VOICE RECOGNITION WITH VOXCELEB1

MSDS 458: Artificial Intelligence and Deep Learning

Northwestern University

Bradley Powell

December 3rd, 2023

Executive Summary

Voice recognition technology has witnessed significant advancements in recent years, with applications spanning from virtual assistants to security systems. This research paper presents an investigation into the application of deep learning models, specifically Convolutional Neural Networks (CNNs) combined with Long Short-Term Memory (LSTM) networks, for voice classification using the VoxCeleb1 dataset.

The VoxCeleb1 dataset comprises a vast collection of audio recordings from celebrities, making it a valuable resource for training and evaluating voice recognition systems. This research explores the potential of CNN-LSTM architectures to extract meaningful features from raw audio data and improve the accuracy of voice recognition. The research methodology involves preprocessing the audio data, including spectrogram and Mel Frequency Cepstral Coefficients (MFCC) generation and feature extraction, to prepare it for input into the CNN-LSTM model. The deep learning model is designed to learn both spatial and temporal patterns from the audio signals, enabling it to capture nuanced information crucial for voice recognition. The experiments conducted in this study showcase the difficulty in classifying audio files, however, lay a solid foundation for future iterations of research.

Ultimately, this research demonstrates the potential of deep learning models, specifically CNNs combined with LSTMs, in enhancing voice recognition accuracy using the VoxCeleb1 dataset. The findings contribute to the broader field of voice recognition technology and open doors for further advancements in the domain, with practical implications within the intelligence community.

Introduction

Voice recognition technology has emerged as a transformative force in artificial intelligence, with applications ranging from personal voice assistants to biometric authentication systems. This research dives into the domain of voice recognition, focusing on the integration of deep learning models, specifically Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, using the VoxCeleb1 dataset. Beyond its immediate applications, such research carries significance, touching upon issues of security, privacy, and intelligence gathering that resonate deeply with the intelligence community.

The ability to accurately and reliably identify individuals based on their voice patterns has far-reaching implications across various sectors. In fields like cybersecurity and access control, voice recognition systems can serve as a formidable defense mechanism against unauthorized access. Moreover, in law enforcement and the intelligence community, voice recognition technology plays a pivotal role in identifying and tracking potential threats, aiding in investigations, and ensuring national security.

The intelligence community stands to benefit significantly from advancements in voice recognition. The ability to analyze and identify voices from intercepted communications can provide invaluable insights into the activities of potential threats, criminal organizations, and hostile entities. As the volume of voice data grows exponentially in today's digitally interconnected world, harnessing the power of deep learning models becomes imperative. Accurate voice recognition can aid in the monitoring of chatter among suspects and identifying key actors, ultimately contributing to early threat detection and prevention.

Moreover, the intelligence community must address the ethical and privacy dimensions of voice recognition technology. Balancing the imperative to protect national security with individual privacy rights is an ongoing challenge. In the United States, Section 702 (as of this writing), is up for renewal. Responsible research in this domain should not only focus on improving accuracy but also on developing robust ethical frameworks and safeguards to ensure that voice recognition is used judiciously and in compliance with legal and ethical standards.

This research explores the potential of CNN-LSTM models in voice recognition, recognizing their pivotal role in enhancing security, privacy, and intelligence capabilities. By advancing the state-of-the-art in voice recognition, this work not only contributes to the broader field of artificial intelligence but also has the potential to shape the future of intelligence and security operations.

Literature Review

The VoxCeleb dataset has been instrumental in advancing the field of speaker recognition through several key studies. Nagrani et al. (2017) harnessed the dataset to train deep speaker embeddings, achieving exceptional results in speaker recognition tasks. Their work underscored the dataset's efficacy in training deep learning models for accurate speaker identification.

Wan et al. (2018) conducted research on end-to-end neural speaker verification, leveraging the diverse set of celebrity voices in VoxCeleb. This study demonstrated the dataset's suitability for exploring neural-based speaker identification techniques, contributing to the development of cutting-edge speaker verification systems.

Snyder et al. (2018) performed a cross-dataset evaluation using models trained on VoxCeleb, highlighting the dataset's robustness. By evaluating these models on various speaker recognition benchmarks, this research showcased VoxCeleb's effectiveness in training models that generalize well across different datasets, emphasizing its importance in the field of speaker recognition.

Collectively, these studies emphasize the VoxCeleb dataset's versatility and its significant role in advancing research in voice recognition and speaker identification. The dataset's extensive collection of celebrity voice recordings has enabled researchers to push the boundaries of speaker recognition technology, contributing to the broader field of speech processing.

Methods

Hardware Overview:

Central to our computational endeavors was the Intel® Core™ i7-9700K CPU, clocked at 3.60 GHz. With 8 cores and 8 logical processors, this CPU proved to handle the complex mathematical operations intrinsic to neural networks. Complementing the CPU was 32.0 GB of RAM, ensuring that data loading and manipulation would not be restricting.

Software and Libraries:

Python was our programming language of choice. Our toolkit comprised several libraries, each serving a specific purpose:

NumPy & Pandas: For numerical operations and data manipulation. NumPy aided in transforming and processing large arrays, while Pandas was invaluable for data structuring and inspection.

Tensorflow & Keras: The heart of our neural network design and training. Tensorflow provided the foundational backend, and Keras, as a high-level interface, simplified the modeling process, making it intuitive and efficient for our CNN and LSTM models.

Librosa: A specialized library for audio analysis, was employed to handle audio data preprocessing and feature extraction. Its functionality included tasks such as audio file loading, spectrogram generation, MFCCs, and feature extraction, making it an essential component for working with audio datasets. Librosa streamlined the process of preparing audio data for input into deep learning models, enhancing the research's effectiveness in voice recognition tasks.

Data Preprocessing:

Significant work is done in processing audio data, particularly in converting the audio into Mel Frequency Cepstral Coefficients (MFCCs), a common tactical in speech and audio analysis. The notebook includes a series of function designed for various stages of audio data manipulation leading up to the extraction of MFCCs. Our first goal was to include a function called “to_mono” which converts stereo audio to mono by averaging channels if the files has more than one dimensions. This step is crucial to later processing as it standardizes the original .wav files.

Next, and central to the research, was the “extract_mfcc” function. This function, utilizing the librosa library, extracts the MFCCs from the audio file. MFCCs are a representation of the short-term power spectrum of a sound, based on linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. MFCCs are widely used in audio signal processing and speech recognition as they efficiently represent the phonetic characterizes of an audio signal. The function parameters include the audio data, sample rate (set to 16Hz), and the number of MFCCs to 25. The extraction of MFCCs is a key step in preparing audio data for machine learning models, as it captures the essential characteristics of the audio signal. Please refer to Figure 1 below for the MCFF diagrams.

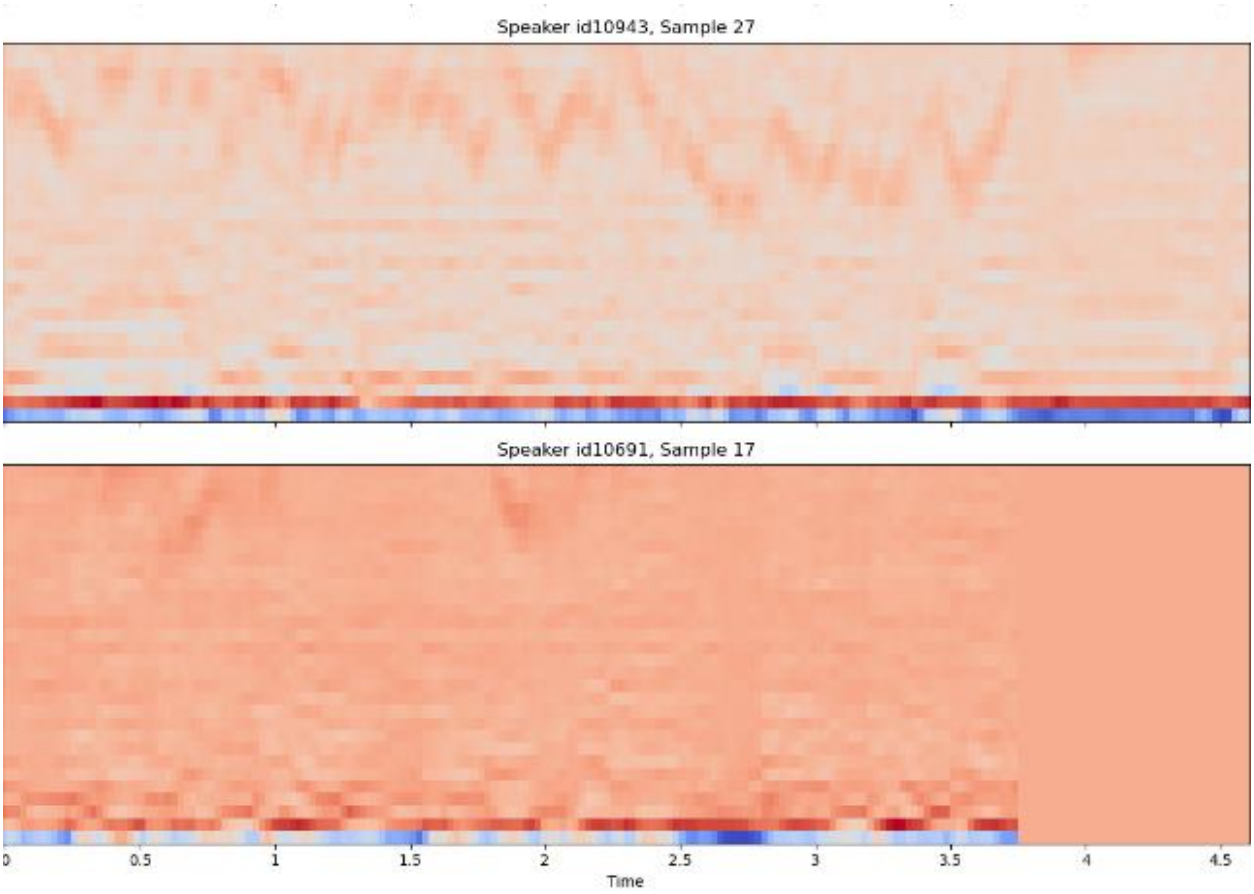

Figure 1: MCFF VoxCeleb1 Images

Another notable function is “sample_audio_files”, designed to sample a specified number of audio files from a directory, allowing for a manageable and representative subset of data to be processed. This function is particularly useful when working with large datasets like the one we have, as it enables efficient handling of data.

Overall, the preprocessing phase demonstrates a comprehensive approach to audio data manipulation. It involves transforming raw audio into a format suitable for analysis (mono conversion), extracting meaningful features (MFCCs and spectrograms), and efficiently handling large datasets through sampling.

Modeling Workflow:

Our approach to neural network modeling was systematic:

Design: Leveraging Keras, we designed the CNN and LSTM model, determining the architecture, layers, neurons, activation functions, and other hyperparameters suitable for the MFCC data.

Train: With the preprocessed data at hand, the training phase focused on adjusting weights and biases to minimize the difference between predicted and actual outputs.

Validate: After training, it was crucial to validate the model on a new subset of data to ascertain its generalization capability and ensure that it doesn't overfit to training samples.

Test: Following validation, we tested the model using a distinct set of unseen audio data. This stage is vital for assessing the model's practical applicability and ensuring consistent performance across varied patterns. Our test results serve as the definitive measure of our model's efficacy.

Model Results

The CNN-LSTM model constructed in the given function is a hybrid neural network that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, aimed at classifying voice data. The model was constructed using numerous techniques and the relation with the success (or lack thereof) will be described below. Please refer to figure 2 for a visual representation of the model.

Convolutional Layer: The initial Conv1D layer, characterized by a specified number of filters and a kernel size, is designed to extract features from the input data. In the context of audio processing, this layer can identify and capture various patterns in the sound wave, such as frequency and amplitude changes. Batch normalization following the convolutional layer speeds up training. The MaxPooling1D layer then reduces the dimensionality of the data, which helps in reducing computational complexity and overfitting by abstracting the features. The Dropout layer following pooling introduces regularization by

randomly dropping a proportion of the network connections (as defined by the dropout rate), which prevents overfitting as well.

LSTM Layer: After the convolutional layers, the model employs a long short-term memory (LSTM) layer. LSTMs are a type of recurrent neural network (RNN) particularly effective in handling sequential data, making them suitable for time-series data like audio. The LSTM can capture temporal dependencies and patterns in the audio signal, which are crucial for understanding speech and sound dynamics. The additional Batch Normalization and Dropout layers here serve to improve training efficiency and prevent overfitting.

Dense and Output Layers: Following the LSTM, the model has a Dense layer with 64 units and ReLU activation. This layer is used to interpret the features extracted and processed by the previous layers. The final output layer is another Dense layer with a softmax activation function, which is typical for multi-class classification tasks. The softmax function outputs a probability distribution over the target classes. Despite this architecture, the model's suboptimal performance (around 22% accuracy) in predicting audio or voices could be due to several factors:

Data Quality and Preprocessing: If the input data is not well-preprocessed or lacks variety, the model might struggle to learn the necessary features. For audio data, aspects like noise removal, normalization, and proper feature extraction (like MFCCs) are critical.

Model Complexity and Overfitting: While the model is complex enough to capture a wide range of features, it might also be prone to overfitting, especially if trained on a limited amount of data. The dropout layers aim to mitigate this, but they might not be sufficient depending on the dataset size and diversity.

Hyperparameter Optimization: The choice of hyperparameters such as the number of filters, kernel size, pool size, LSTM units, and dropout rate significantly impacts the model's performance. Improper tuning of these parameters can lead to suboptimal learning.

In summary, the constructed CNN-LSTM model is theoretically sound for audio classification tasks, combining the feature extraction capabilities of CNNs with the sequential data handling of LSTMs. However, its performance is heavily influenced by the quality of the input data, the complexity and tuning of the model, and the inherent challenges of audio data classification. In this particular research project, the model’s success will be much more dependent on the modification of the data preprocessing stages compared to the actual model architecture.

Conclusions

This level of attention to preprocessing underpins the analytical capabilities required for sophisticated audio analysis and the development of machine learning models tailored for tasks such as speech recognition and audio classification. By focusing on extracting high-quality features from audio, the project sets a precedent for the kind of meticulous data preparation that is essential for reliable signal processing.

Enhanced feature extraction and a deeper understanding of audio file characteristics could significantly bolster the accuracy of voice recognition systems, sound classification, and surveillance technologies. The intelligence community relies heavily on the precision of such systems to interpret audio data, often captured in less-than-ideal circumstances, making the quality of data preprocessing a critical factor. Such advancements would not only elevate the operational effectiveness of intelligence efforts but also increase the speed and reliability with which audio data can be analyzed, potentially transforming the landscape of intelligence gathering and analysis.

Incorporating these enhancements in data processing would ensure that the underpinnings of audio analysis are robust and comprehensive. This foundational strength is vital for the intelligence community, where the stakes are high, and the accuracy of audio analysis can have significant implications. By continuing to prioritize the refinement of audio preprocessing techniques, future research can provide the intelligence community with more powerful tools to meet their unique and challenging objectives.

Appendix

Snyder, Daniel, Daniel Garcia-Romero, Gregory Sell, David Povey, and Sanjeev Khudanpur. 2018. "X-vectors: Robust DNN Embeddings for Speaker Recognition." arXiv:1710.10467.

Wan, Li, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. "Generalized End-to-End Loss for Speaker Verification." arXiv:1710.10467.

Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. 2017. "VoxCeleb: A Large-Scale Speaker Identification Dataset." arXiv:1706.08612.

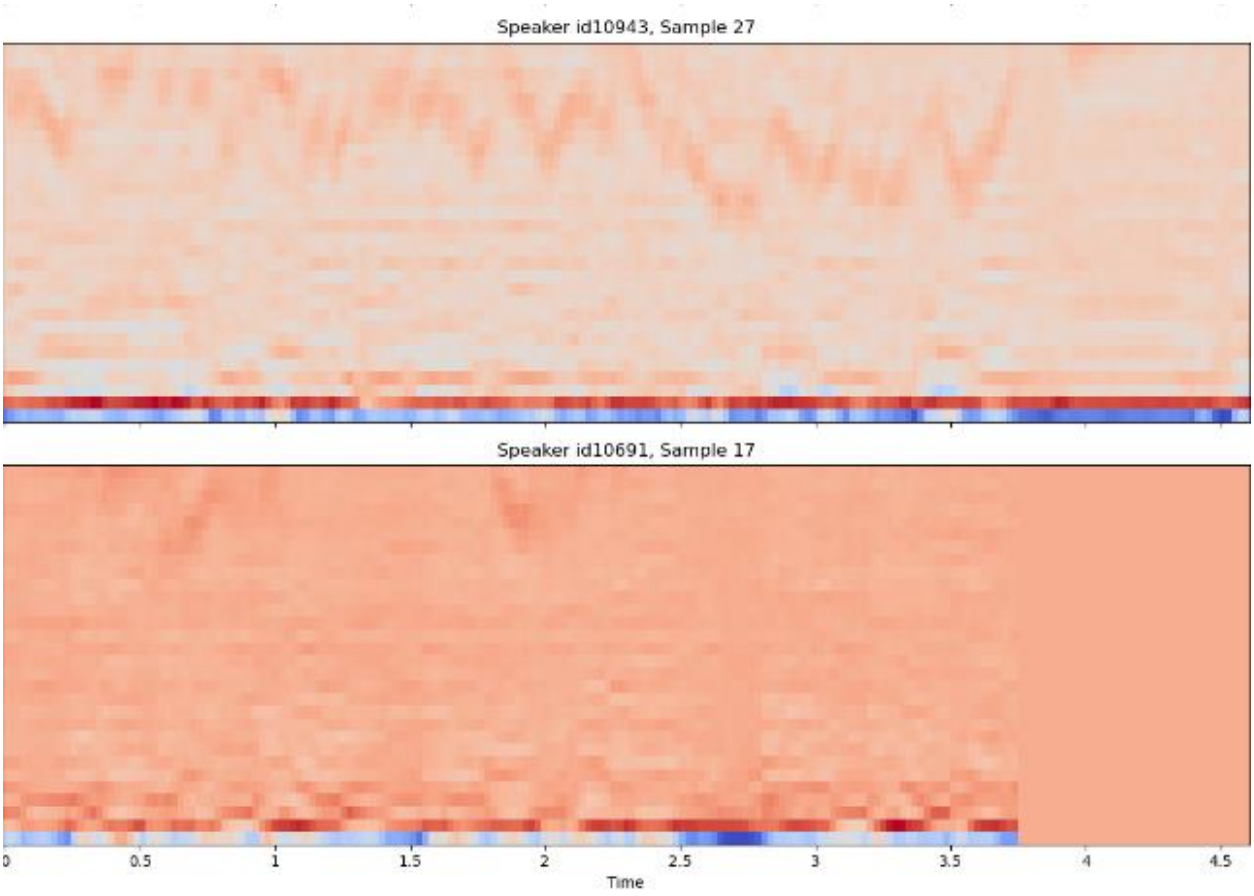

Figure 1: MCFF VoxCeleb1 Images

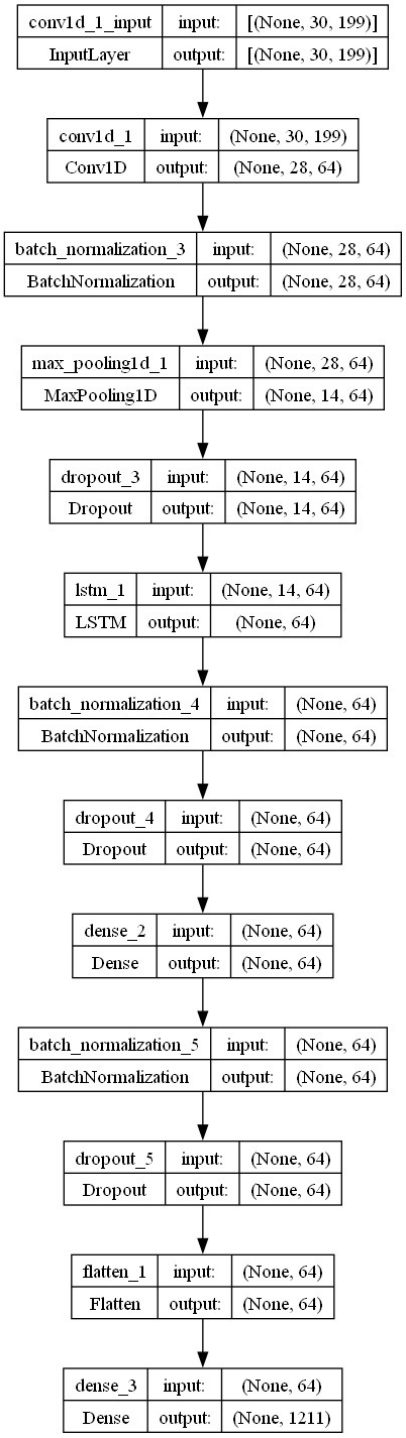

Figure 2: CNN LSTM Architecture (Model 1)