Maastricht University

Faculty of Science and Engineering
Department of Data Science and Knowledge Engineering

Deep, dimensional and multimodal emotion recognition using attention mechanisms

Jan Lucas, Esam Ghaleb[A], Stylianos Asteriadis[A]
[A] Department of Data Science and Knowledge Engineering

Introduction

Problem: Prediction of emotion from a subject using audio and video

Emotion prediction benefits greatly from the fusion of multiple modalities[1]

Emotion can be expressed continuously

- **Arousal**
  Intense or calm emotion
- **Valence**
  positive or negative emotion

Faces are extracted from video and features

Feature extraction using VGGish and VGGFace

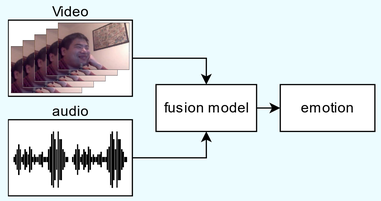


Figure 1: Problem overview

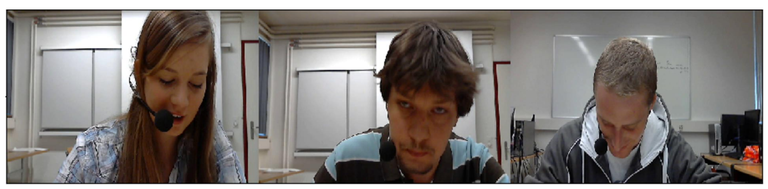


Figure 2: Example frames from the dataset

Model

Builds on DLSTM architecture by Zhao et al.[2]

Two additions:

- **Bimodal Attention Fusion**

$$S_m = C^\top W_m$$

  Fuses outputs of modality-specific DLSMs

  Attends to DLSTM outputs based on their hidden states

- **Embedding attention**

  Uses General Dot-Product attention[3]

  Divides attention over a window of embeddings

  Attends to important video features

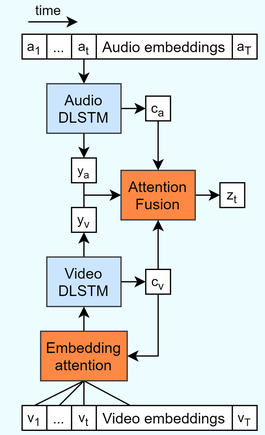


Figure 3: Overview of the proposed model

Experiments

The RECOLA dataset used for the AVEC 2016 competition used for validation

- Training only the video part of the network with and without Embedding Attention.
- Comparing different fusion attention methods by fusing pre-trained DLSMs
  - Compared with two baselines (Output Linear Baseline & Hidden Linear Baseline)
- Comparison to the state of the art by training the model end-to-end.

Evaluation metric: Concordance correlation coefficients (CCC)

Results & Discussion

Embedding Attention

Embedding Attention required use of custom training procedure.

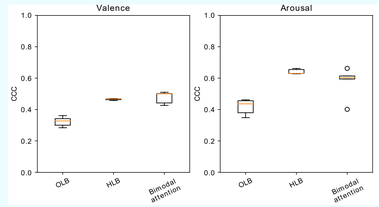
Use of embeddings slightly increased performance of the DLSTM, however not significantly.

	Valence	Arousal
No Embedding Attention	0.36 (±0.07)	0.15 (±0.05)
Embedding Attention	0.39 (±0.06)	0.14 (±0.05)

Fusion Attention

Fusion increases performance of base models

Bimodal Attention does not outperform the baselines



State of the art

- Arousal
  Outperforms baseline
- Valence
  Just below baseline

	Valence	Arousal
Baseline	0.683	0.639
State of the art[4]	0.702	0.82
Proposed	0.62	0.72

Both the baseline as the state of the art incorporated more modalities than just video and audio

Conclusion

Use of attention mechanisms for emotion recognition resulted in:

- promising results, successfully combining information from multiple modalities.
- Future research could expand on this architecture by experimenting with different forms of attention, extra modalities and different feature embeddings.

References

[1] Wu, Z., Zhang, X., Zhi-Xuan, T., Zaki, J., Ong, D.C.: Attending to emotional narratives. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 648–654. IEEE Computer Society (sep 2019).https://doi.org/10.1109/ACII.2019.8925497
[2] Zhao, J., Li, R., Chen, S., Jin, Q.: Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. p. 65–72. AVEC'18, Association for Computing Machinery (2018). https://doi.org/10.1145/3266302.3266313
[3] Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421. Association for Computational Linguistics (Sep 2015). https://doi.org/10.18653/v1/D15-1166
[4] Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., Huang, T.S.: Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. p. 97–104. AVEC'16, Association for Computing Machinery (2016). https://doi.org/10.1145/2988257.2988264

Correspondence to:
Jan Lucas

jan-lucas@hetnet.nl

Department of Data Science and Knowledge Engineering

T +3143 388 3455
F +31 43 388 3455

Department of Data Science and Knowledge Engineering
Maastricht University

P.O. Box 616
6200 MD Maastricht, The Netherlands