Efficient Modelling of Long Temporal Contexts for Continuous Emotion Recognition

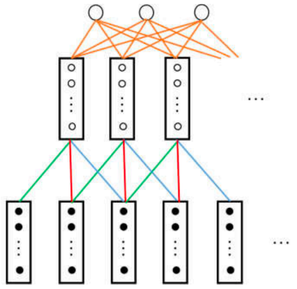
Jyotishman Das (M24CSA013) Pranjal Malik (M24CSA021)
Indian Institute of Technology Jodhpur

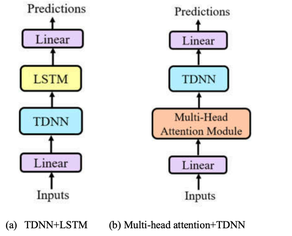
Proposed Methodology:

➢ Continuous Emotion Recognition (CER) is a challenging domain as we aim to track continuous emotional states like arousal and valence over time. This project targets modelling of long-term temporal dependencies using deep learning architectures. We implemented and compared three temporal models — LSTM, TDNN, and Multi-head Attention — and also experimented with hybrid combinations to exploit sequential and contextual information.

➢ Temporal Models
- **LSTM (Long Short-Term Memory)**: Used to track temporal emotions by handling sequential dependencies.
- **TDNN (Time Delay Neural Network)**: Used to capture context by temporal convolutions.
- **Multi-head Attention**: Used to give weightage to temporal information.

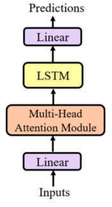
➢ Hybrid Architectures:
- TDNN + LSTM
- Attention + TDNN
- Attention + LSTM
- Attention + TDNN + LSTM



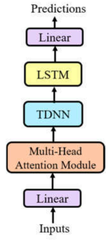


(a) TDNN+LSTM

(b) Multi-head attention+TDNN



(c) Multi-head attention+LSTM



(d) Multi-head attention+TDNN

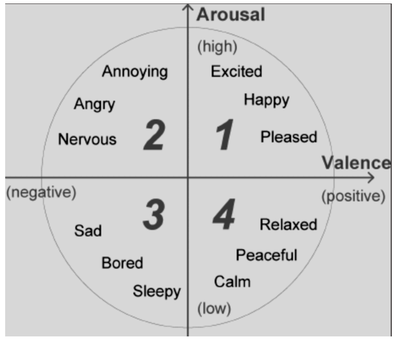
+LSTM

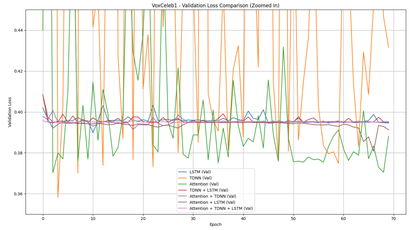
Dataset, Features & Training:

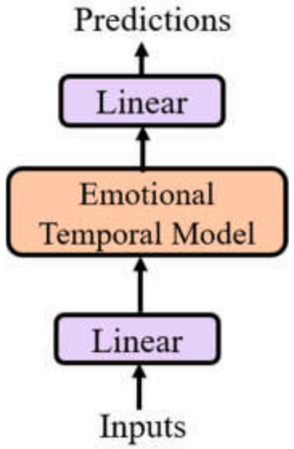
➢ Data Source (used random arousal and valence)
- **Librispeech**
- **VoxCeleb1**
- **VoxCeleb2**

➢ Features
- **Audio**: eGeMAPS (88D, using openSMILE)

➢ Training:
- **Framework**: TensorFlow
- **Optimizer**: Adam
- **Batch Size**: 3
- **Epochs**: 70
- **Evaluation Metric**: MSE, Concordance Correlation Coefficient (CCC)







Experimental Findings:

➢ Observations
- Multi-head Attention performs best among single models.
- Hybrid models outperform standalone ones, especially Attention+TDNN+LSTM.

Limitations and Proposals:

➢ Scarce continuous data availability

➢ Better fusion strategies than normal stacking of models

➢ Can use weighted fusions or use NAS

➢ Can include attention map visualisation for understanding human behaviour in multiple modality framework

➢ Improve upon acoustic only framework and understand behaviour of model with similar labeled text and facial expressions.

Summary

➢ Our Contributions
- Implemented and compared three temporal models: LSTM, TDNN, and Multi-head Attention.
- Evaluated hybrid architectures combining these models.

➢ Key Results
- Hybrid models outperformed individual ones; best CCC achieved with Attention + TDNN + LSTM + Transformer fusion.
- Improved performance on larger datasets
- CCC having positive values implies correlation between prediction and ground truth(random here)

➢ Further Resources
- Code Repository: GitHub Link