# Multimodal Transformer with Learnable Frontend and Self Attention for Emotion Recognition

Paper# 5937

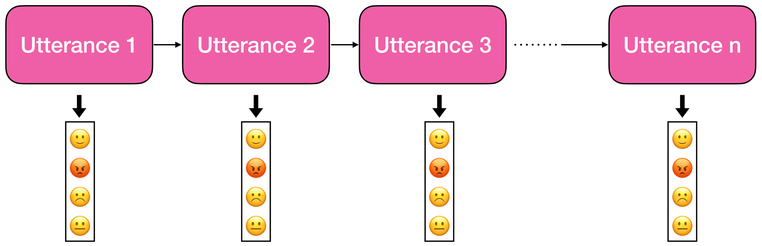
**Soumya Dutta and Sriram Ganapathy**
Learning and Extraction of Acoustic Patterns (LEAP) Lab, Dept. of Electrical Engg.
Indian Institute of Science, Bangalore - 560012, India
Email: {soumyadutta, sriramg}@iisc.ac.in

## At a glance

- This work proposes a novel approach for emotion recognition in conversations which achieves state-of-the-art for the IEMOCAP database
- The model is shown to be more robust to textual errors caused by an automatic speech recognition (ASR) system.

## 2. Emotion recognition in conversations

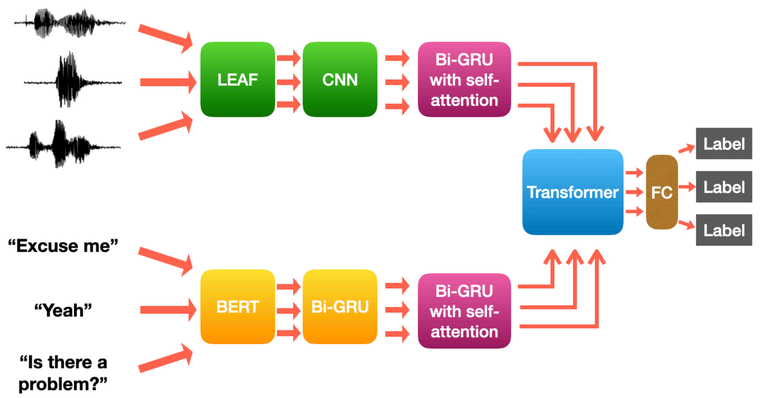
- Each dataset has a number of conversations
- Each conversation has a number of utterances each of which has an emotional label, E.g. happy, angry, sad, neutral




## 3. Our approach

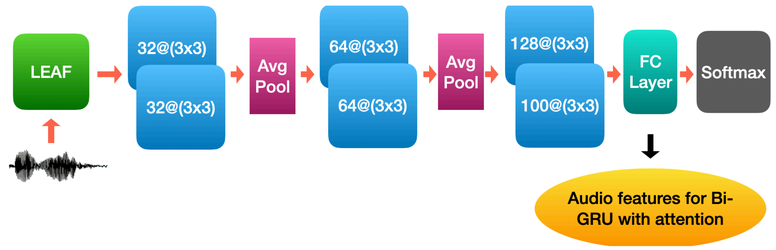
- A learnable frontend used for audio feature extraction
- Effective context addition using self-attention with Bi-GRU network
- Multimodal transformers used for fusion of modalities
- Model trained in a hierarchical manner

## 4. Proposed Model



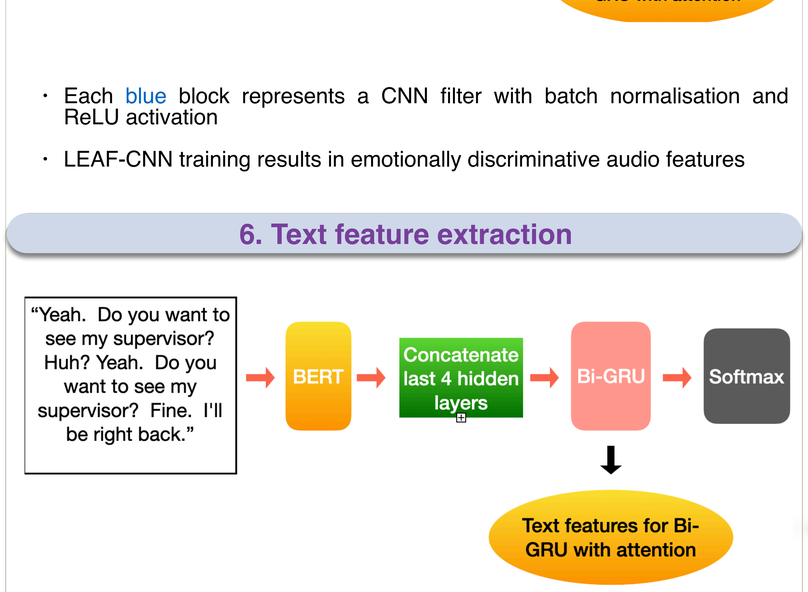

- Audio feature extraction: LEAF-CNN; text feature extraction: BERT-BiGRU
- Model trained hierarchically, E.g. when Bi-GRU with self-attention is trained, LEAF-CNN and BERT-BiGRU networks are frozen

## 5. Audio feature extraction




- Each blue block represents a CNN filter with batch normalisation and ReLU activation
- LEAF-CNN training results in emotionally discriminative audio features

## 6. Text feature extraction




- BERT-BiGRU training results in emotionally discriminative text features
- BERT-base pre-trained model is not frozen during this training

## 7. Multi-utterance self-attention

- Separate BiGRU networks used for text and audio for adding context
- Self-attention employed across utterances for better context modelling
- $O_f$ : forward outputs from Bi-GRU
- $W_f^a$: attention layer parameters
- S : number of utterances in the conversation

$$A_f = (O_f W_f^a)(O_f W_f^a)^T$$

$$A_f^{ij} = \frac{exp(A_f^{ij})}{\sum_{j=1}^{S} exp(A_f^{ij})} \forall i, j \in \{1,2,\ldots,S\}$$

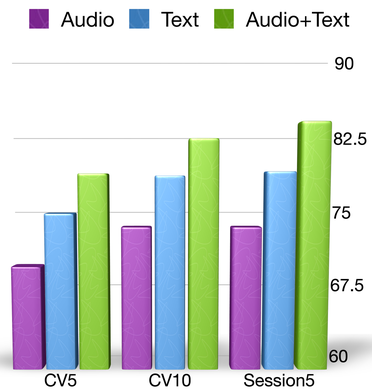
$$O_f^a = A_f O_f$$

## 8. Dataset

- IEMOCAP has 151 recordings - divided into 5 sessions
- Each utterance is labeled one of the four categories - happy, angry, sad and neutral


## 9. Results

- Common test settings - CV5 - 5-fold cross validation, CV10 - 10 fold cross validation, Session 5 as test




| System | Test setting | Modality | Accuracy |
|---|---|---|---|
| Poria et al. | Session 5 | T, A, V | 76.1% |
| Wu et al. | Session 5 | T, A | 83.2% |
| This work | Session 5 | T, A | 83.8% |
| Yoon et al. | CV-5 | T, A | 71.8% |
| Wu et al. | CV-5 | A | 61.3% |
| Wu et al. | CV-5 | T, A | 78.4% |
| This work | CV-5 | T, A | 78.9% |
| Poria et al. | CV-10 | T, A, V | 76.1% |
| Yoon et al. | CV-10 | T, A | 77.6% |
| Li et al. | CV-10 | T, A | 79.2% |
| Wu et al. | CV-10 | T, A | 78.3% |
| This work | CV-10 | T, A | 82.2% |

## 10. Results with ASR transcripts

- Provided transcripts replaced by Google speech-to-text (42% WER)
- Models trained with provided transcripts, tested with ASR transcripts

| System | Test setting | Accuracy |
|---|---|---|
| Wu et al. | CV-5 | 63.5% |
| This work | **CV-5** | **74.9%** |
| This work | CV-10 | 76.2% |
| This work | Session 5 | 77.3% |

## References

•Poria et al. "Context-dependent sentiment analysis in user-generated videos." *ACL* 2017.
•Wu et al."Emotion recognition by fusing time synchronous and time asynchronous representations." *ICASSP 2021*
•Yoon et al. "Speech emotion recognition using multi-hop attention mechanism." ICASSP 2019
•Li et al. "Towards Discriminative Representation Learning for Speech Emotion Recognition." *IJCAI*. 2019.