Interpreting Transformer-Based Models for Speech-Related Task

Mridul Gupta, Siddhant Kumar, Siddhant Mahurkar, Angad Nandwani, Eubin Park, Akshat Gupta (JPMC), Dr. Sining Chen

Columbia University Data Science Institute

JPMorgan Chase & Co.

Objective

To interpret and understand the information captured by Wav2Vec2.0 & HuBERT models for Speech Emotion Recognition (SER) and Keyword Spotting (KS). We extract hidden states from each layer and train classification models to analyze the information gain across the 12 layers of transformer models.

Speech Emotion Recognition (SER)
- To determine the emotion conveyed by speaker
- 7742 clips including 91 actors
- Train-Test-Validation = [90-5-5]

Keyword Spotting (KS)
- To detect when the certain keyword is spoken
- 60,793 audio samples, each of 1 second
- Train-Test-Validation = [84-11-5]

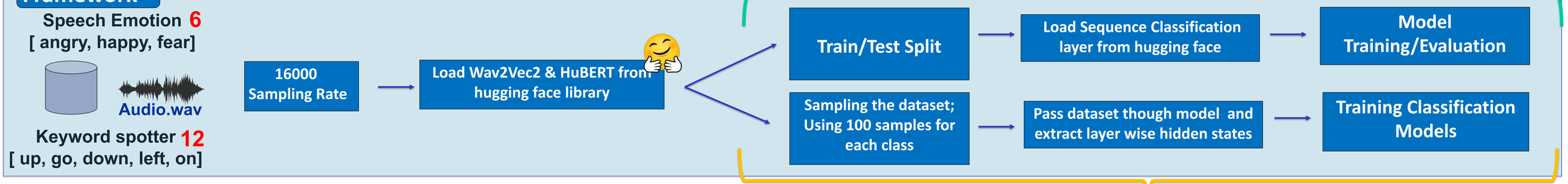
Wav2Vec2.0/HuBERT

- Wav2Vec is a speech encoder that employs semi-supervised training to improve supervised speech recognition. Uses CNN based network to convert the audio signal to a contextualized representation which can be used for down-stream tasks.
- Hidden-Unit BERT (HuBERT) approach for self-supervised speech representation learning, which utilizes an offline clustering step to provide aligned target labels for a BERT-like prediction loss.

TASK	Model	Epochs	Accuracy
KS	Wav2Vec2.0	5	98.23
	HuBERT	5	98.19
SER	Wav2Vec2.0	10	78.00
	HuBERT	5	77.14

Table 1: Transformer Based Models Training Accuracy

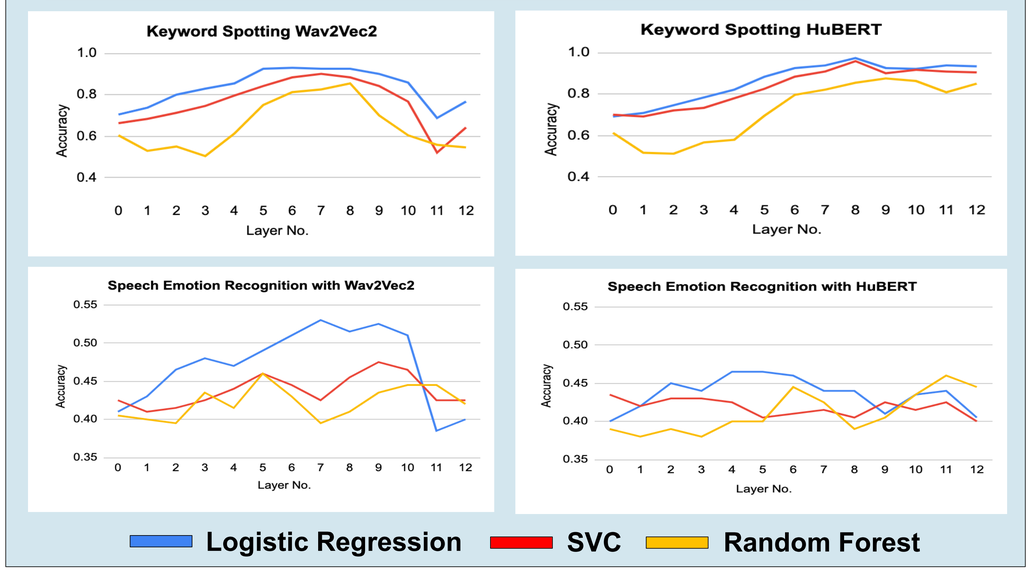
Framework



Conclusion

- For most classes in Keyword Spotting and Speech Emotion Recognition, middle layers are performing better than initial layers.
- Change in accuracy across layers was substantial for Keyword Spotting in comparison to Speech Emotion Recognition.
- "Silence" (_SIL_) keyword and "angry" emotion are accurately classified across all 12 layers
- "Fear" & "Neutral" emotions and "Stop" & "Go" keywords are not correctly classified across all layers.

Result : Layer wise Probing Accuracies



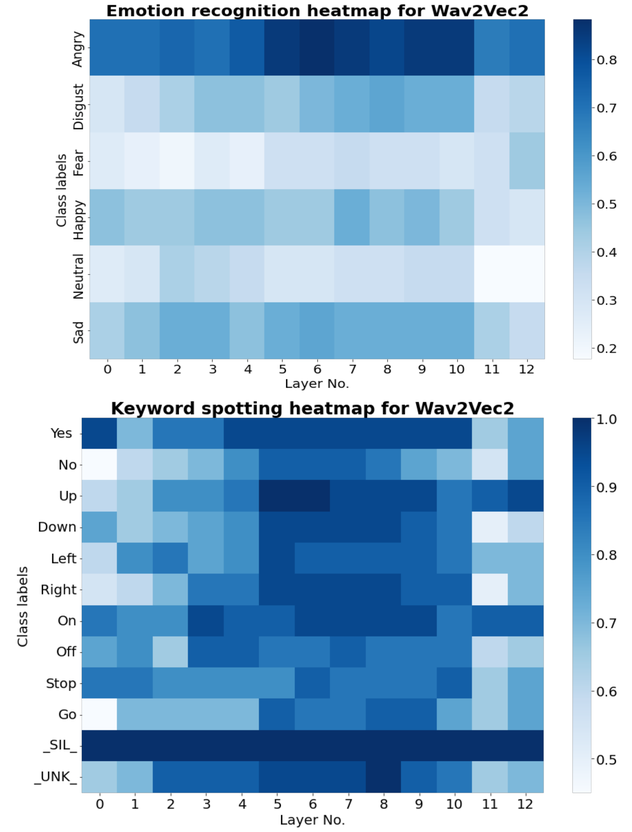


Fig 1 : Layer Wise #True Positive/Class

References

- Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.
- Wang, Y., Boumadane, A., & Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. arXiv preprint arXiv:2111.02735.

Acknowledgement

Sincere thanks to Akshat Gupta, Prof. Sining Chen, and Dr. Eleni Drinea for their constant support, expert guidance with the project and resources.