CIS centre for intelligent sensing

Queen Mary University of London

FONDAZIONE BRUNO KESSLER | ICT CENTER FOR INFORMATION AND COMMUNICATION TECHNOLOGY

IS CROSS-ATTENTION PREFERABLE TO SELF-ATTENTION FOR MULTI-MODAL EMOTION RECOGNITION?

VANDANA RAJAN[1], ALESSIO BRUTTI[2], ANDREA CAVALLARO[1]

[1]CENTRE FOR INTELLIGENT SENSING, QUEEN MARY UNIVERSITY OF LONDON, UK

[2]CENTER FOR INFORMATION AND COMMUNICATION TECHNOLOGY, FONDAZIONE BRUNO KESSLER, ITALY

Contents

Background
- Automatic emotion recognition
- Self attention & cross attention
- Multi-head attention

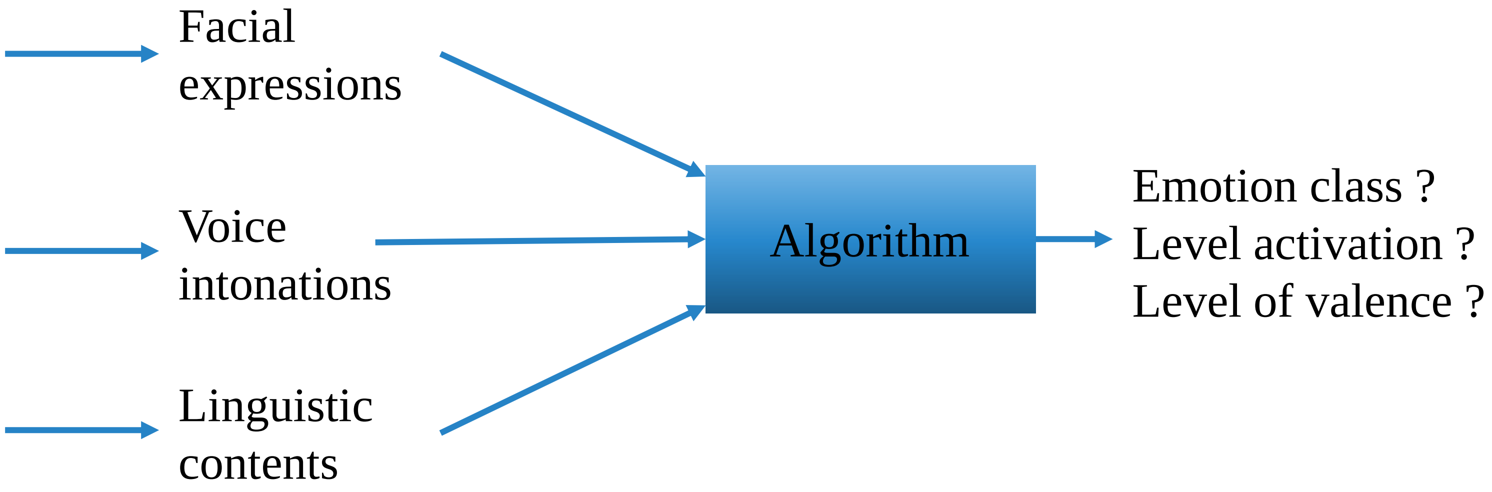
Self and cross-attention models
- Architecture

Validation
- Dataset
- Results

Conclusion

Automatic emotion recognition

Image courtesy: dreamstime.com



Facial expressions

Voice intonations

Linguistic contents

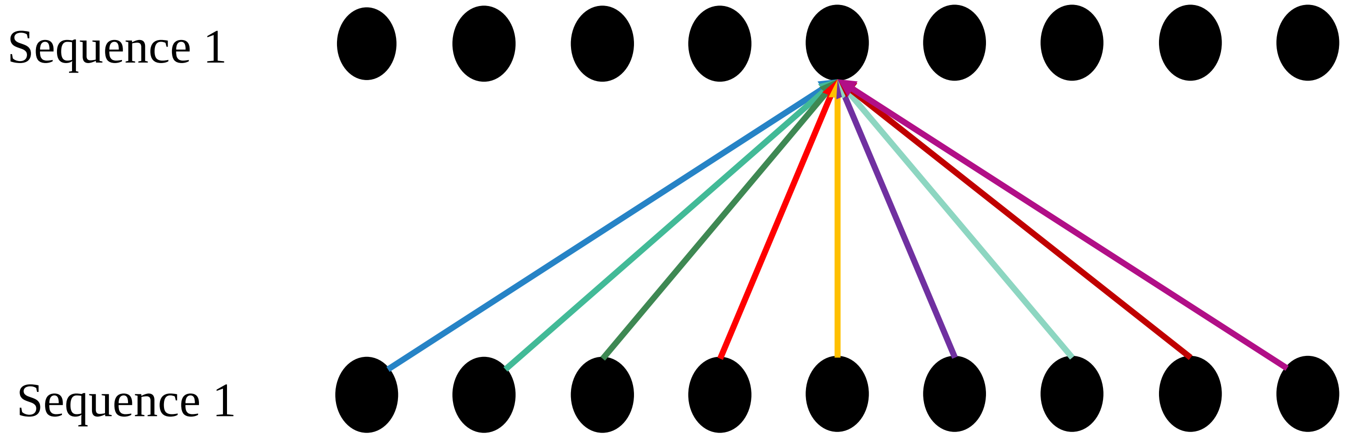
Algorithm

Emotion class ?
Level activation ?
Level of valence ?

Self-attention

Self-attention (intra-attention):

- relates different positions of a sequence to compute a representation of the ***same*** sequence [1].



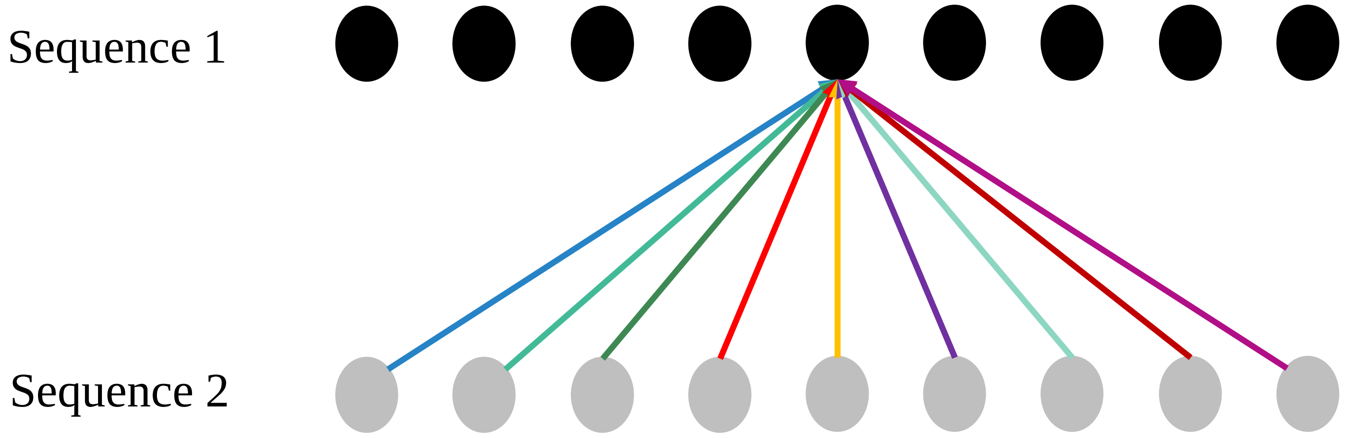
Different colours indicate different weights

[1] Cheng, Jianpeng, Li Dong, and Mirella Lapata. "Long Short-Term Memory-Networks for Machine Reading." *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016.

Cross-attention

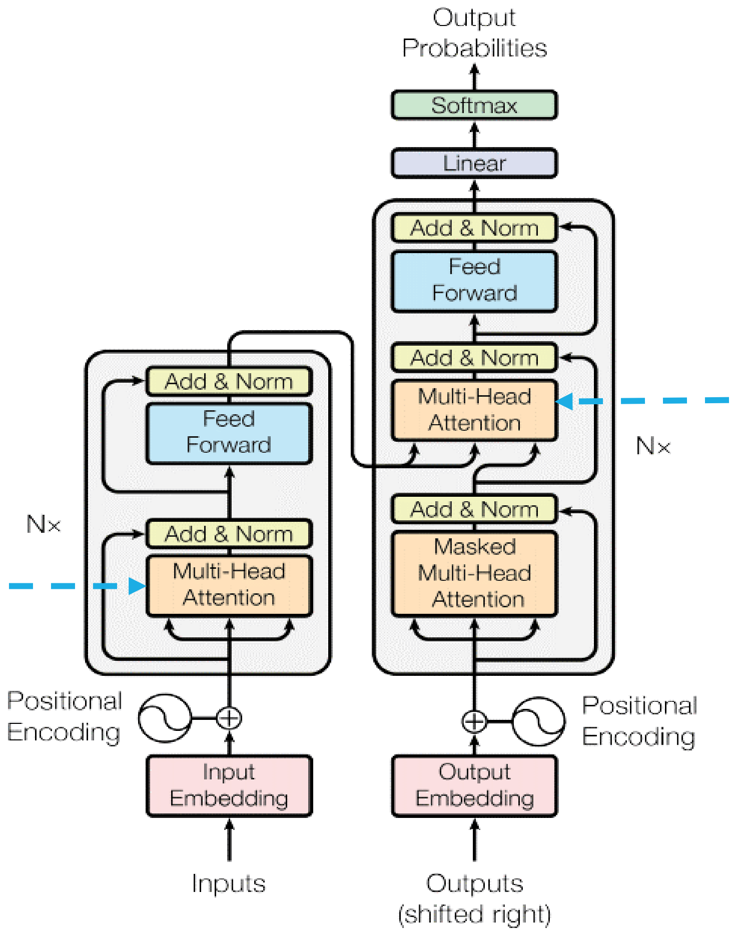
Cross-attention (inter-attention):

- relates different positions of one sequence to compute a representation of ***another*** sequence.



Different colours indicate different weights

Multi-head attention (1/2)



Cross-attention

Self-attention

- Transformers introduced MHA with self and cross-attentions [2]

[2] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Multi-head attention (2/2)

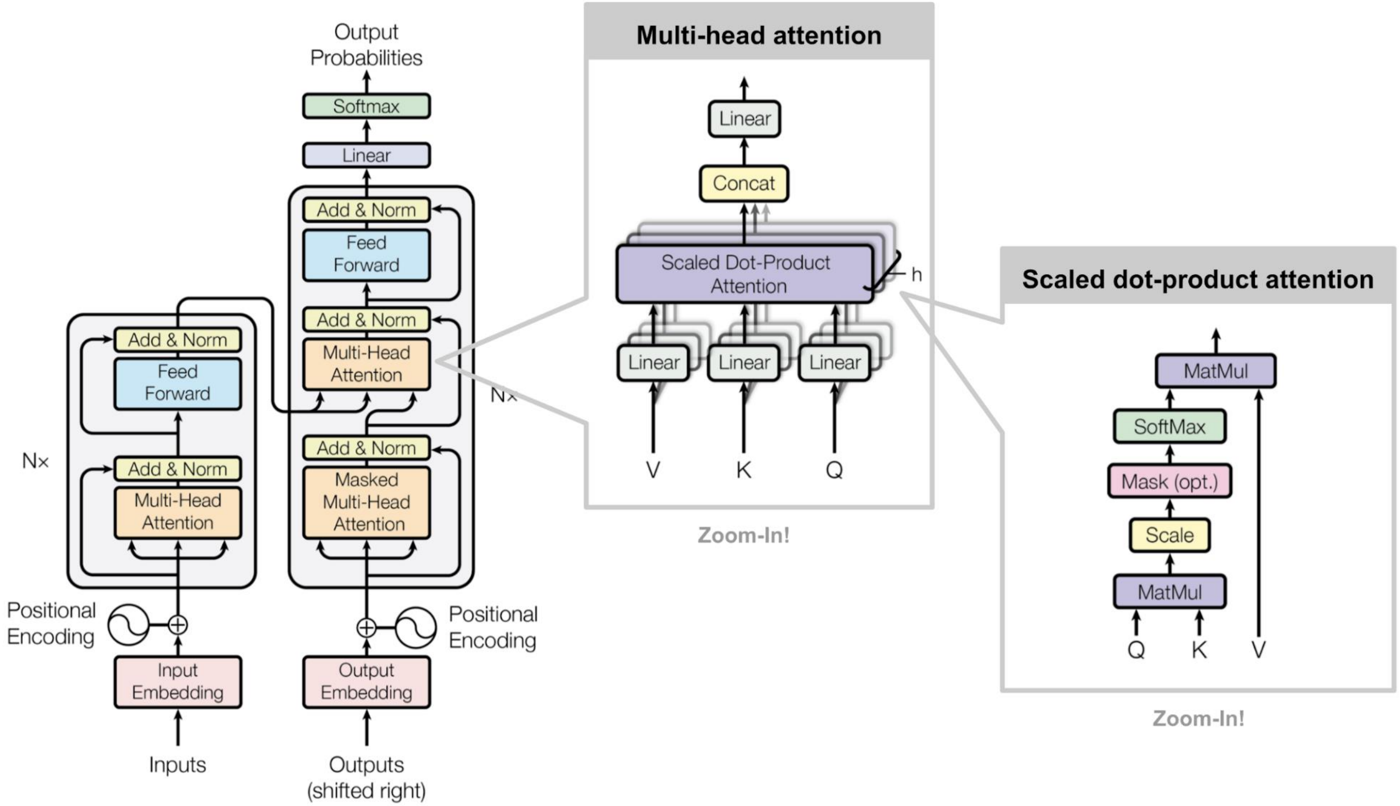


Figure from https://lilianweng.github.io/posts/2018-06-24-attention/

- Self-attention:
  - Intra-modal
  - V, K, Q – same modality
- Cross-attention:
  - Inter-modal [3]
  - V,K – source modality
  - Q – target modality

[3] Tsai, Yao-Hung Hubert, et al. "Multimodal transformer for unaligned multimodal language sequences." *Proceedings of the conference. Association for Computational Linguistics. Meeting*. Vol. 2019. NIH Public Access, 2019.

Self or cross-attention ?

RQ: Is cross-attention preferable to self-attention for MMER ?

Compare 2 models, one made of self (intra-modal) attention, and another made of cross (inter-modal) attention.

Model design

Modality specific encoders

- 1D convolutions – extract local task-relevant components
- Bi-GRU – global sequence modelling

Attention modules

- MHA – varying focus on time instances according to task-relevance

Temporal averaging

- Global representation of entire utterance

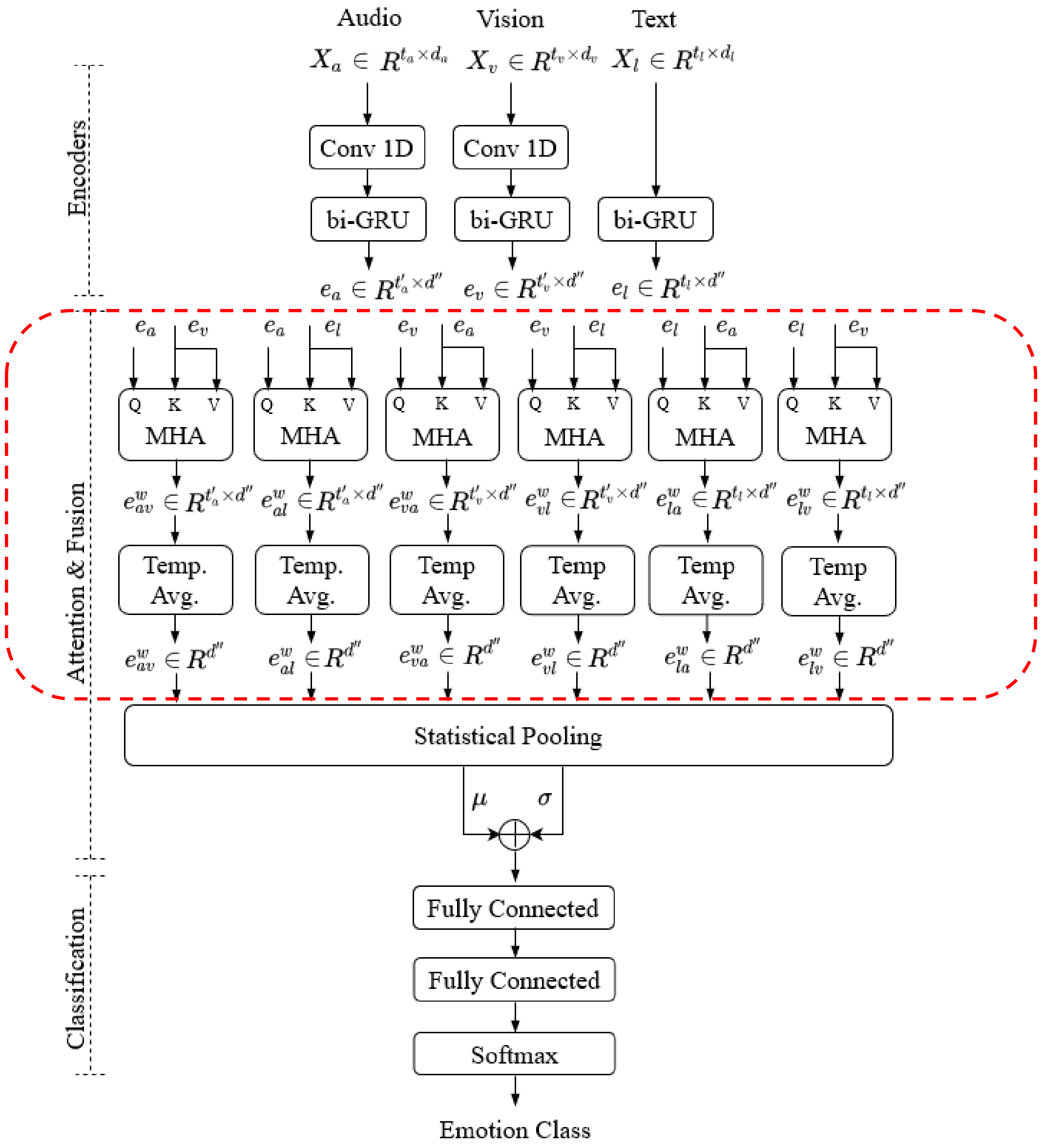
Statistical pooling

- Mean and standard deviation features

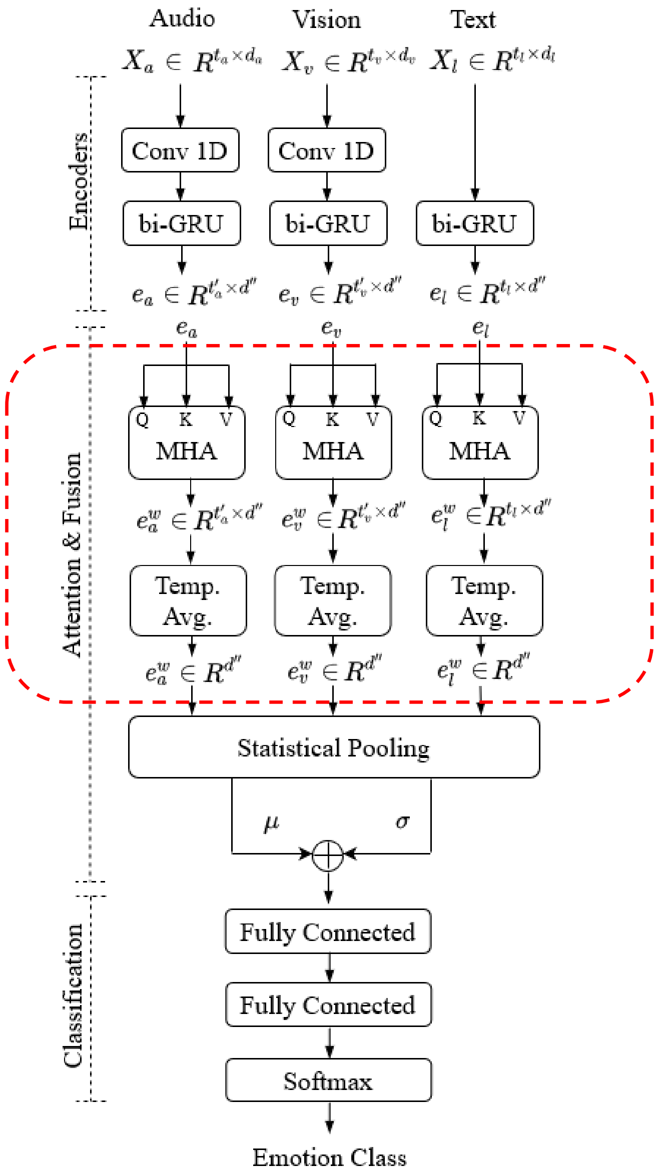
Classifier

- Fully connected layers output predictions

Tri-modal cross-attention model



Tri-modal self-attention model



Validation

Dataset: IEMOCAP [4] (~ 12 hours)

Total num. utterances: 7487

Classes: 7

- 1,103 angry, 1,041 excited, 595 happy, 1,084 sad, 1,849 frustrated, 107 surprise and 1,708 neutral

5 fold cross-validation

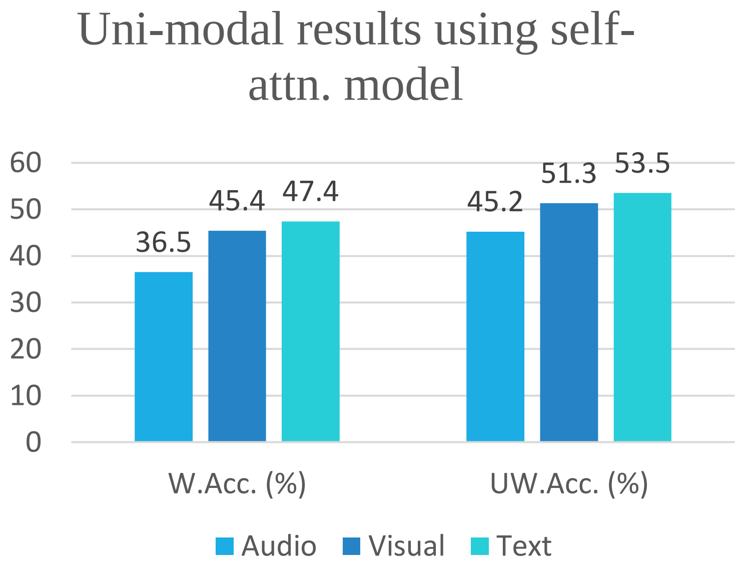
- Train and evaluate models 10 times per fold (with 10 different random seeds)

Metrics:

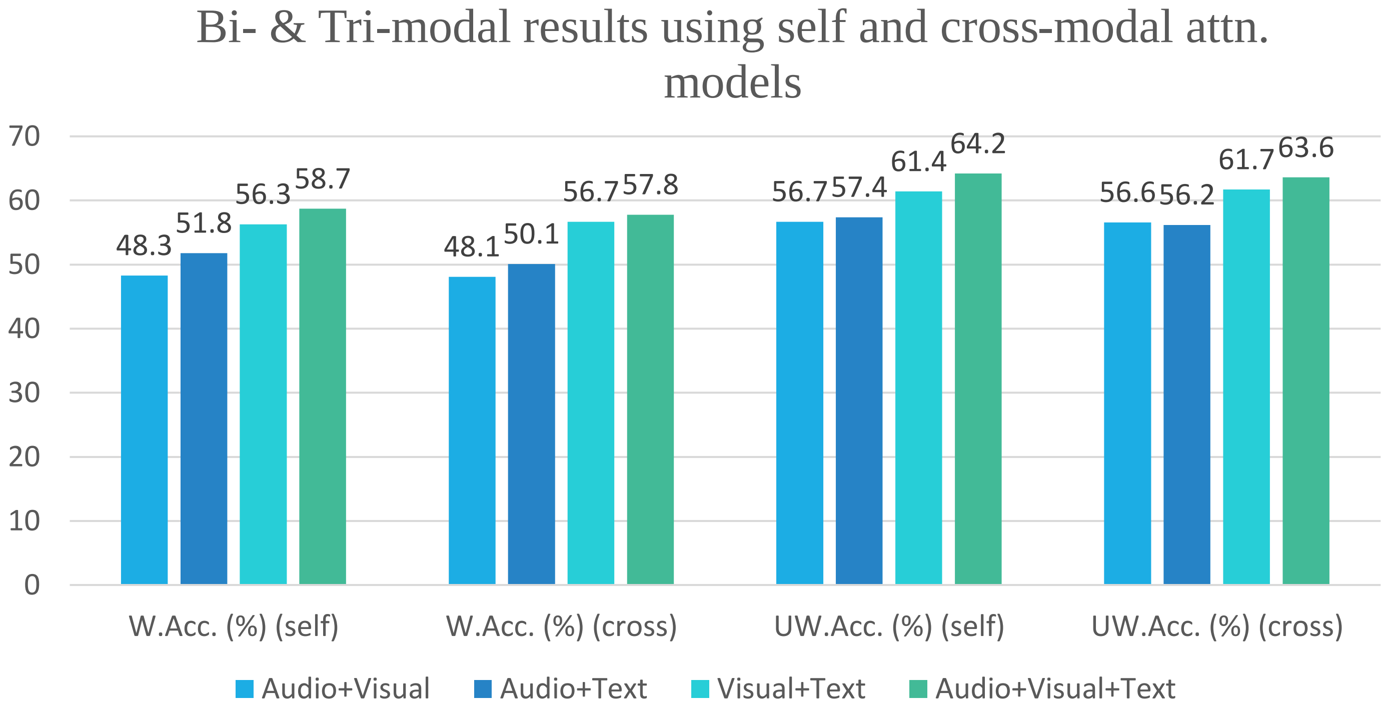
- WA – Weighted Accuracy
- UWA – Un-Weighted Accuracy

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

Results (1/2)



Uni-modal results using self-attn. model



Bi- & Tri-modal results using self and cross-modal attn. models

Tri-modal > Bi-modal > Uni-modal for both fusion models

Results (2/2)

Ablation and model combination

Model	Weighted Accuracy	Unweighted Accuracy
Cross-noSP	.570 (.021)	.634 (.015)
Cross	.578 (.024)	.636 (.012)
Self-noSP	.584 (.021)	.638 (.019)
Self	**.587 (.022)**	**.642 (.019)**
Cross+Self	.585 (.028)	.642 (.020)

Conclusion

No statistically significant difference between self and cross-attn. models

Combination of self and cross-attn did not improve performance

This might indicate that the cross-attn model does not contribute any additional info.

Code

https://github.com/smartcameras/SelfCrossAttn