Egyptian Dialect Voice Recognition Project Documentation

**Project Overview:**

This project focuses on developing a machine learning model capable of recognizing and classifying Egyptian dialect voice recordings. The model is trained on audio files with corresponding transcripts using a convolutional neural network (CNN). The primary goal is to accurately transcribe spoken words and phrases in the Egyptian dialect.

**Model Description:**

**Preprocessing:**

The preprocessing pipeline is crucial for transforming raw audio data into a format suitable for the CNN model. The steps involved are:

1. **Loading WAV Files**: Audio files in WAV format are read and decoded. The audio is then resampled to a consistent sample rate of 12 kHz to standardize the input data.

2. **Generating Spectrograms**: Each audio file is trimmed or padded to ensure a fixed length of 70,000 samples. This standardized length allows the model to process the data uniformly. The audio is then converted into spectrograms using Short-Time Fourier Transform (STFT). Spectrograms represent the intensity of frequencies over time and are used as input features for the CNN.

**Data Preparation:**

The dataset consists of audio files and their corresponding transcripts. During preparation:

- Audio files are preprocessed to generate spectrograms.

- The dataset is split into training and testing sets.

- The model expects a fixed input shape, which is determined by the spectrogram dimensions.

- The number of classes is determined based on the number of unique transcripts in the training data.

**Model Architecture:**

The model is a convolutional neural network (CNN) designed with the following layers:

1. **Convolutional Layers**: These layers extract features from the input spectrograms. The model includes multiple convolutional layers with ReLU activation functions to capture complex patterns in the audio data.

2. **Max Pooling Layers**: These layers reduce the spatial dimensions of the data, which helps in lowering computational requirements and controlling overfitting.

3. **Dense (Fully Connected) Layers**: These layers perform classification based on the features extracted by the convolutional layers. The final dense layer uses a softmax activation function to output probabilities for each class.

**Achieving Best Results:**

The model's performance depends on several factors:

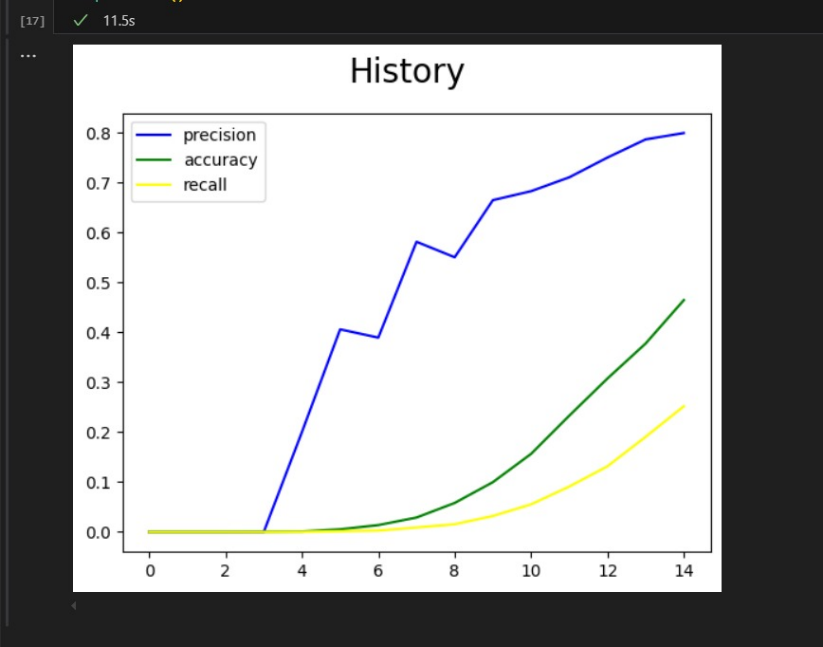
- **Data Augmentation and Preprocessing**: Ensuring consistent audio lengths and converting to spectrograms effectively capture the essential features of the audio data.

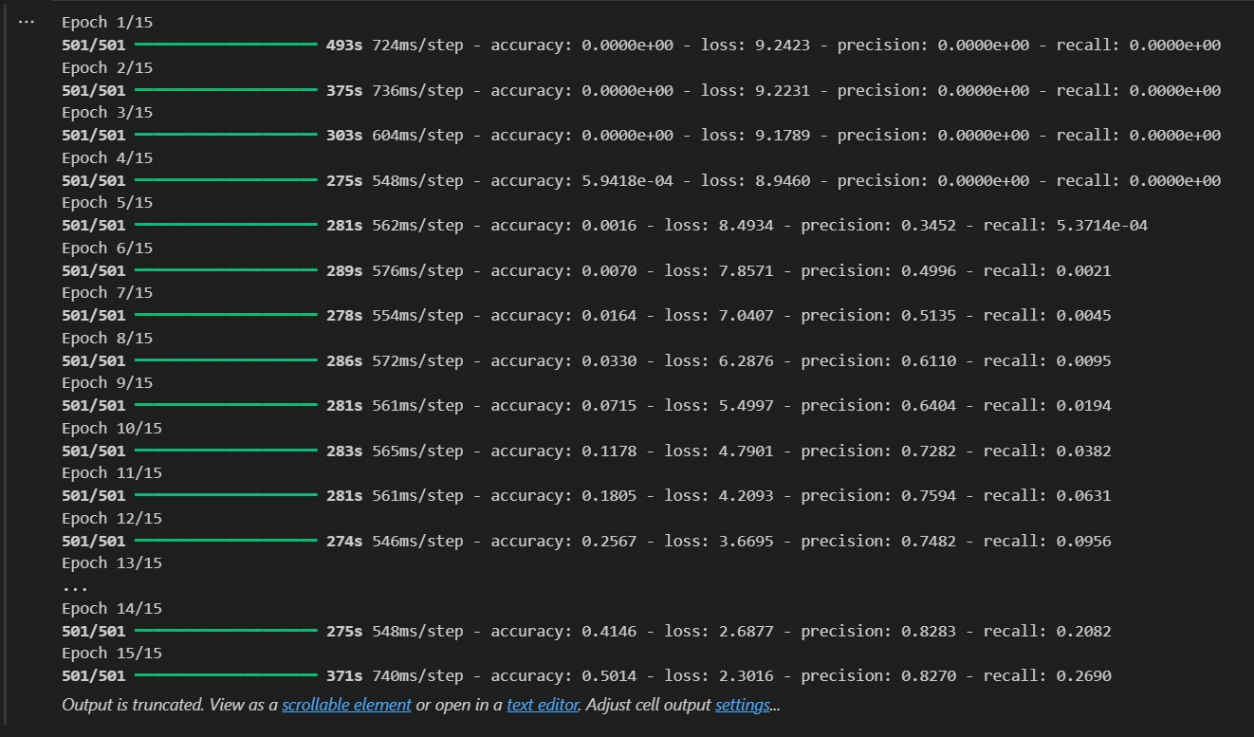
- **Model Architecture**: The chosen architecture, with multiple convolutional and dense layers, is well-suited for handling the complexity of voice data.

- **Training Parameters**: The model is trained using the Adam optimizer and categorical cross-entropy loss function. Metrics such as accuracy, precision, and recall are monitored during training.

**Training Results**

The training process involves fitting the model on the preprocessed training data. The model's performance is evaluated using metrics such as precision, accuracy, and recall. The training history is visualized to understand how these metrics evolve over time, helping in assessing the model's learning progress and identifying any overfitting or underfitting issues.





**Predictions**

Once trained, the model is used to make predictions on the test dataset. The predicted transcripts are compared with the actual transcripts to evaluate the model's accuracy. The results are saved for further analysis.

**Saving the Model**

After training, the model is saved in a format that allows for future use without retraining. This enables deploying the model in production or for further experimentation and improvement.

**Conclusion**

This project demonstrates the effective use of a convolutional neural network for classifying Egyptian dialect voice recordings. By preprocessing the audio data into spectrograms, designing a robust CNN architecture, and training with appropriate parameters, the model achieves notable performance in recognizing and transcribing spoken words in the Egyptian dialect. This approach can be extended to other dialects or languages with similar preprocessing and model training techniques.