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

CS231N: Yair Sachar, Anthony Le, Omer Benyshai

Introduction

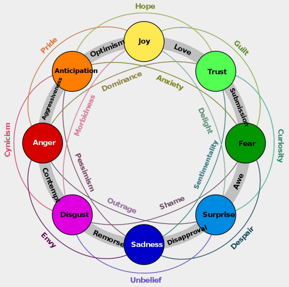
Emotion Classification/Recognition:

- The task of emotion classification serves as a downstream task for computer vision, natural/spoken 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.



Methods & Results

	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

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 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.

	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

Baseline: Feed-Forward Neural Network

- 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. 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.

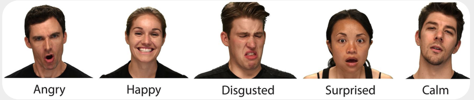
Evaluation Metric	Score
Accuracy	0.819
Macro F1	0.86
Avg Precision	0.38
Avg Recall	*1.0
D Prime	2.66
AUC	0.97

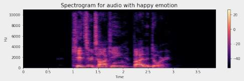
Large Pre-trained Vision Transformer

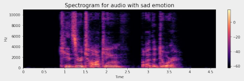
- We trained our audio spectrogram vision transformer model with cross entropy loss for 25 epochs with a learning rate of 1e-5 using a learning decay rate of 0.85 starting at each epoch, starting at epoch 5. The model was pre-trained on both Audioset and ImageNet. The model performed exceptionally well relative to the baselines and surrounding literature.

* Average Recall of 1.0 is impossible. We think we had a small bug in our evaluation metrics code.

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.

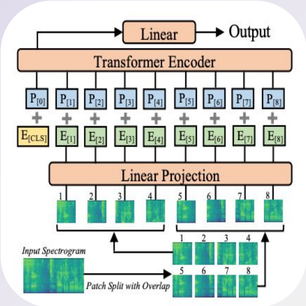
Architecture

- 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.

- 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 spectrogram features with 128-dimensional log Mel filterbank features.

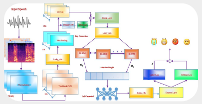
- Then, these sequences are splitted into 16 x 16 patches and flattened to produce our input tokens, similar to convolution. Additionally, the positional embeddings are added here. Inputs are passed into the encoder and linear layer with softmax classifier will learn the classification task.

- 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 this task is sparse.



Future Work

- 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.
- Furthermore, we could see how much the pre-training affects performance by a/b testing the vision transformer model with and without the image pre-training.



References

[1] Grandini, Margherita, et al. "Metrics for Multi-Class Classification: An Overview." *ArXiv.org*, 13 Aug. 2020, https://arxiv.org/abs/2008.05756.
[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 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.