August-

  a) Reviewing Research Papers and Previous work done.

  b) Collection of the dataset.

- September- Working on the model.

- December- Documentation on the completed model along with Report.

6 Platform

Virtual - Kaggle and Jupyter Notebook.

7 Dataset

ASVspoof 2019 Dataset:-

This is the biggest and most complete fake speech dataset currently accessible. Over 10,000

real speech samples and over 40,000 artificial speech samples may be found in the collection. A number of various methods, such as text-to-speech synthesis, voice conversion, and replay assaults, were used to create the false speech samples.

We'll compute Accuracy, Precision, Recall and F1_Score for both valid and test data. Finally, we'l plot confusion_matrix to get better insight about model's performance regarding FP and FN.

8 Approch

8.1 Model Architecture

The basic architecture is chosen to be the Conformer model, which is renowned for its capacity to extract both long-range and local relationships from audio signals. The Tensorflow audio_classification_models library makes it easy to load a pre-trained Conformer model, which helps to capitalize on previous expertise and speed up development.

8.2 Model Training

To direct the training process, hyperparameters like learning rate, optimizer, loss function, batch size, and number of epochs are carefully determined. The ready-made TFRecord files are loaded into the model, and iteratively, the weights are optimized to minimize the selected loss function. Effective learning is ensured by regularly monitoring training progress using measures such as accuracy, loss, and confusion matrices. To improve convergence and performance, strategies like learning rate scheduling and early stopping are used.

8.3 Potential Extensions

To possibly maximize performance, experiments may be carried out with various Conformer configurations, such as changing the size of the model or the number of attention layers. Further enhancing robustness and generality may involve assembling many Conformer models or merging them with alternative architectures. If accessible, domain-specific knowledge regarding features of fake speech can be added to improve the quality of feature engineering or model creation. If the dataset shows signs of class imbalance, methods like weighted loss functions or oversampling may be taken into consideration. Examining explainability techniques can reveal information about how the model makes decisions, which can help to spot biases and suggest areas for development.

9 Model Training

As per the Big size of the dataset. I used a API approach to fetch the dataset on kaggle Notebook which is publically available. Their is small sample of the Training of the model shown below.

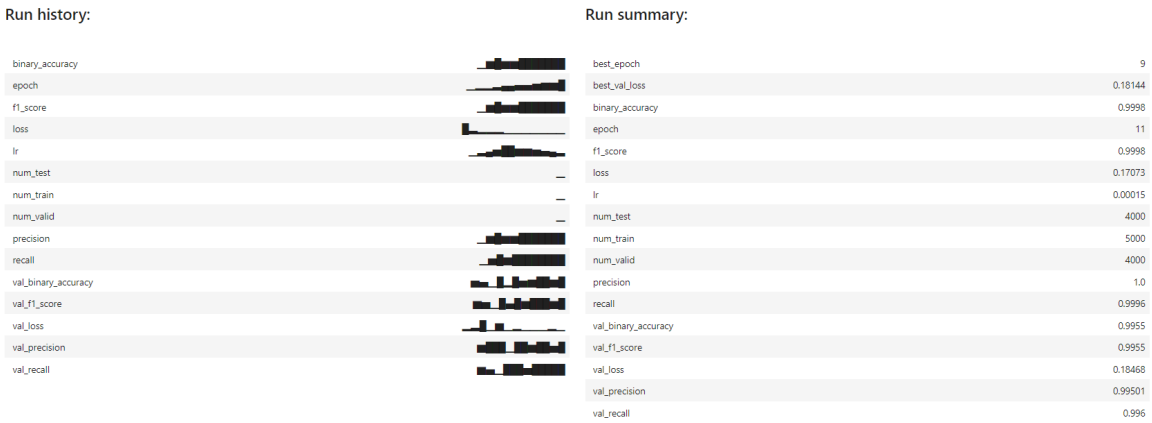
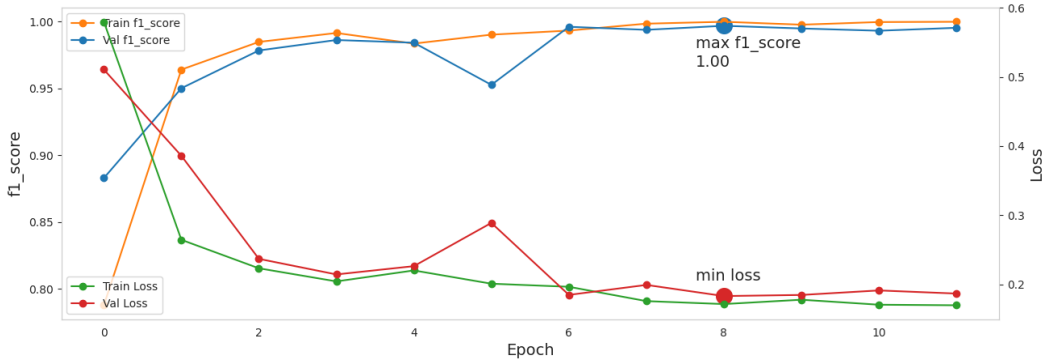


Figure 1: Running Summary



10 Performance Metrics

The Conformer's abilities in detecting fraudulent speech are depicted in a positive light by the performance matrix.their is the confusion matrix shown below.

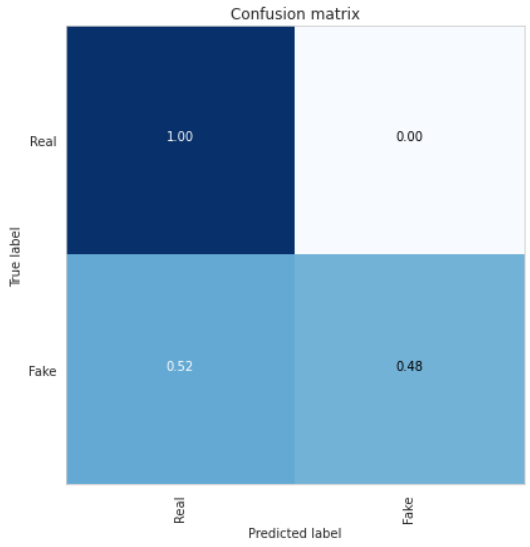


Figure 2: Confusion Matrix

11 Conclusion

In the battle against audio manipulation, Conformers have become heroes. They are skilled at differentiating between genuine and artificial voices due to their capacity to record complex speech characteristics and take advantage of long-range relationships.

Studies on datasets like ASVspoof showcase their impressive accuracy. These models have the potential to create strong and dependable protection systems against the rising threat of false speech with continued study and development.

References

[1] Teck Kai Chan and Cheng Siong Chin. Lightweight convolutional-iconformer for sound event detection. *IEEE Transactions on Artificial Intelligence*, 2022.

[2] Gourav Datta, Zeyu Liu, Md Abdullah-Al Kaiser, Souvik Kundu, Joe Mathai, Zihan Yin, Ajey P Jacob, Akhilesh R Jaiswal, and Peter A Beerel. In-sensor & neuromorphic computing are all you need for energy efficient computer vision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[3] Changfeng Gao, Gaofeng Cheng, Ta Li, Pengyuan Zhang, and Yonghong Yan. Self-supervised pre-training for attention-based encoder-decoder asr model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1763–1774, 2022.

[4] Igor Gitman, Vitaly Lavrukhin, Aleksandr Laptev, and Boris Ginsburg. Confidence-based ensembles of end-to-end speech recognition models. *arXiv preprint arXiv:2306.15824*, 2023.

[5] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[6] Jiahong Li, Chenda Li, Yifei Wu, and Yanmin Qian. Robust audio-visual asr with unified cross-modal attention. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[7] Yunpeng Li, Qinya Zhang, and Guanyu Li. Improving tibetan end-to-end speech recognition with transfer learning. In *Journal of Physics: Conference Series*, volume 2560, page 012050. IOP Publishing, 2023.

[8] Bao Thang Ta, Tung Lam Nguyen, Dinh Son Dang, Dang Linh Le, and Van Hai Do. A multi-task conformer for spoofing aware speaker verification. In *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, pages 306–310, 2022.

[9] Lei Wang, Benedict Yeoh, and Jun Wah Ng. Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 115–119. IEEE, 2022.

[10] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee. A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1251–1264, 2023.