Attention-based Multimodal Speech Emotion Recognition

Luo Li, Yanning Chen, Jingyun Su, Shiyang Xing

Group 26

MSc Data Science and MSc DSML
University College London

March, 2022

Contents

Speech Emotion Recognition (SER)

- Real-time analysis of human speech to detect and classify emotions, enhancing human-computer interaction
- Applications
  - Mental health support
  - Customer service optimization
  - etc
- Machine Learning
  - Advanced algorithms
  - Deep learning
  - Natural language processing (NLP)

Multimodal SER

- Integration of multiple data sources, such as audio, facial expressions, and body language, for comprehensive emotion recognition
- Improved Performance
  - Enhanced accuracy
  - Contextual understanding
  - etc
- Expansive Applications
  - Supports sentiment analysis
  - Remote education
  - etc

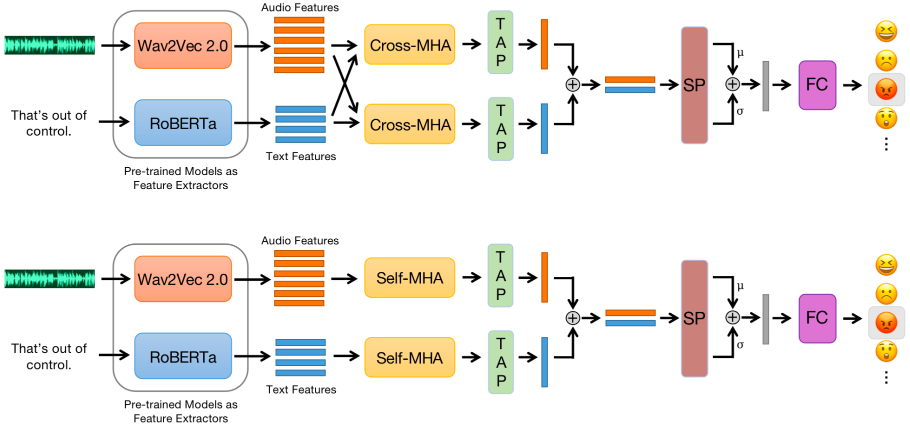
Contents

Background

- SOTA methods focus on two main sub-areas
  - Representation of multimodal data
  - Feature fusion
- Initiative: A recent paper compared the effectiveness of self- and cross-attention on traditional features [1]
- Research Gap: No one compared the effect of self- and cross-attention on self-supervised features
- Idea: Pre-trained models for feature extraction (representation), compare self- and cross-attention for feature fusion

Contents

Proposed Models



IEMOCAP Dataset

- 12 hours of emotional interactions in scripted/unscripted settings
- Recordings from 5 male and 5 female speakers
- Speech audio clips and ground-truth text transcripts with emotion labels

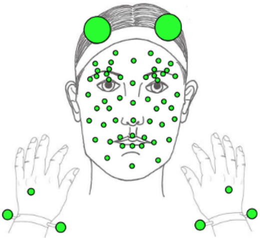


Figure 1. Marker layout. In the recording, fifty-three markers were attached to the face of the subjects. They also wore wristbands (two markers) and headband (two markers). An extra marker was also attached on each hand.

Training

- Trained the models on each fold of data for a maximum of 50 epochs
- Two strategies used
  - Learning rate scheduling
  - Early stopping
- Evaluated the models using the four metrics after each epoch on both the validation and test sets

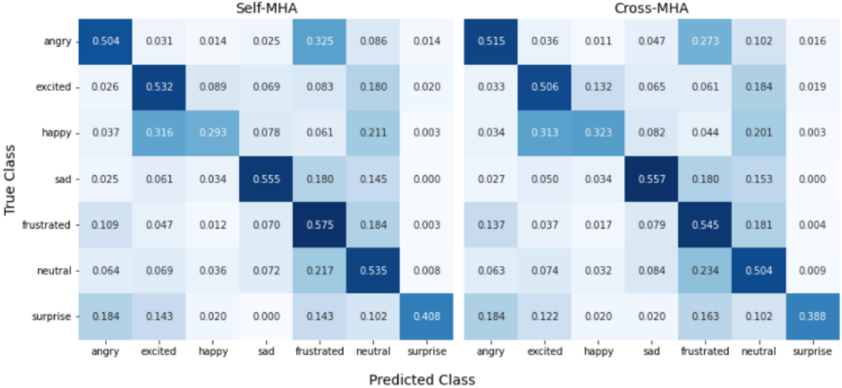
Contents

Experimental Results

	WA	UWA	WF1	UWF1
Random	0.138 ± 0.002	0.135 ± 0.004	0.151 ± 0.002	0.126 ± 0.002
Audio	0.398 ± 0.018	0.342 ± 0.006	0.367 ± 0.016	0.335 ± 0.007
Text	0.506 ± 0.027	**0.492 ± 0.029**	0.498 ± 0.025	0.475 ± 0.022
MDRE [x]	0.491 ± 0.039	0.466 ± 0.056	0.482 ± 0.035	0.470 ± 0.041
Self-MHA	**0.522 ± 0.014**	0.486 ± 0.020	**0.519 ± 0.016**	**0.488 ± 0.014**
Cross-MHA	0.509 ± 0.009	0.477 ± 0.027	0.509 ± 0.009	0.478 ± 0.024

- Self-attention model outperforms the cross-attention model
- Self-attention model outperformed all four baseline models
- Text modality can provide loads of valuable emotion information

Experimental Results



- Frequent confusion between some emotion classes
  - Angry vs Frustrated
  - Happy vs Excited
- Relatively poor performance on recognizing the class "surprise"

Ablation Study

	WA	UWA	WF1	UWF1
Self-noSP	0.514 ± 0.022	0.464 ± 0.042	0.507 ± 0.027	0.473 ± 0.046
Self-CLS	0.505 ± 0.008	0.461 ± 0.019	0.504 ± 0.009	0.468 ± 0.021
Self-MHA	**0.522 ± 0.014**	**0.486 ± 0.020**	**0.519 ± 0.016**	**0.488 ± 0.014**

	WA	UWA	WF1	UWF1
Cross-noSP	0.498 ± 0.005	0.442 ± 0.021	0.493 ± 0.005	0.448 ± 0.024
Cross-CLS	**0.518 ± 0.011**	0.451 ± 0.020	**0.512 ± 0.010**	0.456 ± 0.022
Cross-MHA	0.509 ± 0.009	**0.477 ± 0.027**	0.509 ± 0.009	**0.478 ± 0.024**

- Decreased performance without statistical pooling layer
- Use BERT's CLS token as the text feature
  - Decreased performance for self-attention model
  - Increased weighted accuracy and weighted F1 score for cross-attention model

Contents

Limitations

Self-attention is more effective for multi-modal emotion recognition

- Limited dataset (IEMOCAP only)
- Suboptimal hyperparameter tuning

Future Work

- Improve the model's ability to distinguish between confused emotion classes
- Integrate video components in multi-modal models

Contents

References

[1] Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4693–4697. IEEE, 2022.