MULTIMOADAL TRANSFORMER FOR UNALIGNED MULTIMODAL LAGUAGE SEQUENCES

2021.08.12 REVIEW

CONTENTS

INTRODUCTION

- multimodal : language, vision, audio

- issue

1) inherent data non-alignment due to variable sampling rates for the sequences from each modality

2) long-range dependencies between elements across modalities

- propose : Multimodal Transformer (MulT)

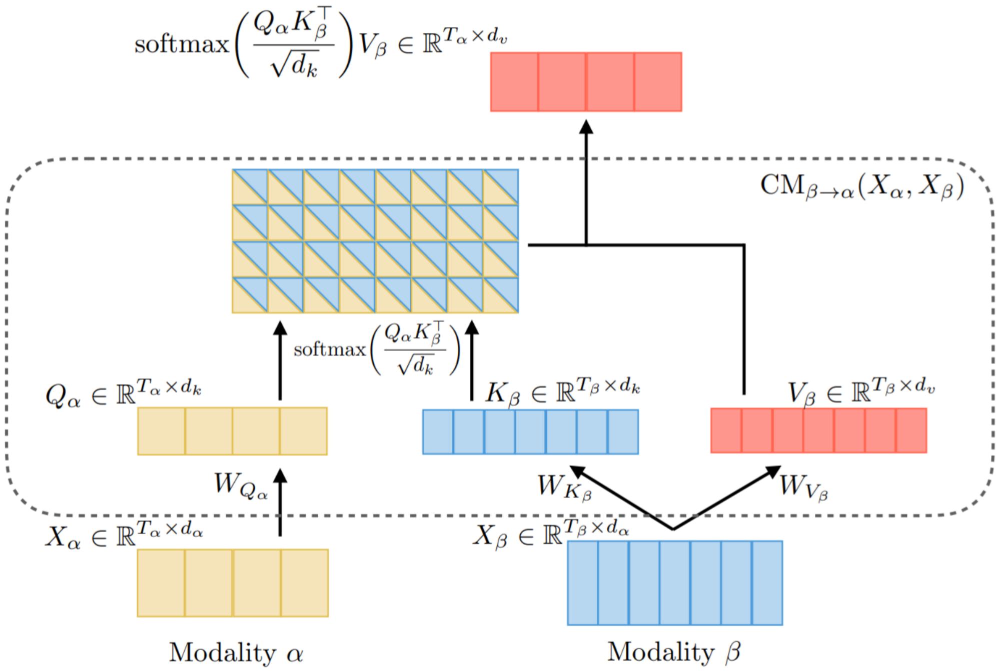
RELATED WORKS

1) Human multimodal language analysis

2) Transformer Network

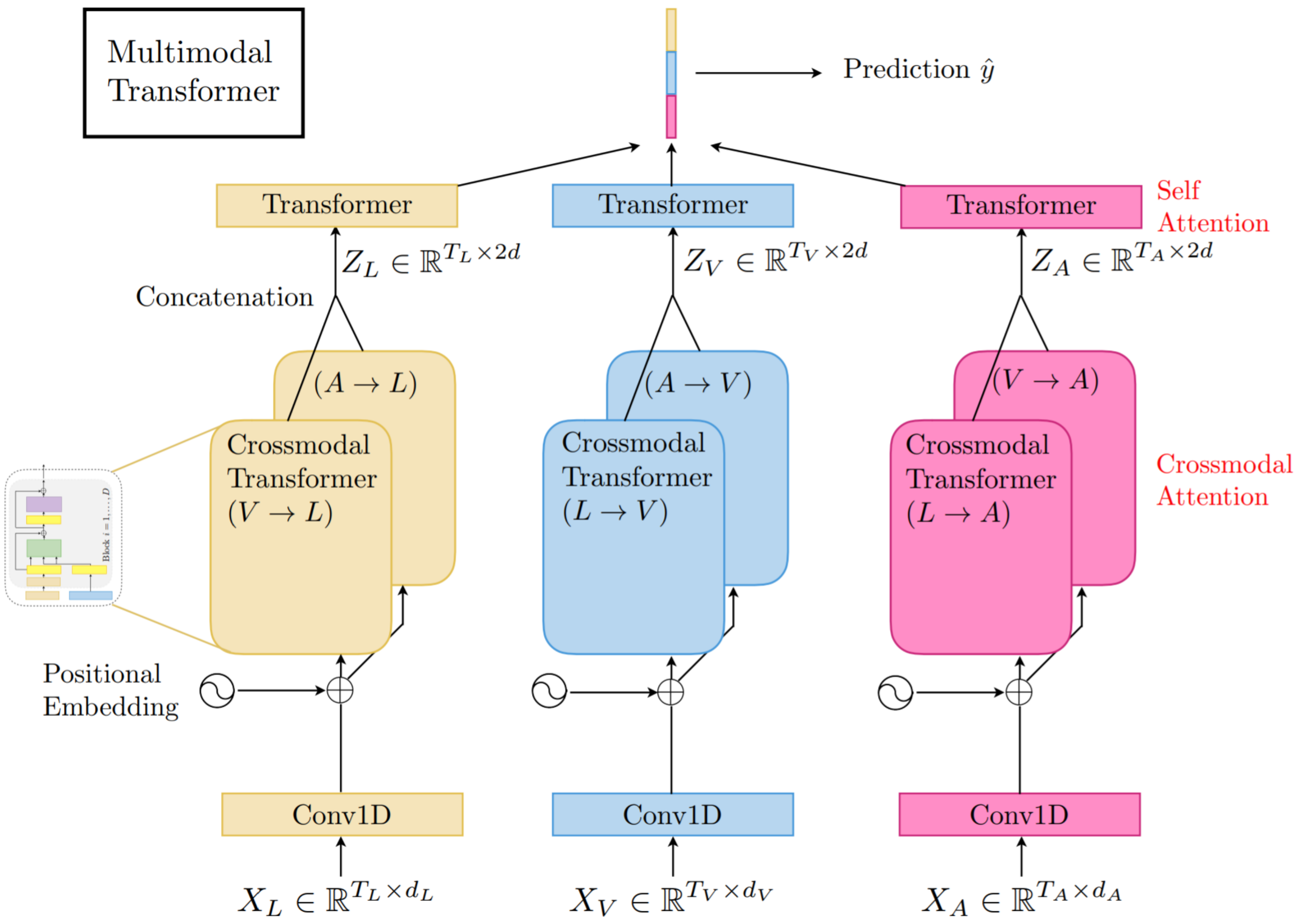
PROPOSED METHOD

- crossmodal attention



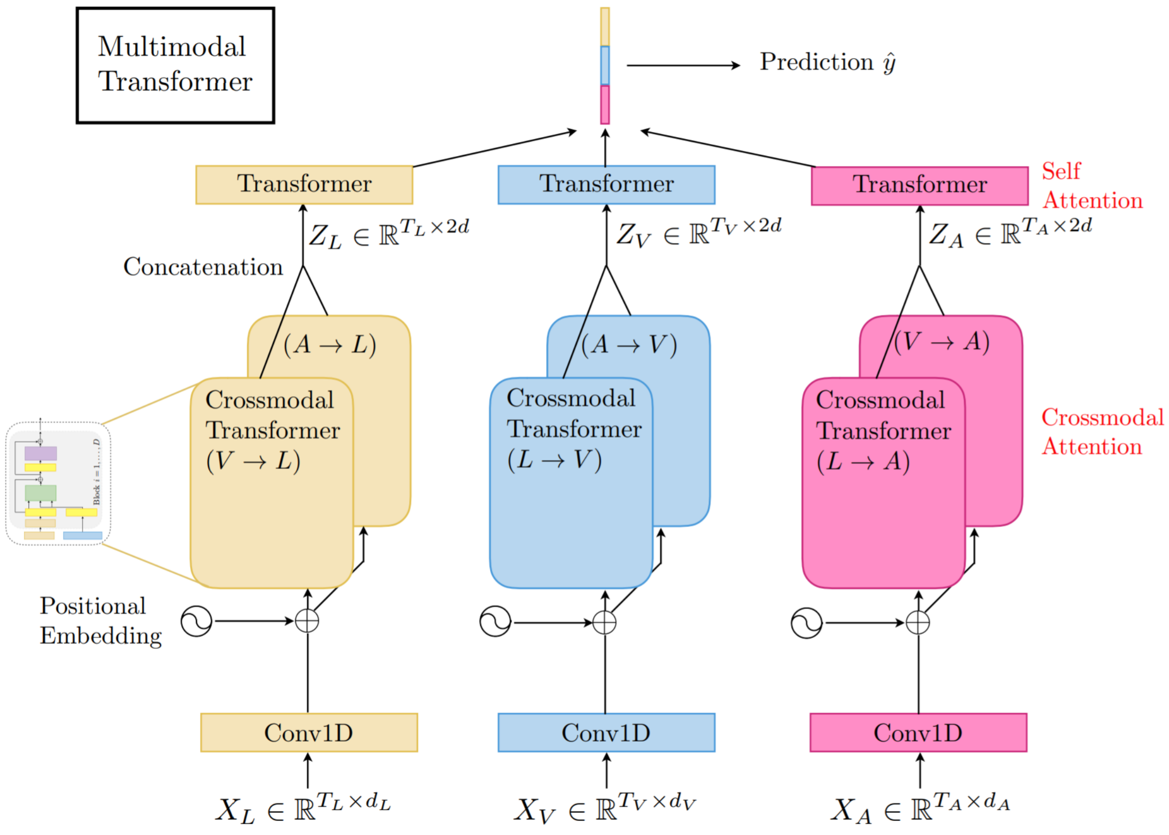
(a) Crossmodal attention $\mathrm{CM}_{\beta\to\alpha}(X_\alpha, X_\beta)$ between sequences X_α, X_β from distinct modalities.

PROPOSED METHOD



PROPSOED METHOD
(1) temporal convolution (CONV1D)

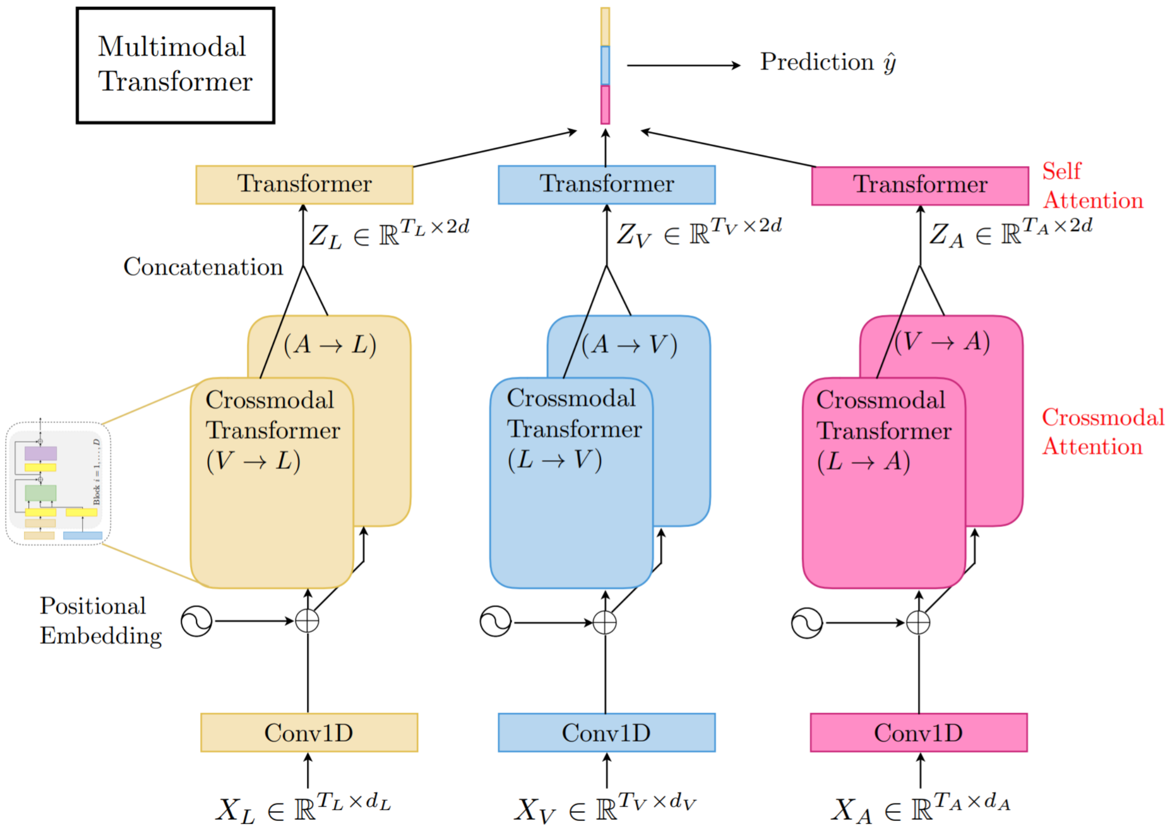
- contain the local structure of the sequence
- poject to the same dimension d



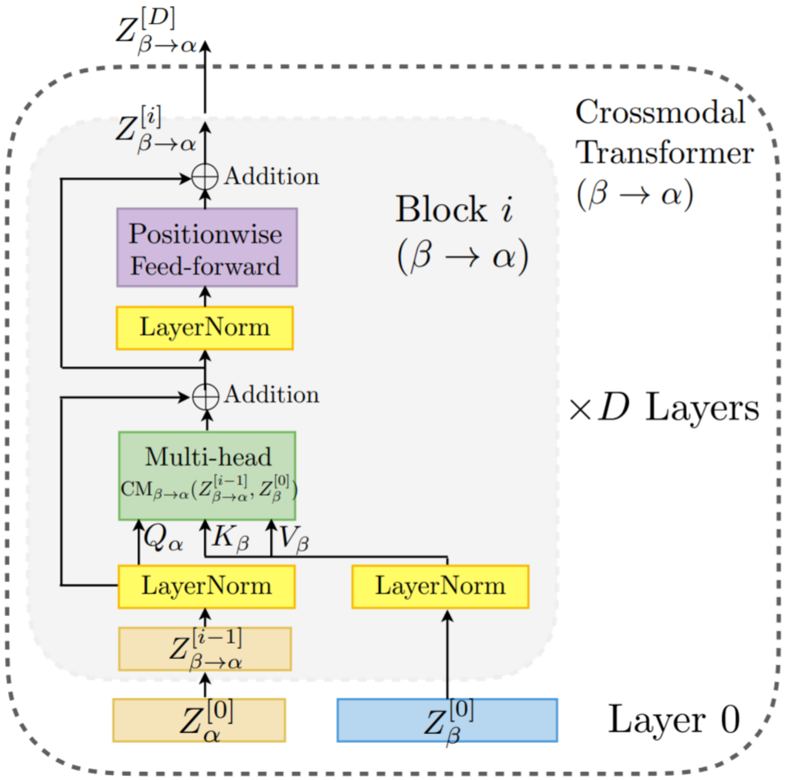
PROPSOED METHOD
(2) Positional embedding

$$\mathrm{PE}[i, 2j] = \sin\left(\frac{i}{10000^{\frac{2j}{d}}}\right)$$

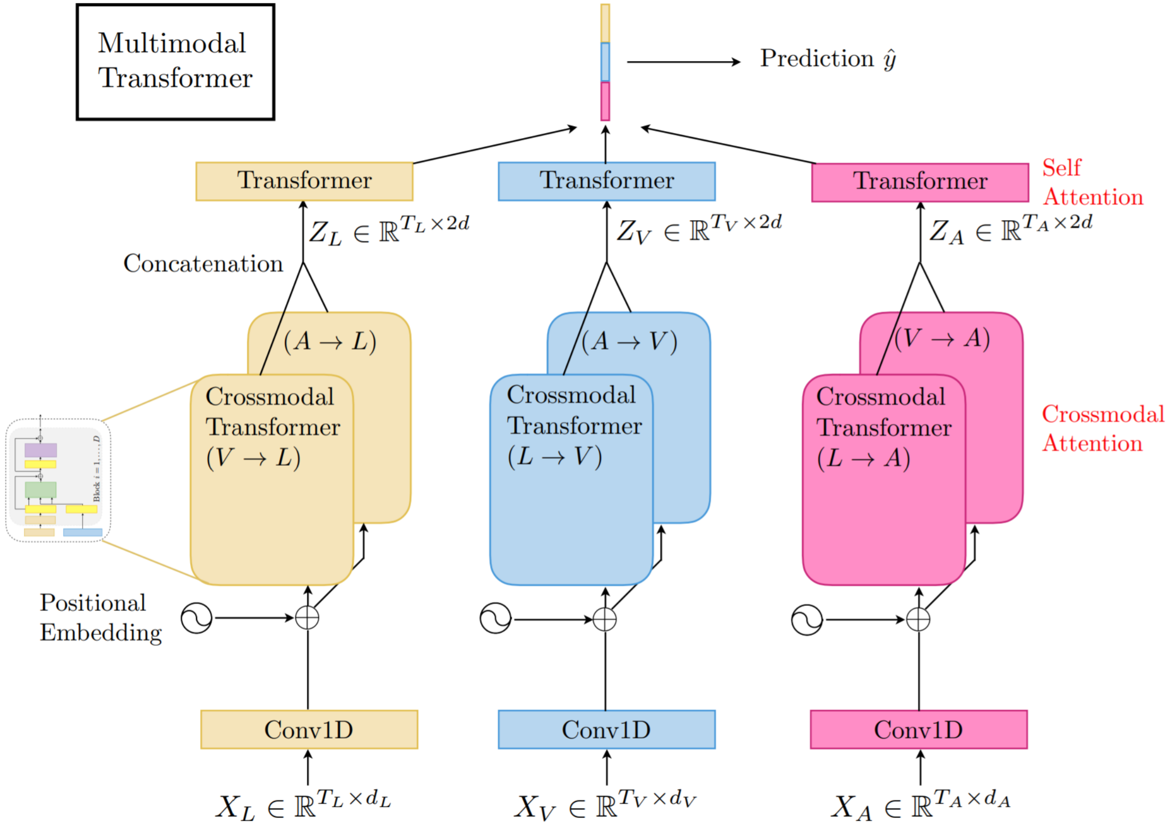
$$\mathrm{PE}[i, 2j+1] = \cos\left(\frac{i}{10000^{\frac{2j}{d}}}\right)$$



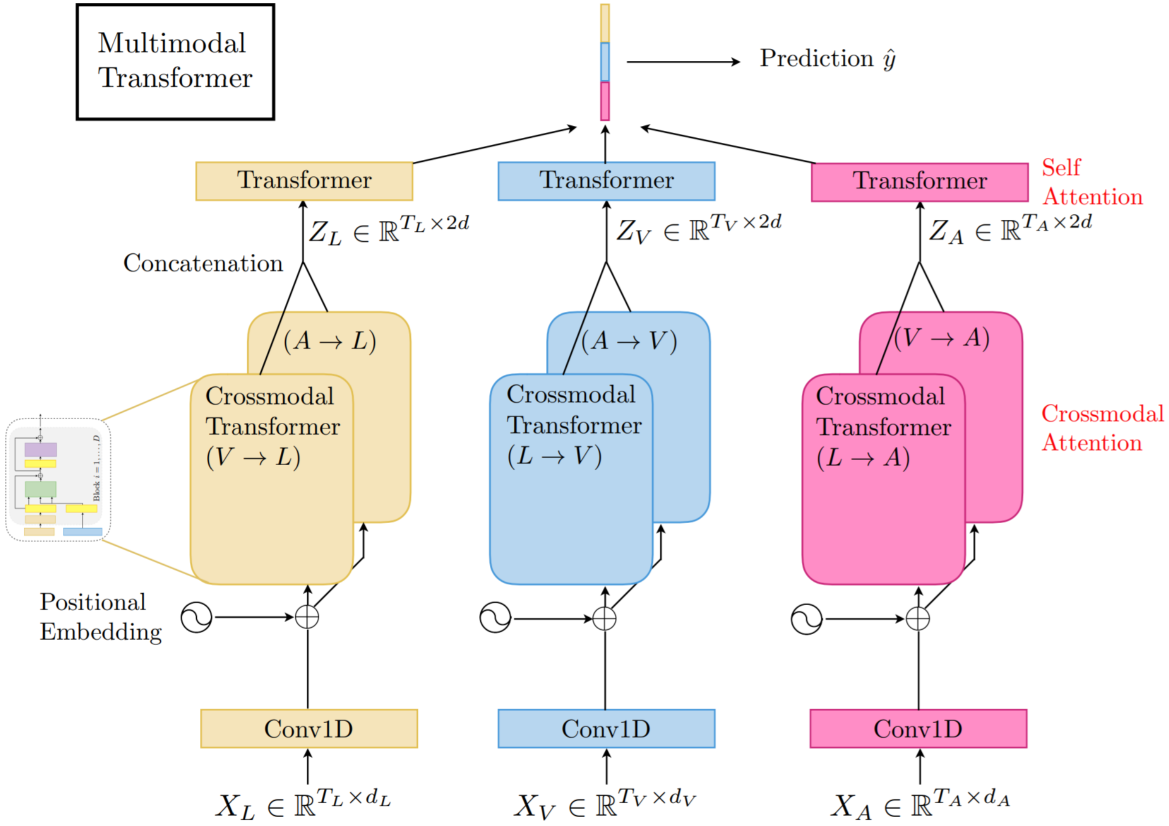
PROPOSED METHOD
(3) Crossmodal Transformers



(b) A crossmodal transformer is a deep stacking of several crossmodal attention blocks.



PROPOSED METHOD
(4) Self-Attention Transformers and Prediction



EXPERIMENT

Table 1: Results for multimodal sentiment analysis on CMU-MOSI with aligned and non-aligned multimodal sequences. ^{h} means higher is better and $^\ell$ means lower is better. EF stands for early fusion, and LF stands for late fusion.

Metric	Acc_7^h	Acc_2^h	F1^h	MAE^ℓ	Corr^h
(Word Aligned) CMU-MOSI Sentiment					
EF-LSTM	33.7	75.3	75.2	1.023	0.608
LF-LSTM	35.3	76.8	76.7	1.015	0.625
RMFN (Liang et al., 2018)	38.3	78.4	78.0	0.922	0.681
MFM (Tsai et al., 2019)	36.2	78.1	78.1	0.951	0.662
RAVEN (Wang et al., 2019)	33.2	78.0	76.6	0.915	**0.691**
MCTN (Pham et al., 2019)	35.6	79.3	79.1	0.909	0.676
MulT (ours)	**40.0**	**83.0**	**82.8**	**0.871**	**0.698**
(Unaligned) CMU-MOSI Sentiment					
CTC (Graves et al., 2006) + EF-LSTM	31.0	73.6	74.5	1.078	0.542
LF-LSTM	33.7	77.6	77.8	0.988	0.624
CTC + MCTN (Pham et al., 2019)	32.7	75.9	76.4	0.991	0.613
CTC + RAVEN (Wang et al., 2019)	31.7	72.7	73.1	1.076	0.544
MulT (ours)	**39.1**	**81.1**	**81.0**	**0.889**	**0.686**

Table 2: Results for multimodal sentiment analysis on (relatively large scale) CMU-MOSEI with aligned and non-aligned multimodal sequences.

Metric	Acc_7^h	Acc_2^h	F1^h	MAE^ℓ	Corr^h
(Word Aligned) CMU-MOSEI Sentiment					
EF-LSTM	47.4	78.2	77.9	0.642	0.616
LF-LSTM	48.8	80.6	80.6	0.619	0.659
Graph-MFN (Zadeh et al., 2018b)	45.0	76.9	77.0	0.71	0.54
RAVEN (Wang et al., 2019)	50.0	79.1	79.5	0.614	0.662
MCTN (Pham et al., 2019)	49.6	79.8	80.6	0.609	0.670
MulT (ours)	**51.8**	**82.5**	**82.3**	**0.580**	**0.703**
(Unaligned) CMU-MOSEI Sentiment					
CTC (Graves et al., 2006) + EF-LSTM	46.3	76.1	75.9	0.680	0.585
LF-LSTM	48.8	77.5	78.2	0.624	0.656
CTC + RAVEN (Wang et al., 2019)	45.5	75.4	75.7	0.664	0.599
CTC + MCTN (Pham et al., 2019)	48.2	79.3	79.7	0.631	0.645
MulT (ours)	**50.7**	**81.6**	**81.6**	**0.591**	**0.694**

EXPERIMENT

Table 3: Results for multimodal emotions analysis on IEMOCAP with aligned and non-aligned multimodal sequences.

Task	Happy		Sad		Angry		Neutral	
Metric	Acc^h	$F1^h$	Acc^h	$F1^h$	Acc^h	$F1^h$	Acc^h	$F1^h$
(Word Aligned) IEMOCAP Emotions								
EF-LSTM	86.0	84.2	80.2	80.5	85.2	84.5	67.8	67.1
LF-LSTM	85.1	86.3	78.9	81.7	84.7	83.0	67.1	67.6
RMFN (Liang et al., 2018)	87.5	85.8	83.8	82.9	85.1	84.6	69.5	69.1
MFM (Tsai et al., 2019)	90.2	85.8	**88.4**	**86.1**	**87.5**	86.7	72.1	68.1
RAVEN (Wang et al., 2019)	87.3	85.8	83.4	83.1	**87.3**	86.7	69.7	69.3
MCTN (Pham et al., 2019)	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0
MulT (ours)	**90.7**	**88.6**	86.7	**86.0**	**87.4**	**87.0**	**72.4**	**70.7**
(Unaligned) IEMOCAP Emotions								
CTC (Graves et al., 2006) + EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
LF-LSTM	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2
CTC + RAVEN (Wang et al., 2019)	77.0	76.8	67.6	65.6	65.0	64.1	**62.0**	**59.5**
CTC + MCTN (Pham et al., 2019)	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
MulT (ours)	**84.8**	**81.9**	**77.7**	**74.1**	**73.9**	**70.2**	**62.5**	**59.7**

EXPERIMENT

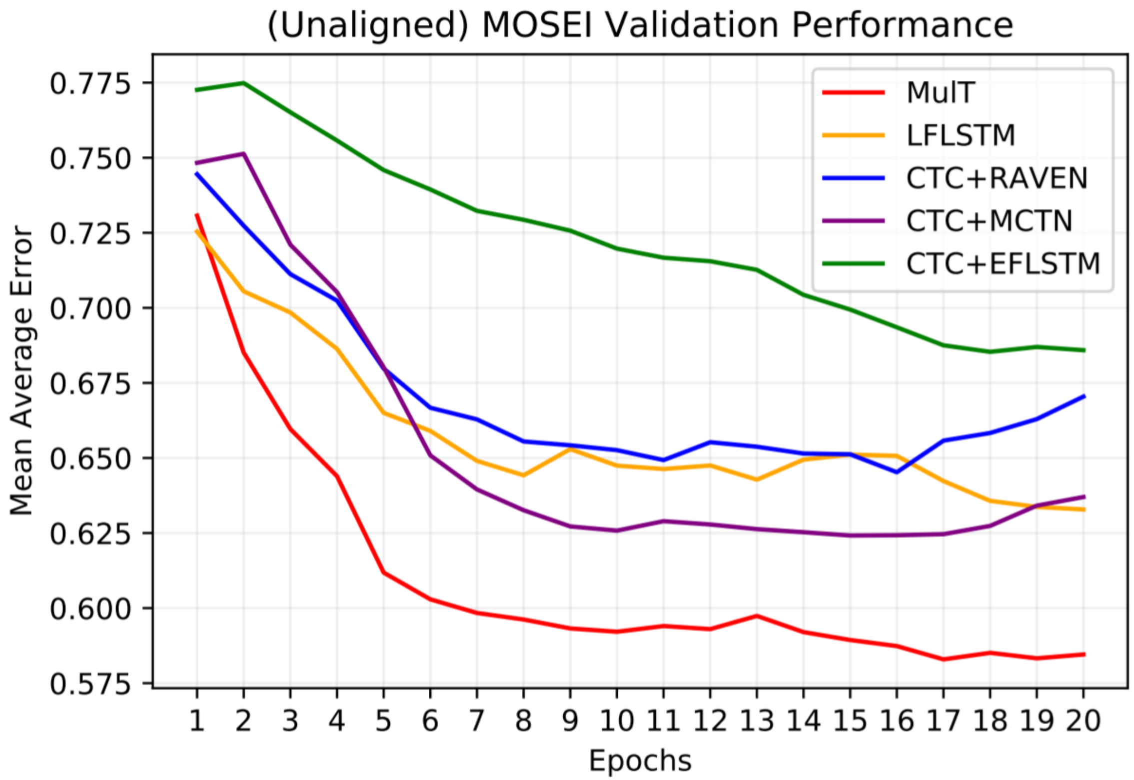


Figure 5: Validation set convergence of MulT when compared to other baselines on the unaligned CMU-MOSEI task.

EXPERIMENT

Table 4: An ablation study on the benefit of MulT's cross-modal transformers using CMU-MOSEI.).

Description	(Unaligned) CMU-MOSEI Sentiment				
	Acc_7^h	Acc_2^h	$\mathrm{F1}^h$	MAE^ℓ	Corr^h
Unimodal Transformers					
Language only	46.5	77.4	78.2	0.653	0.631
Audio only	41.4	65.6	68.8	0.764	0.310
Vision only	43.5	66.4	69.3	0.759	0.343
Late Fusion by using Multiple Unimodal Transformers					
LF-Transformer	47.9	78.6	78.5	0.636	0.658
Temporally Concatenated Early Fusion Transformer					
EF-Transformer	47.8	78.9	78.8	0.648	0.647
Multimodal Transfomers					
Only $[V, A \rightarrow L]$ (ours)	**50.5**	80.1	80.4	0.605	0.670
Only $[L, A \rightarrow V]$ (ours)	48.2	79.7	80.2	0.611	0.651
Only $[L, V \rightarrow A]$ (ours)	47.5	79.2	79.7	0.620	0.648
MulT mixing intermediate-level features (ours)	50.3	80.5	80.6	0.602	0.674
MulT (ours)	**50.7**	**81.6**	**81.6**	**0.591**	**0.691**

EXPERIMENT

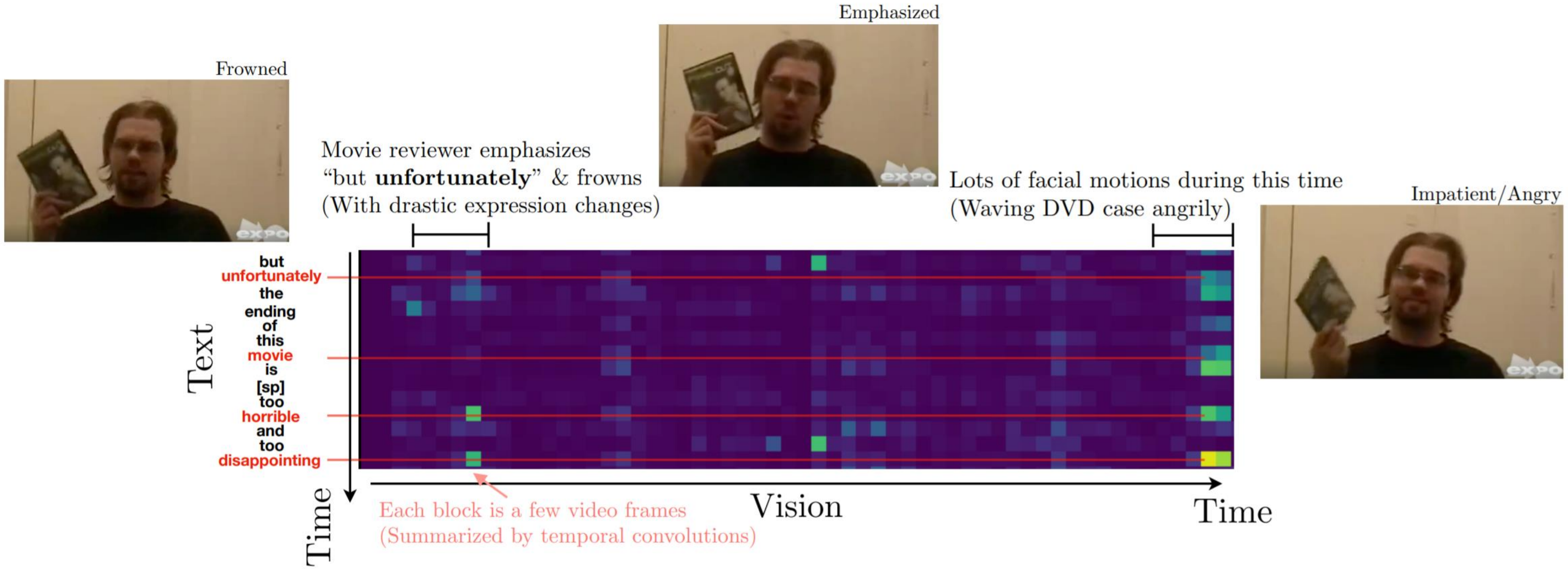


Figure 6: Visualization of sample crossmodal attention weights from layer 3 of $[V \rightarrow L]$ crossmodal transformer on CMU-MOSEI. We found that the crossmodal attention has learned to correlate certain meaningful words (e.g., “movie”, “disappointing”) with segments of stronger visual signals (typically stronger facial motions or expression change), despite the lack of alignment between original L/V sequences. Note that due to temporal convolution, each textual/visual feature contains the representation of nearby elements.