Detecting Emotion from Greek Speech Audio Using a Fine-tuned Wav2Vec2 Model

Kiel Hizon
(SN: 2013-17614)
Electrical and Electronics Engineering Institute
University of the Philippines Diliman

ABSTRACT

Fine tuning of publicly released large models has been a popular approach in deep learning. In this project, we fine-tuned the Greek tuned version of XLS-R was further fine tuned to classify emotion from Greek utterances. Using only about 400 data samples across 5 emotions, the model was able to achieve 88% accuracy.

1. INTRODUCTION

The large amount of data needed to train machine learning models has been a hindrance for developers with limited resources. In the domain of natural language processing, and speech processing, this is shown when trying to adapt a model to your local languages. Fine-tuning and transfer learning has been a very powerful advancement in the field in the recent years. Large research laboratories, release their models which are trained on terabytes of data using clusters of compute units publicly. Individual developer can then fine tune those models using smaller data sets and less compute power and time to their specific needs while achieving good results.

In this project, we fine-tuned a speech-to-text model already trained for the Greek language, to an even smaller data set to classify emotion from Greek audio. Using only 400 audio samples for training, the model was able to achieve 88% accuracy.

2. RELATED WORK

Hand-crafted features Earlier machine learning models require a combination of handcrafted features such as zero-crossing, amplitude envelope, significant frequency components etc. as inputs to machine learning models such Random forests, and Support Vector Machines. In this approach, the research is optimizing two things: feature extraction, and the model.

Deep Learning Deep learning models, "learn" the features from the data. Convolutional Neural Networks, for example, which was very popular in image recognition, extract features at sub-regions of the image which enable this model to preserve local context. Audio signals can be converted to an image-like representation using spectrogram or Mel Frequency Cepstral Coefficients. In this form, it can be used as an input to a CNN to perform audio-classification.

Transformers Transformers take their inputs as a sequence of tokens. Using its attention mechanism it can learn to which tokens should are more significant in that given sequence. For Natural language processing, this is very useful to represent context in sentences or paragraphs. Transformer models has since been adapted to other domains such as image, audio, and video. Additionally, research laboratories publish their models and weights on multiple platforms and frameworks which make adoption of state-of-the-art models relatively easier.

Wav2Vec2 XLS-R Facebook AI released XLS-R as their large pre-trained multilingual speech model. It is pre-trained on 436k hours of speech from 128 languages. Its pre-training data set included VoxPopuli, MLS, CommonVoice[1], BABEL, and VoxLingua107. They have released three sizes of the model, 300M, 1B, and 2B. Which stand for the number of parameters of the model. Furthermore, language specific models have been developed from this model, which includes the Greek model that was used in this project.

3. METHODOLOGY

Data Set The Acted Emotional Speech Dynamic Database is a publicly available speech emotion recognition data set. It contains about 500 utterances of acted emotional speech in the Greek language.The database utterances with five emotions: anger, disgust, fear, happiness, and sadness[3]. It was originally constructed to help reduce the absence of publicly available high-quality speech data sets in Greek.

Architecture In this project we started with an XLS-R that is already further trained on Greek speech data. Since the model was originally intended for speech-to-text application, we attach a classifier head of neural networks to the output of the original model. The final output will be probabilities for each of the five classes, the highest one will be the prediction of the model. The project and code was based on the google colab notebook by Mehrdad Farahani[2]. Minor changes introduced included the conversion of the training section from a notebook to a script such that it can be run without the risk of interruption.

Training 20% of the data was reserved for testing. Audio processing on the data is limited to resampling to 16kHz, and normalization. The remaining training data is passed through a processor (already pre-trained for Greek) which acts as feature extractor and tokenizer. The XLS-R-Greek was trained with its feature extraction segment frozen. These sections are CNNs which are already trained on bigger data sets and we don't want to skew them to our limited data.

Additionally, this would also reduce training time. One pass on the training data (epoch) took 60 training steps. We trained the model for 3 epochs (180 training steps).

4. RESULTS

Since the categories are fairly balanced, simple accuracy was deemed a sufficient metric. The loss of the model for the training data is recorded. Every 20 steps, the model is evaluated on the test set, and accuracy is measured. Note that this evaluation step does not affect the weights of the model. Figure 4 below show the decrease on the training loss and increase on the test accuracy as the steps progress. After 180 steps, the evaluation accuracy was at 88%.

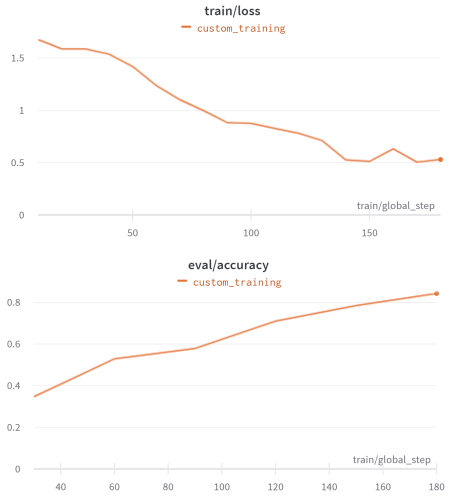

Figure 1: Train loss and Test accuracy per training step

On further testing, we saw that "happiness" seem to be a weakness of the label. The confusion matrix below on Fig. 4 show that although the overall performance of the model is good, some happiness labels are predicted to be angry. Some disgust labels are also labeled as happiness.

5. CONCLUSION

In this project, we have demonstrated the ability to generate specific-language and specific-task models using small data sets using transfer learning. By taking advantage of the pre-trained XLS-R, and then the further tuned XLS-R Greek, we were able to build an emotion classifier model using only 400 utterances for training data. Further research can done on adapting this approach on other low resource languages and distiling the models into smaller models that more apt for end device deployment.

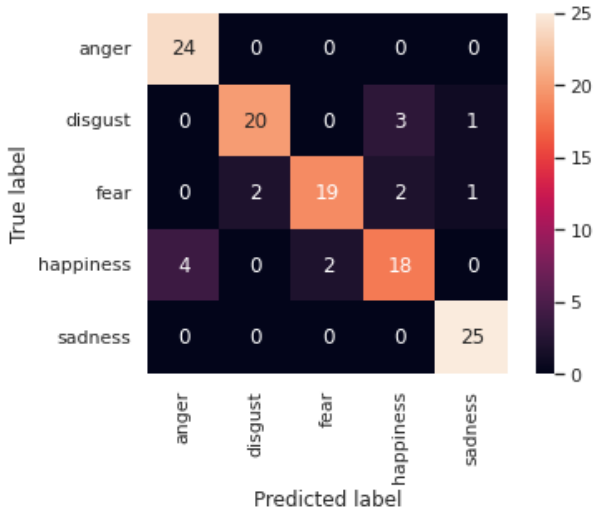

Figure 2: Confusion Matrix

References

[1] Rosana Ardila et al. "Common voice: A massively-multilingual speech corpus". In: *arXiv preprint arXiv:1912.06670* (2019).

[2] Mehrdad Fahrahani. *WAV2VEC2-xlsr-greek-speech-emotion-recognition*. URL: `https://huggingface.co/m3hrdadfi/wav2vec2-xlsr-greek-speech-emotion-recognition`.

[3] Nikolaos Vryzas et al. "Speech emotion recognition for performance interaction". In: *Journal of the Audio Engineering Society* 66.6 (2018), pp. 457–467.