Using Large-Pretrained CV Transformers for Speech-Audio Image Spectrogram Representations: Emotion Recognition

Yair Shachar
Anthony Le
Omer Benyshai

Stanford
CS 231N
Final Report

Abstract

Emotion classification serves as a useful downstream task of large audio models. For instance, It has the capability of improving further downstream tasks like sentiment analysis by providing richer feature representations. This paper aims to increase performance in emotion classification through a combination of attention based and CNN architectures on audio spectrogram images. Recent works have also shown the use of CNN architectures on audio spectrogram images performing at state-of-the-art level, while attention based neural architectures have also led to benchmark performance increase across several domains. We provide in-depth analysis on state-of-the-art deep neural network architectures alongside benchmarking some of these architectures on a downstream task, mainly emotion classification from audio and/or audio spectrograms. This project shows how large pre-trained attention based models outperform standard baseline models in the task of emotion recognition and classification from audio image spectrogram features. We show our best performing model as the Audio Spectrogram Transformer that receives a macro F1 score of and a validation accuracy of 85%. This shows improved performance of large attention based models over other ML techniques for emotion recognition.

1. Introduction

We will be investigating state of the art end to end systems that classify specific emotions from audio spectrograms. One of the largest uses of spoken and natural language processing is audio classification, which can be used to automatically detect domain-specific audio classes from audio files. We aim to compare more recent transformer/attention based models against state of the art CNN models to see which model is best suited for this domain specific task of emotion classification. This problem is interesting because there hasn't been much research regarding applying these large pre-trained CV models to audio spectrogram data in the context of emotions. We think that it will be interesting to see if transformer based models can find more nuanced patterns within the spectrograms to provide state of the art emotion classification. To start, we began with a literature survey around the use of deep neural network architectures on audio spectrogram data, for brevity we have included a list of some of the papers we read. Many papers were exploring the idea of using large-pretrained computer vision models on audio spectrograms for downstream tasks such as audio classification or even speech recognition and translation.

1.1. Language: English

1.2. Problem Statement

The problem of audio classification has been previously solved by using mel-spectrogram features and MFCC's and classifiers based on GMM's or SVM's. More recent approaches leverage deep neural networks and large pretrained models to solve this problem. Unfortunately, most of the work done on audio classification is done on domains outside of spoken human language, as the two largest benchmarks for this overall problem are AudioSet and ESC-50 which are open domain classes and environmental animal classes, respectively. We plan to use a large pre-trained audio classification/image classification model and finetune this model to our emotion classification training dataset. Because of this, it could be interesting in exploring multiple baseline models on emotion classification. We will test methods we learned in class such as SVM's and DNN's on basic audio spectrogram representations, and then will extend to larger models shown to have greater success classifying audio image spectrograms. We hypothesize that a larger, more complex model, even if pre-trained on non-domain specific audio/image data, will perform significantly better on low-resource emotion classification

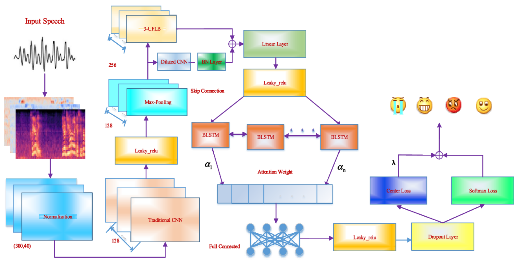
Touching on the problem of low resource tasks, the currently available emotion classification audio datasets are quite small relative to other audio classification tasks; Where most emotion classification audio datasets have <100 hours of audio, large audio classification datasets have >5000 hours of audio. We will be using the RAVDESS dataset, which will be discussed next.

1.3 Related Works

Deep Learning Methods [6]

The major deep learning techniques used for the task of SER (Speech Emotion Recognition), are CNNs, LSTMs and attention mechanisms. Here we will mention the outstanding ones in each category.

CNN Architectures [5]



Zhang et al. [11] have developed an Emotion recognition system based on AlexNet and fine-tuned using samples from EMO-DB. Using this system, they can recognize three classes of emotions (angry, sad, and happy) plus a neutral category. They have demonstrated that their system can achieve accuracies over 80% with EMO-DB, about 20% more than the baseline SVM standard of that time. The capability of deep convolutional neural networks comes with the cost of exponentially more variables to tune, and this means more samples are needed to train the system. However, in SER, usually, the numbers of the samples are limited to thousands. This makes solutions based on deep convolutional networks more prone to overfitting.

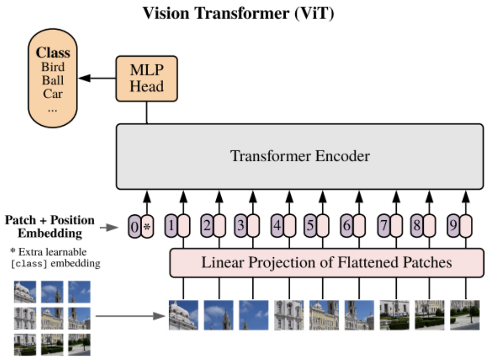
LSTMs

Xie et al. [9] introduced a system based on two layers of modified LSTMs with 512 and 256 hidden units, followed by a layer of attention weighting on both time dimension and feature dimension and two fully connected layers at the end. They have experimented with five combinations of their proposed methods, LSTM with Time attention, LSTM with feature attention, LSTM with both time and feature attention, LSTM with modified forget gate, and LSTM with modified forget gate and time and feature attention. Additionally, as the results on their English speech dataset eNTERFACE, they have reached 89.6% UAR accuracy which they claim is the best result on that dataset.

The exciting capabilities of LSTMs, however, come with the cost of more processing power and exponential memory requirements. They also, similar to CNNs, need a vast number of training samples to tune their large number of variables.

In the attention mechanism, the classifier regards the given samples' specific locations based on the attention weights assigned to each part of the data, which contains an emotionally salient portion, thereby exploiting the non-uniform distribution of emotion over the utterance for every sample. However, contemporary works demonstrated the power of the attention mechanism by combining it with CNN and LSTM based architectures. We will discuss what we found for transformers in our literature review next.

AST [3]


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Yuan et al. [3] show that the use of vision transformers [10] for performing sequential image processing tasks can be used for sequential processing of audio spectrograms for audio classification tasks and they do so without the use of any CNN mechanism.

In their transformer architecture, they split each spectrogram into a sequence of 16x16 patches which were then projected into 1D embedding vectors concatenated with learnable positional embeddings and added a classification token that was embedded as well. All of these embeddings are feeded into a transformer encoder and the output of the classification token is then feeded into a linear layer which performs the classification.

In their work, they show that no tuning of the internal architecture is needed even when processing different lengths of audio data which makes their architecture very inviting to experiment with.

2. Data

Ravdess:

The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and the song contains calm, happy, sad, angry, and fearful emotions. The data split we used only consisted of speech actors, no songs, and contained over 7,000 <5 second audio clips.

Due to the simplicity and how well classified this dataset was, we decided to use it for our experiments. In our experiments we used only the raw audio of the speech and disregarded the video data.

Other Datasets:

Other datasets that we experimented and considered were:

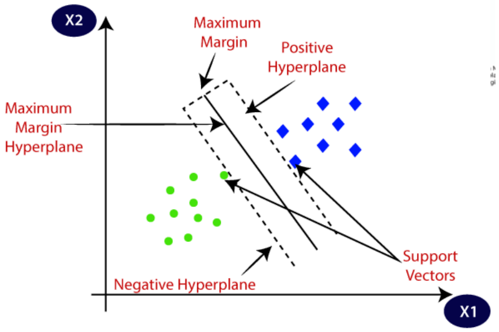
- **CREMA-D:** A crowd-sourced emotional dataset consisting of multi modal examples. Contains around 7,000 labeled utterances from actors.
- **EMODB:** 535 utterances across seven emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral.
- **CMU-Multimodal SDK:** 65 hours of annotated video from more than 1000 speakers, 250 topics, and 6 emotions.

We chose to use the RAVDESS dataset because it was the most accessible out of the few options we had. RAVDESS contained emotional states that we found to be the most inclusive and general. In the future, we would like to extend our work to other emotional classification benchmark datasets, or perhaps even pre-process and aggregate the aforementioned datasets into a large emotional classification benchmark.

3. Methods

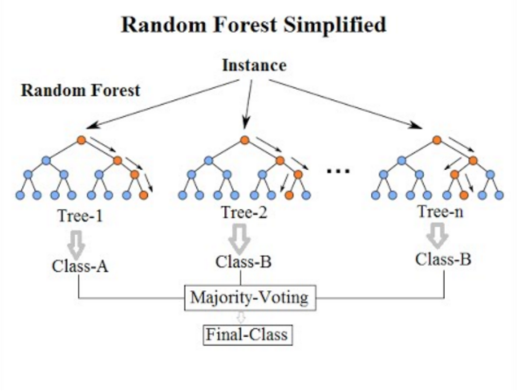
3.1 Baselines

Support Vector Machines



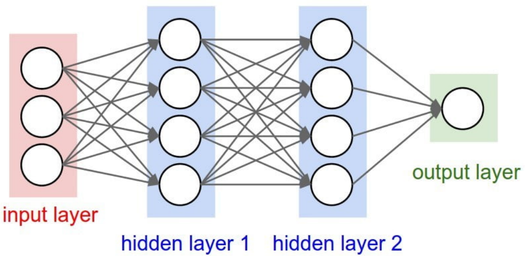
Support vector machines are a primitive model used as a baseline in this project. In order to see how low resource models performed we used a basic SVM to predict emotion on the mel-spectrogram image features of our training data audio. Support vector machines have the objective of separating feature vectors in N-dimensional space with the use of hyper-planes. These hyper planes form decision boundaries that are used to classify the N-dimensional feature representations.

Random Forest



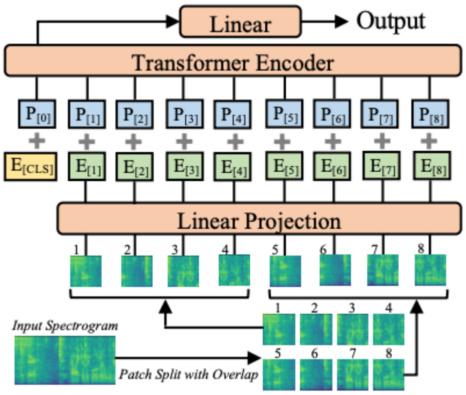
Random forest classifiers were also used as a baseline method for the task of emotion classification. Random forest classifiers use a large number of individual decision trees that come together and take a majority vote in ensembling to predict the correct class. The low correlation between each of the individual trees makes this work by allowing an ensembling algorithm to help the classifier protect itself from each of the individual errors made by some of the trees. Our model uses 100 individual estimators to predict the 8 emotion classes contained in RAVDESS.

Deep Neural Networks



As we have seen in class, deep neural network architectures can perform much better than simple random forest or svm's because the hidden layers and weights are able to capture much more complex relationships between the feature representations. We use a 5 layer fully connected neural network with Adam optimizer, batch normalization, dropout, and a softmax classifier. This is a standard architecture for a neural network classification task.

3.2 Large-Pretrained Transformer Models



Throughout our literature survey we found that recent developments in transformer models may be able to outperform large pre trained CNN's in the task of audio classification. These large CNN models aim to learn direct mappings from the audio spectrogram features into desired labels. We hypothesized that the sequential nature of LSTM's with large pre trained CNN's would do great at this task, but that transformer models may have the edge in this specific audio related task. We believed this because transformer and attention based models have been shown to perform well in downstream NLP tasks [8]. In an ideal setting we would have loved to compare large pre-trained models with LSTMS's, but due to time, virtual machine memory, and GPU constraints we chose to implement only the transformer model. Since audio is sequential in nature, we proposed using a large transformer model that had been pre-trained on similar data, as we know pre-training with domain specific data is usually known to boost performance with domain-specific tasks.

To this end we used the Audio Spectrogram Transformer model [3] as our base model. This model is pre-trained on the Audioset dataset which contains over 2 million 10-second sound and video clips from youtube with 632 different classes of audio in their ontology. We hoped that this pre-training on audio image spectrograms would help the model in the downstream task of predicting emotions from audio image spectrograms.

Our model uses a canonical transformer model where inputs are tokenized, positional embeddings are added, and the sequences are fed into the mode. The model first transforms our audio into image mel-spectrogram feature representations with 128-dimensional log Mel filterbank features. Then, these sequences are splitted into 16 x 16 patches and flattened to produce our input tokens, similar to convolution. Additionally, the positional embeddings are added here. Inputs are passed into the encoder and linear layer with softmax classifier will learn the classification task.

The novel idea of this approach is that Image-Net and Audioset data is used to pretrain ViT (Vision Transformer), allowing us to use less data. As we have talked about previously, this helps with our task in that domain-specific data for emotion classification is sparse.

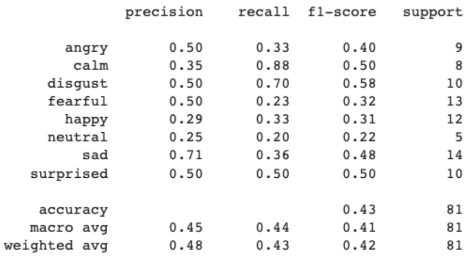
4. Experiments

4.1 Metrics

For our experiments we will be evaluating classification with 3 metrics. First, we will have our validation and test accuracy measures showing the overall correctness of our predictions on the 8 emotion classes. Secondly, we will present precision, recall, and F1 scores for each class's predictions. This will help us identify which classes the model is struggling with and in what ways we might be lacking in both our data or model implementation. Lastly, we will use the AUC ROC curve metric for our AST model with a one vs all methodology to compare how well each class is discriminated by the model.

4.1 Baselines

Support Vector Machines:



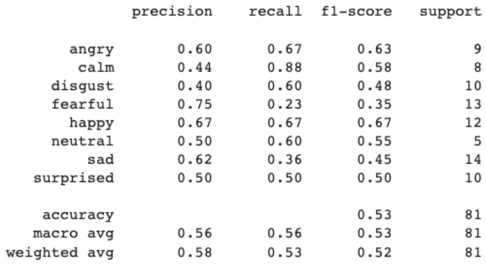
```
              precision    recall  f1-score   support

       angry       0.50      0.33      0.40         9
        calm       0.35      0.88      0.50         8
     disgust       0.50      0.70      0.58        10
     fearful       0.50      0.23      0.32        13
       happy       0.29      0.33      0.31        12
     neutral       0.25      0.20      0.22         5
         sad       0.71      0.36      0.48        14
   surprised       0.50      0.50      0.50        10

    accuracy                           0.43        81
   macro avg       0.45      0.44      0.41        81
weighted avg       0.48      0.43      0.42        81
```

These are the results from our support vector machines with a linear kernel. It received the lowest accuracy and macro F1 score as expected.

Random Forests:



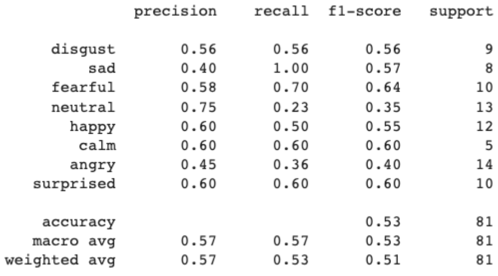
```
              precision    recall  f1-score   support

       angry       0.60      0.67      0.63         9
        calm       0.44      0.88      0.58         8
     disgust       0.40      0.60      0.48        10
     fearful       0.75      0.23      0.35        13
       happy       0.67      0.67      0.67        12
     neutral       0.50      0.60      0.55         5
         sad       0.62      0.36      0.45        14
   surprised       0.50      0.50      0.50        10

    accuracy                           0.53        81
   macro avg       0.56      0.56      0.53        81
weighted avg       0.58      0.53      0.52        81
```

These are the results from our random forest classifier with 100 individual estimators with majority vote ensembling. We saw that this was the best performing number of individual estimators for this classification task as we tried several other hyperparameters for this. Surprisingly, this baseline method performed just as well as the simple neural network.

Simple Neural Network:



```
              precision    recall  f1-score   support

     disgust       0.56      0.56      0.56         9
         sad       0.40      1.00      0.57         8
     fearful       0.58      0.70      0.64        10
     neutral       0.75      0.23      0.35        13
       happy       0.60      0.50      0.55        12
        calm       0.60      0.60      0.60         5
       angry       0.45      0.36      0.40        14
   surprised       0.60      0.60      0.60        10

    accuracy                           0.53        81
   macro avg       0.57      0.57      0.53        81
weighted avg       0.57      0.53      0.51        81
```

These are the results from our simple neural network trained on different hidden layer numbers and sizes, different optimizers, and different learning rates. We found that the best performing simple neural network consisted of 5 hidden layers with size 300 hidden units, Adam optimizer, a learning rate of 1e-3. We also added batch normalization and a dropout rate of 0.5. We saw the limitations of our data when implementing this simple neural network. As we increased the number of training epochs, the loss and validation accuracy would plateau, signaling to us either that the model is not complex enough or that we have a shortage in data. Either way, our unique pre-trained transformer model aimed at combating these two problems.

4.2 Large Pre-trained Transformer Models

We trained our audio spectrogram vision transformer model with cross entropy loss for 25 epochs with a learning rate of 1e-5 using a learning decay rate of 0.85 starting at each epoch, starting at epoch 5. We set our max length for samples to be 512 samples with 128 mel fbank features, additionally we masked 24 out of the 128 mel fbanks and up to 96 of the time frames. The model was pre-trained on both Audioset and ImageNet. The model performed exceptionally well relative to the baselines and surrounding literature. Below are our recorded eval metrics on the validation set. As we can see, we have a very high accuracy and f1 score, with much worse precision than recall for most of the classes. This might mean that our model over guesses for the correct class when predicting that specific class. A model with no false negatives has a recall of 1, meaning it correctly identifies that class. Since our model had an average of 1, this means that there were way more false positives per class than intended.

Evaluation Metric	Score
Accuracy	0.819
Macro F1	0.86
Avg Precision	0.38
Avg Recall	1.0
D Prime	2.66
AUC	0.97

From this we can see that large pre-trained transformer models outperform simple baselines in a low-resource downstream task. We can apply larger models pre-trained on similar data to boost performance when data is sparse or unavailable.

Unfortunately due to time and GPU constraints, we were unable to test our theory of pre-training vs no pre-training with the transformer model. We would have liked to run monroe experiments where the Vision Transformer model had never seen audio image spectrograms before, in order to compare the effects of large pre-training on audio image spectrograms.

Lastly, we would have also liked to see how our model does against other emotion classification benchmark datasets in the future. As we received astounding results from this small RAVDESS dataset. It would be interesting to see how a larger dataset could boost performance and possibly remedy our low precision performance.

5. Conclusion

As we have seen, large pre-trained transformer models outperform all other baselines because of the additive property of pre-training and the helpfulness of the attention mechanism. We saw significant increase in performance with a relatively small dataset (<7000 overall examples in RAVDESS). We have shown how large pre-trained models, especially with vision transformer architectures, can improve the low-resource task of emotion classification of audio image spectrograms. This proves our hypothesis that pre-training can also help in downstream audio tasks.

Furthermore, if time permitted, future work would be done in comparing large pre-trained CNN's and LSTM's to compare them against transformer models when classifying audio image spectrograms for emotion classification. This would be an interesting study to see which model can perform the best on sequential audio image spectrogram data.

References

[1] METRICS FOR MULTI-CLASS CLASSIFICATION

[2] Pengcheng Li and Yan Song and Ian Mcloughlin and Wu Guo and Lirong Dai.
An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition

[3] AST: Audio Spectrogram Transformer Yuan Gong, Yu-An Chung, James Glass

[4] Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the INTERSPEECH 2019: Training Strategy for Speech Emotion Recognition, Graz, Austria, 15–19 September 2019.

[5] Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D and 2D CNN LSTM networks. Elsevier Biomed. Signal Process. Control 2019, 47, 312–323.

[6] Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensors 2021, 21, 1249.

[7] Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, J. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018.

[8] Vaswani, Ashish, et al. "Attention Is All You Need." *ArXiv.org*, 6 Dec. 2017, https://arxiv.org/abs/1706.03762.

[9] Xie, Yue, et al. "Speech Emotion Classification Using Attention-Based LSTM." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, 2019, pp. 1675–1685., https://doi.org/10.1109/taslp.2019.2925934.

[10] Ranftl, Rene, et al. "Vision Transformers for Dense Prediction." *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, https://doi.org/10.1109/iccv48922.2021.01196.

[11] Zhang, Shiqing, et al. "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching." IEEE Transactions on Multimedia 20.6 (2017): 1576-1590.