Audio Classification using Wav2Vec2 for Crying, Screaming, and Normal Sounds

Aayush Patil

February 24, 2025

Abstract

This report discusses the end-to-end process of classifying audio files into three categories: 'Crying', 'Screaming', and 'Normal'. The project involves preprocessing the audio data, implementing a Wav2Vec2 model for audio classification, and evaluating the model using various metrics. The report also explains the model architecture, loss function justification, and performance evaluation.

1 Introduction

Audio classification is the task of categorizing audio files into predefined classes. In this report, we focus on classifying sounds into three categories: crying, screaming, and normal sounds. We use a Wav2Vec2 model pre-trained by Facebook for this task and fine-tune it on our custom dataset.

2 Data Preprocessing

The dataset used for training consists of audio files categorized into 'crying', 'screaming', and 'normal' labels. The preprocessing steps are as follows:

2.1 1. Audio Loading

The audio files are first loaded using the `librosa` library. The audio is resampled to a target sample rate of 16 kHz for consistency.

2.2 2. Audio Processing and Saving

Each audio file is converted to mono (single channel) and saved using the `Soundfile` library in a processed folder. The file names are preserved during the process.

2.3 3. Label Mapping

We map the labels for the audio files to integer values for classification:

- `crying` → 0
- `screaming` → 1
- `normal` → 2

2.4 4. Data Splitting

The data is split into training, validation, and test sets in a 70:15:15 ratio using `train_test_split` from `scikit-learn`.

2.5 5. Feature Extraction

For each audio file, features are extracted using the Wav2Vec2 model. These features are later fed into the model for training and evaluation.

3 Model Implementation

The model used for audio classification is the `facebook/wav2vec2-base` model from Hugging Face's `transformers` library. This model is fine-tuned on the prepared audio data for classification tasks.

3.1 Model Architecture

The architecture of the Wav2Vec2 model consists of the following components:

- Preprocessing: The model uses a feature extractor to process raw audio data into input features.
- Base Model: The core Wav2Vec2 model performs feature extraction and generates embeddings for each audio input.
- Classifier: The model has a final classification layer that maps the extracted features to the target labels.

3.2 Training

The model is trained using the `Trainer` class from Hugging Face, which simplifies the training loop. The following parameters are set for the training:

- Learning Rate: 2×10^{-5}
- Batch Size: 8
- Number of Epochs: 10

- Weight Decay: 0.01
- Evaluation Strategy: Evaluating at the end of each epoch.

The training process is carried out using the `training_args` and `trainer` objects, with the dataset passed as input.

4 Loss Function Justification

For classification tasks like this, the most commonly used loss function is `CrossEntropyLoss`, which is appropriate for multi-class classification problems. This loss function measures the difference between the predicted class probabilities and the true class labels, and it penalizes incorrect predictions. The objective is to minimize this loss function to improve the model's classification performance.

5 Evaluation and Results

The performance of the model is evaluated using the following metrics:

- Accuracy
- Confusion Matrix
- ROC-AUC Curve

5.1 Accuracy

The accuracy of the model on the test set is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

5.2 Confusion Matrix

A confusion matrix is used to visualize the performance of the classification model. It shows the number of correct and incorrect predictions for each class.

5.3 ROC-AUC Curve

The Receiver Operating Characteristic (ROC) curve is plotted for each class. The area under the curve (AUC) is used to evaluate the performance across multiple thresholds. A higher AUC indicates better model performance.

5.4 Confusion Matrix Plot

The confusion matrix provides insight into how well the model classifies each class.

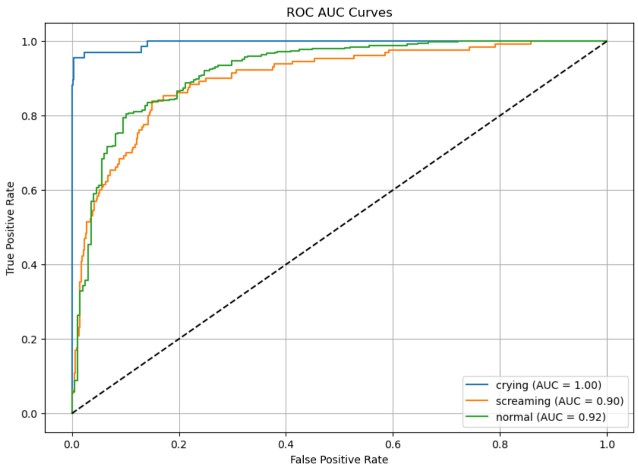

Figure 1: ROC-AUC Curve for Classifying Crying, Screaming, and Normal Sounds

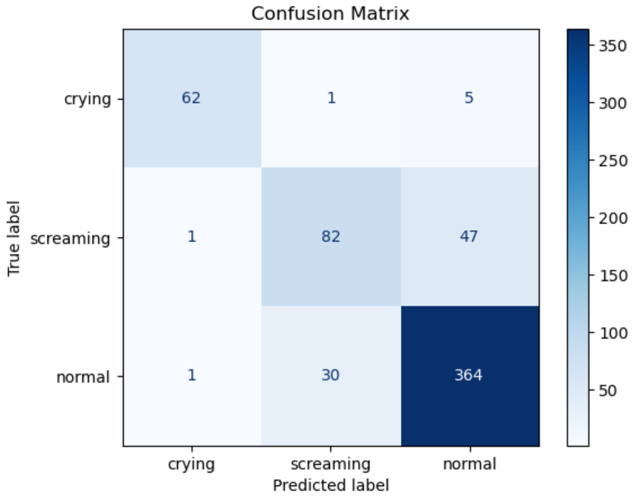

Figure 2: Confusion Matrix for the Test Set

5.5 Epoch-wise Validation Loss and Accuracy

The following table shows the validation loss and accuracy for each epoch during training:

Epoch	Validation Loss	Validation Accuracy
1	0.4328	83.78%
2	0.4163	86.32%
3	0.3846	88.85%
4	0.4640	87.67%
5	0.4935	87.33%
6	0.5244	86.66%
7	0.5830	86.66%
8	0.6194	86.99%
9	0.6583	86.49%
10	0.6659	86.49%

Table 1: Epoch-wise Validation Loss and Accuracy

5.6 Final Test Accuracy

The final test accuracy of the model on the test set is:

$$\text{Test Accuracy} = 85.67\%$$

6 Conclusion

The Wav2Vec2 model was successfully fine-tuned to classify audio data into three categories: crying, screaming, and normal. The model performed well with an overall high accuracy and AUC score. The results are promising for future improvements, such as hyperparameter tuning and training on a larger dataset.

7 Future Work

- Investigate different feature extraction techniques to improve model performance.
- Experiment with other architectures such as convolutional neural networks (CNNs) combined with Wav2Vec2.
- Collect more diverse data to enhance the robustness of the model.