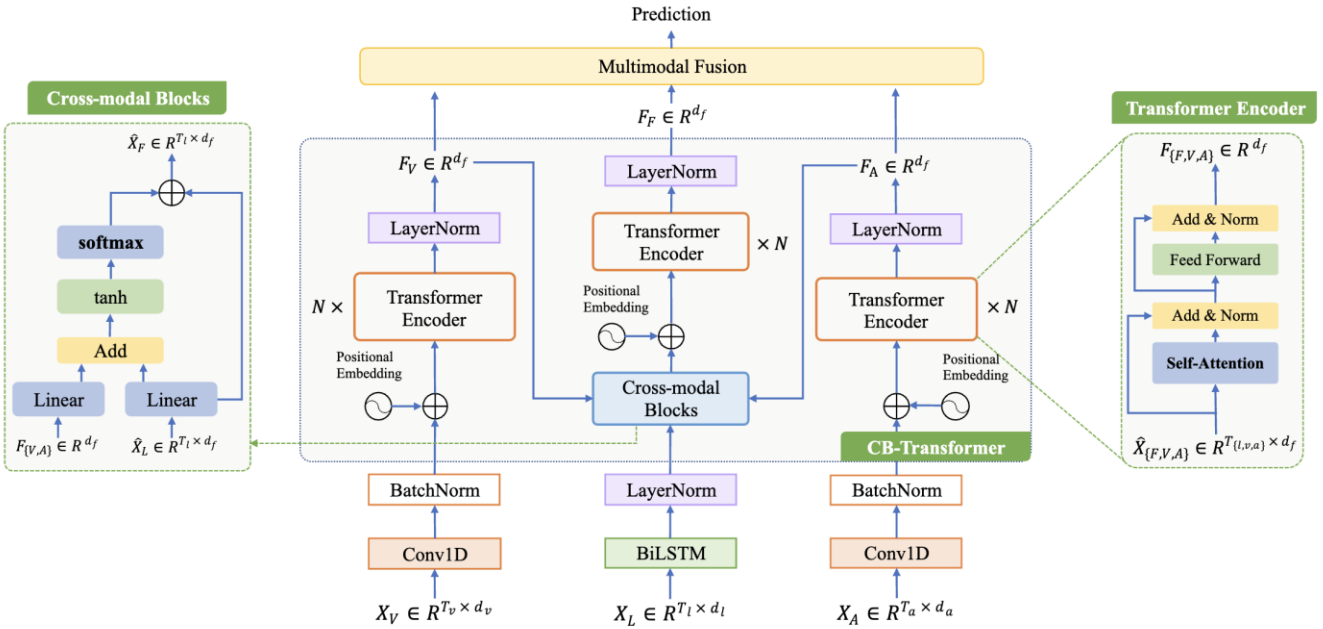
LMR-CBT: Learning Modality-fused Representations with CB-Transformer for Multimodal Emotion Recognition from Unaligned Multimodal Sequences

Ziwang FU, Feng LIU, Qing XU, Xiangling FU, Jiayin QI

Frontiers of Computer Science, DOI: 10.1007/s11704-023-2444-y

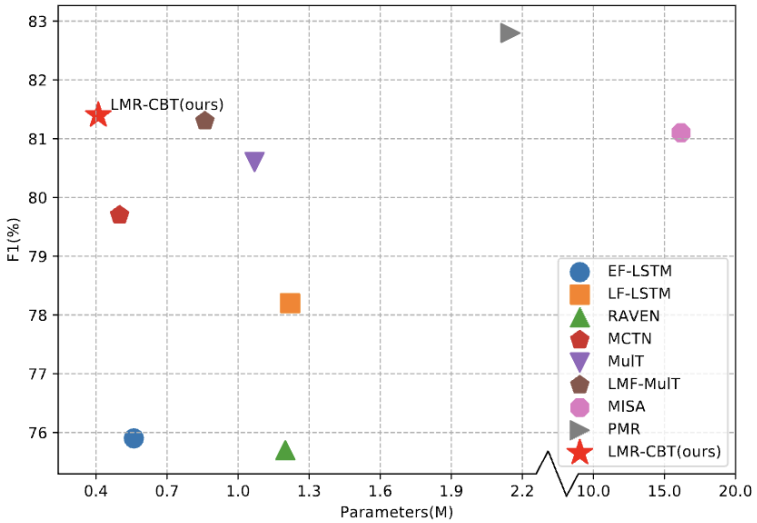
Problems & Ideas

- Problems of multimodal emotion recognition approaches:
  - Existing approaches use directional pairwise attention or a message hub to fuse language, visual, and audio modalities.
  - These fusion methods are often quadratic in complexity with respect to the modal sequence length, bring redundant information and are not efficient.
- Ideas: We propose an efficient neural network to learn modality-fused representations with CB- Transformer (LMR-CBT) for multimodal emotion recognition from unaligned multimodal sequences (only 0.41M), which can effectively fuse the interactive information of the three modalities.



Main Contributions

- Contributions:
  - We design an innovative asymmetric transformer with cross-modal blocks (CB-Transformer) to achieve complementary learning of different modalities, which is mainly divided into local temporal learning, cross-modal feature fusion and global self-attention representations. The CB-Transformer can adequately represent the fused features without losing the original features and can efficiently handle unaligned multimodal sequences.
  - We obtain a better trade-off between the performance and the efficiency on three challenging datasets. Compared with the existing state-of-the-art methods, LMR-CBT achieves comparable or even higher performance with a minimal number of parameters.



Setting	Method	Acc_7(%)	Acc_2(%)	$F1$(%)
Unaligned	EF-LSTM	46.3	76.1	75.9
	LF-LSTM	48.8	77.5	78.2
	RAVEN	45.5	75.4	75.7
	MCTN	48.2	79.3	79.7
	MulT (1.07M)*	50.4	80.7	80.6
	LMF-MulT (0.86M)	49.3	80.8	81.3
	MISA (15.9M)‡	52.1	80.7	81.1
	PMR (2.15M)†	51.8	83.1	82.8
	LMR-CBT (0.41M)	**51.8**	**80.9**	**81.5**
	LMR-CBT (1.23M)	**51.9**	**82.7**	**82.8**