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

CS231N: Yair Sachar, Anthony Le, Omer Benyshai

Introduction

**Emotion Classification/Recognition:**
- The task of emotion classification serves as a downstream task for computer vision, natural/speoken language processing alike.
- Our project aimed at classifying audio samples containing human speech by transforming audio files into image-like representations and using large pre-trained computer vision models for classification.
- Large CNN/LSTM models have shown good performance in classification of audio image spectrogram features, but more recent work cites vision transformers as a possible improvement for audio classification.
- We investigate vision transformers against simple baselines.

**Dataset**
RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song contains 7356 files containing short video and audio clips of actors vocalizing two lexically-matched statements in a neutral North American Accent. Each example was rated 10 times on emotional validity, intensity, and genuineness, for 8 different emotions: calm, happy, sad, angry, fearful, surprised, and disgusted. For the scope of this project we only used the speech files, giving us a total of 2880 audio/video files for our model. We split our dataset into training and validation sets with 80% being the training data and 20% being validation.

**Architecture**
- The model is pre-trained on the ImageNet dataset which contains 1.2 million different visual objects that are semantically classified into 1000 different classes of objects in their ontology.
- Our model uses a canonical transformer architecture with cross attention and positional embeddings added, positional embeddings are added, and the sequences are split into 16 x 10-second sound and video clips from the RAVDESS dataset which contains over 2 million sequential audio image spectrogram data.
- The novel idea of this approach is that transformers and visual data are extremely effective at modeling sequential 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.
- Furthermore, we could see how much the pre-training affects performance by ab testing the vision transformer model with and without the image pre-training.

**References**

We trained our audio spectrogram vision transformer model with cross entropy loss for 25 epochs with a learning rate of 0.05 using a learning decay rate of 0.85 starting at each epoch, starting at epoch 5. The model was pre-trained on both Audioset and ImageNet. The model performed exceptionally well relative to the baselines and surrounding literature.

**Methods & Results**

**Baseline: Feed-Forward Neural Network**
- The best performing simple neural network consisted of 5 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 Baselines: SVM's and Random Forest**
- These are the results from our random forest classifier with 100 individual estimators with majority vote ensembling. We see 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.

**Large Pre-trained Vision Transformer**
- Future work would be done in comparing large pre-trained CNN's and LSTMs 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.
- Furthermore, we could see how much the pre-training affects performance by ab testing the vision transformer model with and without the image pre-training.

**Evaluation Metric:**

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.919</td>
</tr>
<tr>
<td>Macro F1</td>
<td>0.86</td>
</tr>
<tr>
<td>Avg Precision</td>
<td>0.38</td>
</tr>
<tr>
<td>Avg Recall</td>
<td>0.9</td>
</tr>
<tr>
<td>D Prece</td>
<td>2.86</td>
</tr>
<tr>
<td>AUC</td>
<td>0.97</td>
</tr>
</tbody>
</table>

**Future Work**

**References**

8. [CS231N: Yair Sachar, Anthony Le, Omer Benyshai](https://www.youtube.com/watch?v=dQw4w9WgXcQ).
9. [CS231N: Yair Sachar, Anthony Le, Omer Benyshai](https://www.youtube.com/watch?v=dQw4w9WgXcQ)