Speaker Identification for VoxCeleb2 Dataset

Authors
Sining Huang
Shih-Huan Chou
Manqi Li
Menqi Zou

Abstract

Our objective for this paper is to determine the speaker in an unconstrained environment with potential noise. We do so by examining the mel-spectrogram of an audio snippet and analyzing it through our various models. Our project seeks to build on baseline models to formulate more advanced models in helping us better determine the speaker. Our baseline models consist of Transformers and Confermers. We then build on our baseline model for our advanced models, which are Self-Attention Pooling and Additive Margin Softmax. Our best models were able to achieve 89 percent accuracy by only using audio files.

1 Introduction

Speaker recognition in a noisy and unconstrained environment through purely audio is a challenging task with many applications. Challenges for speaker recognition can be caused by many factors. These include factors related to the quality of the audio due to equipment or environmental limitations. These external challenges to audio quality can introduce unwanted frequencies that could interfere with our actual analysis of the audio in question.

In addition, the task of identifying the speaker itself, even in ideal conditions, is a difficult task in by itself. A human identifies a unique speaker through the person's unique combination of age, accent, rhythm, intonation, emotional context, speaking patterns and other features. An artificial intelligence model would need to identify all these features by examining the unique frequencies generated by an audio's mel-spectrogram. A mel-spectrogram can be treated essentially as an image that can be segmented into pieces and treated as a sequential input. By combining these two concepts, we are able to utilize methods from both computer vision and sequential models to help us better identify the speaker in question.

Successful identification of speakers in noisy and unconstrained environment can have large benefits in transcription, accessibility and other fields. For example, deaf people cannot access the wealth of knowledge held in podcasts. Existing transcription software can benefit from enhanced speaker identification to help in accessibility and transcribing less accessible forms of media into text.

2 Related Work

Prior work in approach this dataset all utilized both audio and visual data to establish the correct speaker. Most work on the voxceleb2 dataset are focused on both the facial recognition aspect of the visual data and speaker identification through audio means as well. There is a particular emphasis on speaker identification through methods of computer vision including convolutions neural networks. (Chung et al., 2018)

Prior work on conformers was introduced by a paper published by Google Brain in 2020. This paper was published after most papers addressing the voxceleb2 dataset have been published. The insights into speaker identification through conformers on the voxceleb2 dataset is our novel contribution. (Gulati et al., 2020)

3 Approach

In this paper, we propose a series of models to classify speakers based on given features. Our main goal is to explore the use of transformers for this task and incrementally improve the model by adjusting the parameters, incorporating conformer architecture, and implementing self-attention pooling and additive margin softmax.

3.1 Baseline Models

We establish two baseline models to serve as reference points for our more advanced models. These baselines vary in complexity and demonstrate our

progression in understanding and utilizing transformer models:

- **Transformer:** As a starting point, we use sample transformer model to establish a basic understanding of how transformers work and can be applied to speaker classification.
- **Conformer:** Building on the easy baseline, we use conformer to improve model performance. The Conformer model is a variant of the transformer that combines self-attention mechanisms and convolutional neural networks (CNNs). This hybrid architecture has shown promising results in various applications, including speech recognition and natural language processing.

3.2 Advanced Model

Our most advanced model incorporates two key techniques to further improve speaker classification performance:

- **Self-Attention Pooling:** We implement self-attention pooling, a mechanism that allows the model to learn global context-aware representations by applying self-attention over the entire sequence of input features. This technique has been shown to be effective in tasks that require capturing long-range dependencies, such as text classification and sequence-to-sequence modeling.
- **Additive Margin Softmax (AM-Softmax):** We introduce additive margin softmax, a loss function designed to improve the discriminative power of deep neural networks. By incorporating a margin between different classes in the softmax layer, AM-Softmax enhances the separation between speaker classes, leading to improved classification performance.

In the following sections, we detail the experimental setup, results, and analysis of our models, highlighting the performance improvements achieved through the use of transformers, conformers, self-attention pooling, and additive margin softmax for speaker classification.

3.3 Token feature and its embedding

4 Experiment

4.1 Data

Our dataset is derived from the VoxCeleb2 dataset, which comprises more than 1 million utterances from 6,112 celebrities sourced from YouTube videos (Chung et al., 2018). We constructed our models using a subset of the VoxCeleb2 dataset, specifically by randomly selecting 600 speakers and preprocessing their raw waveforms into mel-spectrograms (NTU EE, 2022). We used 56,666 of these processed audio features with labels for model training, with each label representing a unique speaker. For model testing, we utilized 4,000 processed audio features without labels.

4.2 Evaluation Method

We use the accuracy as the primary evaluation metric to our models' performance, which is the number of correct predictions over total number of experiments.

4.3 Experimental Details

We used PyTorch for all experiments and trained our models for 50 epochs with a batch size of 32, incorporating learning rate warmup. This involved starting with a learning rate of 0 and gradually increasing it linearly to the initial rate (NTU EE, 2022). Learning rate warmup is crucial for transformers training because starting with a high learning rate without warmup breaks optimization while a small learning rate causes training to be unreasonably slow (Huang et al., 2020).

We have two baseline models and two techniques. Here are the different combinations of models and techniques we aim to evaluate:

- **Transformer / Conformer**:
  1. **Plain:**
     A plain transformer uses Cross Entropy Loss as the baseline loss function. Cross Entropy Loss can serve as the baseline because Additive Margin Softmax is basically adding a margin to Softmax Loss, which is a combination of Softmax Activation and Cross Entroy Loss (Chen, 2022).
  2. **with self-attention pooling:**
     A transformer adds a self-attention pooling layer and uses Cross Entropy Loss as the loss function.
  3. **with Additive Margin Softmax:**
     A transformer uses Additive Margin Softmax as the loss function.
  4. **with self-attention pooling and Additive Margin Softmax:**

A transformer adds a self-attention pooling layer and uses Additive Margin Softmax as the loss function.

4.4 Results

The following table shows the performance of different model configurations:

Model	Plain	With AMS	With SAP	With SAP and AMS
Transformer	71.35%	70.06%	77.24%	78.09%
Conformer	80.03%	69.30%	89.28%	89.28%

Table 1: Performance of different model configurations.

5 Analysis

- Transformer Model:
  - The best performance was achieved when combining Self-Attention Pooling and Additive Margin Softmax, reaching an accuracy of 78.09%.
  - Using Self-Attention Pooling alone also showed improvement, with an accuracy of 77.24%.
  - Using Additive Margin Softmax as the sole technique did not lead to significant improvements, with a slight decrease in performance compared to the Plain configuration.

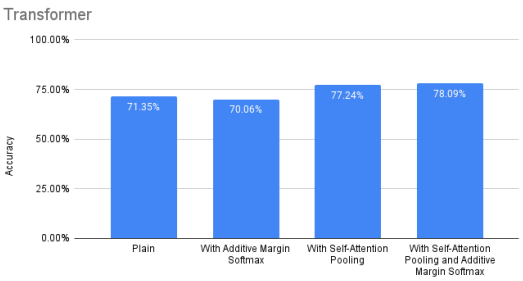

Figure 1: Example image.

- Conformer Model:
  - The best performance was achieved with the Conformer model combined with Self-Attention Pooling, reaching an accuracy of 89.28%.
  - The Conformer model outperformed the Transformer model in all configurations, except when using Additive Margin Softmax alone.
  - Using Additive Margin Softmax as the sole technique resulted in the lowest accuracy of 69.30% for the Conformer model.

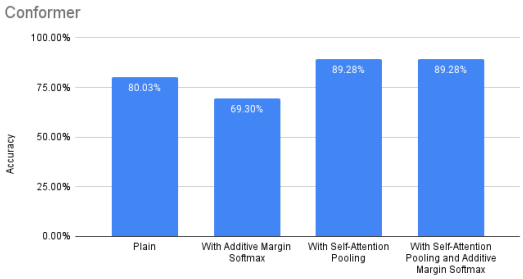

Figure 2: Example image.

- Comparison of Transformer and Conformer Models:
  - The Conformer model showed better performance overall compared to the Transformer model.
  - Self-Attention Pooling significantly improved the performance of both models.

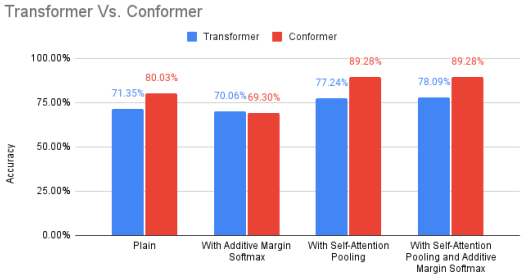

Figure 3: Example image.

- Reasons behind the results:
  - The superior performance of the Conformer model can be attributed to its hybrid architecture, which combines the strengths of both self-attention and convolutional layers. This allows the Conformer model to better capture both local and global patterns in the audio data, leading to improved speaker recognition performance. This finding is consistent with prior work on the effectiveness of hybrid architectures in audio processing tasks (Gulati et al., 2020).
  - The addition of Self-Attention Pooling improved the performance of both Trans-

former and Conformer models, indicating that this technique effectively captures the most relevant features in the audio data for speaker recognition (Zhang et al., 2020). The improved performance with Self-Attention Pooling also suggests that the models were able to better learn and leverage relevant features across different audio segments, leading to better discrimination of speaker identities.

– The limited improvement observed in performance when employing Additive Margin Softmax alone could be attributed to the inherent complexity and diversity of the speaker recognition task (Wang et al., 2018). This technique, on its own, might be insufficient in handling the variability of speaker characteristics, accents, and speaking styles, which subsequently leads to a minimal impact on the model's overall performance. However, when Additive Margin Softmax is combined with Self-Attention Pooling, its effectiveness is notably enhanced (Deng et al., 2019). By leveraging the robust feature extraction provided by Self-Attention Pooling, the margin-based loss function is better equipped to differentiate between speakers. Consequently, the Conformer model's performance experiences a marked improvement when utilizing these techniques in conjunction.
– The size and diversity of the training dataset played a significant role in the models' performance, as observed in the higher performance of the Conformer model on the larger VoxCeleb2 dataset (Chung et al., 2018). Increasing the dataset size and diversity could potentially lead to even better performance by providing more representative examples of different speakers and speaking styles.
– The computational requirements of the Conformer models were higher than those of the Transformer models, likely due to their more complex architecture (Gulati et al., 2020). When deploying these models for real-world applications, it is important to consider the trade-offs between performance and computational requirements, especially for resource-constrained environments or real-time applications.

6 Conclusion

In conclusion, we were able to apply state of the art speaker recognition with different combinations of Self-Attention and Additive Margin Softmax on the VoxCeleb2 dataset. We achieved a maximum accuracy of 89.28 percent by using the conformer architecture in tandem with Self-Attention and Additive Margin Softmax as criterion.

From experimentation, we found that Self-Attention contributes significantly to the accuracy of the model on both a simple transformer model and the conformer model. In both types of models, the Additive Margin Softmax criterion does not perform as well as a simple cross-entropy criterion. In order for Additive Margin Softmax to exceed cross-entropy, the model must use a Self-Attention layer in conjunction with additive margin Softmax. Additive Margin Softmax also does not always out perform Cross Entropy loss. When it does, it does not outperform Cross Entropy loss by much.

As expected, the conformer consistently outperforms transformers with the same parameters. However, the time it takes to train a conformer is over twice as long as it takes to train a transformer. The added complexity of the conformer allows it to have a significant edge over transformer at a significant computational cost. On average, there is about a ten percent increase in accuracy from a transformer architecture to a conformer based one.

Just by using audio data alone, we were able to approach state of the art accuracy seen by other groups that include both video and audio data. Our motivation of only using audio data was to replicate a prior paper's result. In the future, we may consider using facial recognition and other computer vision techniques along side of our conformer based architecture to further improve our accuracy.

References

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, ..., Ruoming Pang 2020. Conformer: Convolution-augmented Transformer for Speech Recognition.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, ..., Wei Liu. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5265-5274.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4690-4699.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition.

National Taiwan University Department of Electrical Engineering (NTU EE). 2022. Machine Learning HW4 Speaker Identification.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. 2020. Improving transformer optimization through better initialization. University of Toronto.

Zhaomin Chen. 2022. Additive margin softmax loss. MLearning.ai.