Multimodal Transformer for Unaligned Multimodal Language Sequences

Yao-Hung Hubert Tsai[†*], **Shaojie Bai**[†*], **Paul Pu Liang**[†],
J. Zico Kolter[†‡], **Louis-Philippe Morency**[†], **Ruslan Salakhutdinov**[†]
[†] Carnegie Mellon University, [‡] Bosch Center for AI
{yaohungt,shaojieb,pliang,zkolter,morency,rsalakhu}@cs.cmu.edu

Abstract

Human language is often multimodal, which comprehends a mixture of natural language, facial gestures, and acoustic behaviors. However, two major challenges in modeling such multimodal human language time-series data exist: 1) inherent data non-alignment due to variable sampling rates for the sequences from each modality; and 2) long-range dependencies between elements across modalities. In this paper, we introduce the Multimodal Transformer (MulT) to generically address the above issues in an end-to-end manner without explicitly aligning the data. At the heart of our model is the directional pairwise crossmodal attention, which attends to interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another. Comprehensive experiments on both aligned and non-aligned multimodal time-series show that our model outperforms state-of-the-art methods by a large margin. In addition, empirical analysis suggests that correlated crossmodal signals are able to be captured by the proposed crossmodal attention mechanism in MulT.

1 Introduction

Human language possesses not only spoken words but also nonverbal behaviors from vision (facial attributes) and acoustic (tone of voice) modalities (Gibson et al., 1994). This rich information provides us the benefit of understanding human behaviors and intents (Manning et al., 2014). Nevertheless, the heterogeneities across modalities often increase the difficulty of analyzing human language. For example, the receptors for audio and vision streams may vary with variable receiving frequency, and hence we may not obtain optimal mapping between them. A frowning face may relate to a pessimistically word spoken in the past.

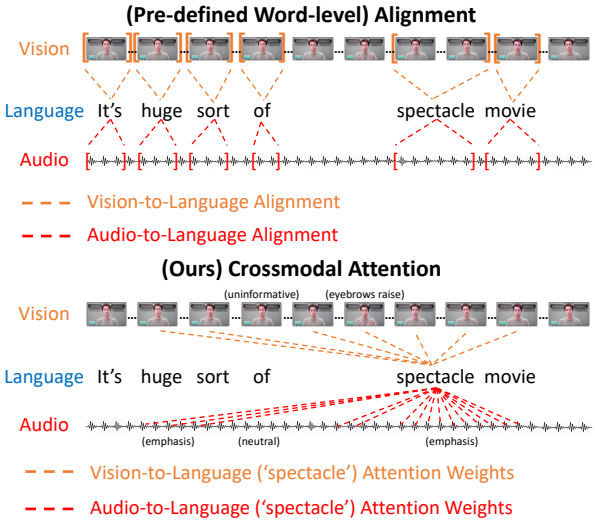


Figure 1: Example video clip from movie reviews. [Top]: Illustration of word-level alignment where video and audio features are averaged across the time interval of each spoken word. [Bottom] Illustration of crossmodal attention weights between text ("spectacle") and vision/audio.

That is to say, multimodal language sequences often exhibit "unaligned" nature and require inferring long term dependencies across modalities, which raises a question on performing efficient multimodal fusion.

To address the above issues, in this paper we propose the Multimodal Transformer (MulT), an end-to-end model that extends the standard Transformer network (Vaswani et al., 2017) to learn representations directly from unaligned multimodal streams. At the heart of our model is the crossmodal attention module, which attends to the crossmodal interactions at the scale of the entire utterances. This module latently adapts streams from one modality to another (e.g., vision $\rightarrow$ language) by repeated reinforcing one modality's features with those from the other modalities, re-

*equal contribution.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6558–6569
Florence, Italy, July 28 - August 2, 2019.

gardless of the need for alignment. In comparison, one common way of tackling unaligned multimodal sequence is by forced word-aligning before training (Poria et al., 2017; Zadeh et al., 2018a,b; Tsai et al., 2019; Pham et al., 2019; Gu et al., 2018): manually preprocess the visual and acoustic features by aligning them to the resolution of words. These approaches would then model the multimodal interactions on the (already) aligned time steps and thus do not directly consider long-range crossmodal contingencies of the original features. We note that such word-alignment not only requires feature engineering that involves domain knowledge; but in practice, it may also not always be feasible, as it entails extra meta-information about the datasets (e.g., the exact time ranges of words or speech utterances). We illustrate the difference between the word-alignment and the crossmodal attention inferred by our model in Figure 1.

For evaluation, we perform a comprehensive set of experiments on three human multimodal language benchmarks: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b), and IEMOCAP (Busso et al., 2008). Our experiments show that MulT achieves the state-of-the-art (SOTA) results in not only the commonly evaluated word-aligned setting but also the more challenging unaligned scenario, outperforming prior approaches by a margin of 5%-15% on most of the metrics. In addition, empirical qualitative analysis further suggests that the crossmodal attention used by MulT is capable of capturing correlated signals across asynchronous modalities.

2 Related Works

Human Multimodal Language Analysis. Prior work for analyzing human multimodal language lies in the domain of inferring representations from multimodal sequences spanning language, vision, and acoustic modalities. Unlike learning multimodal representations from static domains such as image and textual attributes (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012), human language contains time-series and thus requires fusing time-varying signals (Liang et al., 2018; Tsai et al., 2019). Earlier work used early fusion approach to concatenate input features from different modalities (Lazaridou et al., 2015; Ngiam et al., 2011) and showed improved performance as compared to learning from a single modality. More recently, more advanced models were proposed to learn representations of human multimodal language. For example, Gu et al. (2018) used hierarchical attention strategies to learn multimodal representations, Wang et al. (2019) adjusted the word representations using accompanying non-verbal behaviors, Pham et al. (2019) learned robust multimodal representations using a cyclic translation objective, and Dumpala et al. (2019) explored cross-modal autoencoders for audio-visual alignment. These previous approaches relied on the assumption that multimodal language sequences are already aligned in the resolution of words and considered only short-term multimodal interactions. In contrast, our proposed method requires no alignment assumption and defines crossmodal interactions at the scale of the entire sequences.

Transformer Network. Transformer network (Vaswani et al., 2017) was first introduced for neural machine translation (NMT) tasks, where the encoder and decoder side each leverages a *self-attention* (Parikh et al., 2016; Lin et al., 2017; Vaswani et al., 2017) transformer. After each layer of the self-attention, the encoder and decoder are connected by an additional decoder sublayer where the decoder attends to each element of the source text for each element of the target text. We refer the reader to (Vaswani et al., 2017) for a more detailed explanation of the model. In addition to NMT, transformer networks have also been successfully applied to other tasks, including language modeling (Dai et al., 2018; Baevski and Auli, 2019), semantic role labeling (Strubell et al., 2018), word sense disambiguation (Tang et al., 2018), learning sentence representations (Devlin et al., 2018), and video activity recognition (Wang et al., 2018).

This paper absorbs a strong inspiration from the NMT transformer to extend to a multimodal setting. Whereas the NMT transformer focuses on unidirectional *translation* from source to target texts, human multimodal language time-series are neither as well-represented nor discrete as word embeddings, with sequences of each modality having vastly different frequencies. Therefore, we propose not to explicitly translate from one modality to the others (which could be extremely challenging), but to *latently* adapt elements across modalities via the attention. Our model (MulT) therefore has no encoder-decoder structure, but it

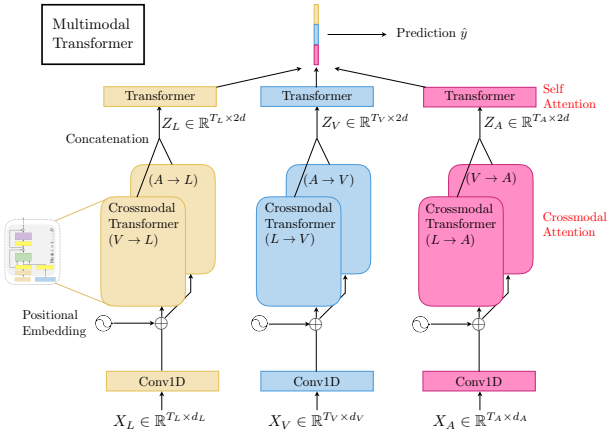


Figure 2: Overall architecture for MulT on modalities (L, V, A). The crossmodal transformers, which suggests latent crossmodal adaptations, are the core components of MulT for multimodal fusion.

is built up from multiple stacks of pairwise and bidirectional crossmodal attention blocks that directly attend to low-level features (while removing the self-attention). Empirically, we show that our proposed approach improves beyond standard transformer on various human multimodal language tasks.

3 Proposed Method

In this section, we describe our proposed Multimodal Transformer (MulT) (Figure 2) for modeling unaligned multimodal language sequences. At the high level, MulT merges multimodal time-series via a feed-forward fusion process from multiple directional pairwise crossmodal transformers. Specifically, each crossmodal transformer (introduced in Section 3.2) serves to repeatedly reinforce a *target modality* with the low-level features from another *source modality* by learning the attention across the two modalities' features. A MulT architecture hence models all pairs of modalities with such crossmodal transformers, followed by sequence models (e.g., self-attention transformer) that predicts using the fused features.

The core of our proposed model is crossmodal attention module, which we first introduce in Section 3.1. Then, in Section 3.2 and 3.3, we present in details the various ingredients of the MulT architecture (see Figure 2) and discuss the difference between crossmodal attention and classical multimodal alignment.

3.1 Crossmodal Attention

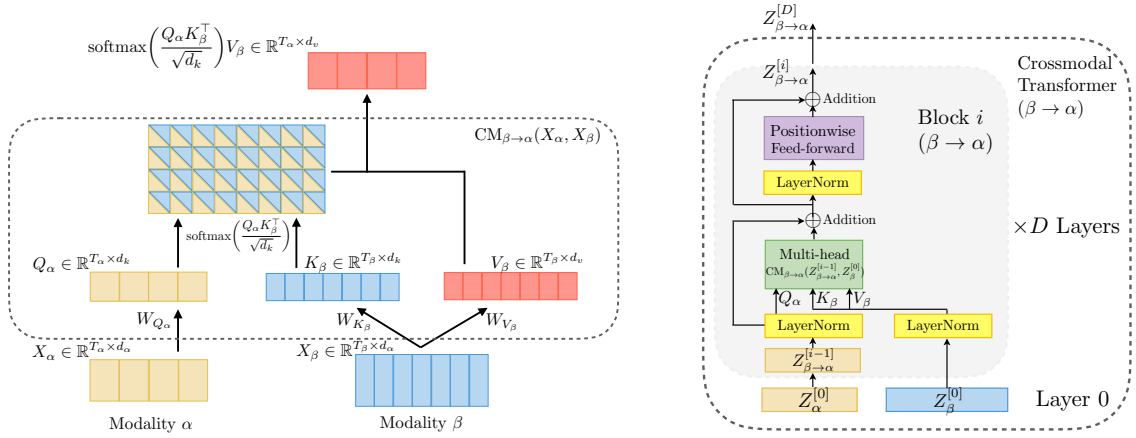
We consider two modalities α and β, with two (potentially non-aligned) sequences from each of them denoted $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, respectively. For the rest of the paper, $T_{(\cdot)}$ and $d_{(\cdot)}$ are used to represent sequence length and feature dimension, respectively. Inspired by the decoder transformer in NMT (Vaswani et al., 2017) that translates one language to another, we hypothesize a good way to fuse crossmodal information is providing a latent adaptation across modalities; i.e., β to α. Note that the modalities consider in our paper may span very different domains such as facial attributes and spoken words.

We define the Querys as $Q_\alpha = X_\alpha W_{Q_\alpha}$, Keys as $K_\beta = X_\beta W_{K_\beta}$, and Values as $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}, W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are weights. The latent adaptation from β to α is presented as the crossmodal attention $Y_\alpha := \text{CM}_{\beta \to \alpha}(X_\alpha, X_B) \in \mathbb{R}^{T_\alpha \times d_v}$:

$$\begin{aligned} Y_\alpha &= \text{CM}_{\beta \to \alpha}(X_\alpha, X_\beta) \\ &= \text{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right) V_\beta \\ &= \text{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta}. \end{aligned} \tag{1}$$

Note that Y_α has the same length as Q_α (i.e., T_α), but is meanwhile represented in the feature space of V_β. Specifically, the scaled (by $\sqrt{d_k}$) softmax in Equation (1) computes a score matrix $\text{softmax}(\cdot) \in \mathbb{R}^{T_\alpha \times T_\beta}$, whose (i, j)-th entry measures the attention given by the i-th time step of modality α to the j-th time step of modality β. Hence, the i-th time step of Y_α is a weighted summary of V_β, with the weight determined by i-th row in $\text{softmax}(\cdot)$. We call Equation (1) a *single-head* crossmodal attention, which is illustrated in Figure 3(a).

Following prior works on transformers (Vaswani et al., 2017; Chen et al., 2018; Devlin et al., 2018; Dai et al., 2018), we add a residual connection to the crossmodal attention computation. Then, another positionwise feed-forward sublayer is injected to complete a *crossmodal attention block* (see Figure 3(b)). Each crossmodal attention block adapts directly from the low-level feature sequence (i.e., $Z_\beta^{[0]}$ in Figure 3(b)) and does not rely on self-attention, which makes it different from the NMT encoder-decoder architecture (Vaswani et al., 2017; Shaw et al., 2018) (i.e., taking intermediate-level features). We argue that performing adaptation



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

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

Figure 3: Architectural elements of a crossmodal transformer between two time-series from modality α and β.

from low-level feature benefits our model to preserve the low-level information for each modality. We leave the empirical study for adapting from intermediate-level features (i.e., $Z_\beta^{[i-1]}$) in Ablation Study in Section 4.3.

3.2 Overall Architecture

Three major modalities are typically involved in multimodal language sequences: language (L), video (V), and audio (A) modalities. We denote with $X_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ the input feature sequences (and the dimensions thereof) from these 3 modalities. With these notations, in this subsection, we describe in greater details the components of Multimodal Transformer and how crossmodal attention modules are applied.

Temporal Convolutions. To ensure that each element of the input sequences has sufficient awareness of its neighborhood elements, we pass the input sequences through a 1D temporal convolutional layer:

$$\hat{X}_{\{L,V,A\}} = \mathrm{Conv1D}(X_{\{L,V,A\}}, k_{\{L,V,A\}}) \in \mathbb{R}^{T_{\{L,V,A\}} \times d} \tag{2}$$

where $k_{\{L,V,A\}}$ are the sizes of the convolutional kernels for modalities $\{L, V, A\}$, and d is a common dimension. The convolved sequences are expected to contain the local structure of the sequence, which is important since the sequences are collected at different sampling rates. Moreover, since the temporal convolutions project the features of different modalities to the same dimension d, the dot-products are admittable in the crossmodal attention module.

Positional Embedding. To enable the sequences to carry temporal information, following (Vaswani et al., 2017), we augment positional embedding (PE) to $\hat{X}_{\{L,V,A\}}$:

$$Z^{[0]}_{\{L,V,A\}} = \hat{X}_{\{L,V,A\}} + \mathrm{PE}(T_{\{L,V,A\}}, d) \tag{3}$$

where $\mathrm{PE}(T_{\{L,V,A\}}, d) \in \mathbb{R}^{T_{\{L,V,A\}} \times d}$ computes the (fixed) embeddings for each position index, and $Z^{[0]}_{\{L,V,A\}}$ are the resulting low-level position-aware features for different modalities. We leave more details of the positional embedding to Appendix A.

Crossmodal Transformers. Based on the crossmodal attention blocks, we design the crossmodal transformer that enables one modality for receiving information from another modality. In the following, we use the example for passing vision (V) information to language (L), which is denoted by "$V \to L$". We fix all the dimensions ($d_{\{\alpha,\beta,k,v\}}$) for each crossmodal attention block as d.

Each crossmodal transformer consists of D layers of crossmodal attention blocks (see Figure 3(b)). Formally, a crossmodal transformer computes feed-forwardly for $i = 1, \ldots, D$ layers:

$$\begin{aligned} Z^{[0]}_{V\to L} &= Z^{[0]}_L \\ \hat{Z}^{[i]}_{V\to L} &= \mathrm{CM}^{[i],\mathrm{mul}}_{V\to L}(\mathrm{LN}(Z^{[i-1]}_{V\to L}), \mathrm{LN}(Z^{[0]}_V)) + \mathrm{LN}(Z^{[i-1]}_{V\to L}) \\ Z^{[i]}_{V\to L} &= f_{\theta^{[i]}_{V\to L}}(\mathrm{LN}(\hat{Z}^{[i]}_{V\to L})) + \mathrm{LN}(\hat{Z}^{[i]}_{V\to L}) \end{aligned} \tag{4}$$

where f_θ is a positionwise feed-forward sublayer parametrized by θ, and $\mathrm{CM}^{[i],\mathrm{mul}}_{V\to L}$ means a multi-head (see (Vaswani et al., 2017) for more details) version of $\mathrm{CM}_{V\to L}$ at layer i (note: d should be

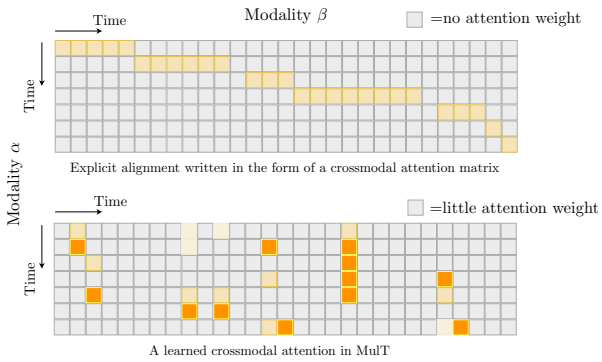


Figure 4: An example of visualizing alignment using attention matrix from modality β to α. Multimodal alignment is a special (monotonic) case for crossmodal attention.

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 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.

divisible by the number of heads). LN means layer normalization (Ba et al., 2016).

In this process, each modality keeps updating its sequence via low-level external information from the multi-head crossmodal attention module. At every level of the crossmodal attention block, the low-level signals from source modality are transformed to a different set of Key/Value pairs to interact with the target modality. Empirically, we find that the crossmodal transformer learns to correlate meaningful elements across modalities (see Section 4 for details). The eventual MulT is based on modeling every pair of crossmodal interactions. Therefore, with 3 modalities (i.e., L, V, A) in consideration, we have 6 crossmodal transformers in total (see Figure 2).

Self-Attention Transformers and Prediction. As a final step, we concatenate the outputs from the crossmodal transformers that share the same target modality to yield $Z_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times 2d}$. For example, $Z_L = [Z_{V\to L}^{[D]}; Z_{A\to L}^{[D]}]$. Each of them is then passed through a sequence model to collect temporal information to make predictions. We choose the self-attention transformer (Vaswani et al., 2017). Eventually, the last elements of the sequences models are extracted to pass through fully-connected layers to make predictions.

3.3 Discussion about Attention & Alignment

When modeling unaligned multimodal language sequences, MulT relies on crossmodal attention blocks to merge signals across modalities. While the multimodal sequences were (manually) aligned to the same length in prior works before training (Zadeh et al., 2018b; Liang et al., 2018; Tsai et al., 2019; Pham et al., 2019; Wang et al., 2019), we note that MulT looks at the non-alignment issue through a completely different lens. Specifically, for MulT, the correlations between elements of multiple modalities are purely based on attention. In other words, MulT does not handle modality non-alignment by (simply) aligning them; instead, the crossmodal attention encourages the model to directly attend to elements in other modalities where strong signals or relevant information is present. As a result, MulT can capture long-range crossmodal contingencies in a way that conventional alignment could not easily reveal. Classical crossmodal alignment, on the other hand, can be expressed as a special (step diagonal) crossmodal attention matrix (i.e., monotonic attention (Yu et al., 2016)). We illustrate their differences in Figure 4.

4 Experiments

In this section, we empirically evaluate the Multimodal Transformer (MulT) on three datasets that are frequently used to benchmark human multimodal affection recognition in prior works (Pham et al., 2019; Tsai et al., 2019; Liang et al., 2018). Our goal is to compare MulT with prior competitive approaches on both *word-aligned* (by word, which almost all prior works employ) and *unaligned* (which is more challenging, and which MulT is generically designed for) multimodal language 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**

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

4.1 Datasets and Evaluation Metrics

Each task consists of a *word-aligned* (processed in the same way as in prior works) and an *unaligned* version. For both versions, the multimodal features are extracted from the textual (GloVe word embeddings (Pennington et al., 2014)), visual (Facet (iMotions, 2017)), and acoustic (COVAREP (Degottex et al., 2014)) data modalities. A more detailed introduction to the features is included in Appendix.

For the word-aligned version, following (Zadeh et al., 2018a; Tsai et al., 2019; Pham et al., 2019), we first use P2FA (Yuan and Liberman, 2008) to obtain the aligned timesteps (segmented w.r.t. words) for audio and vision streams, and we then perform averaging on the audio and vision features within these time ranges. All sequences in the word-aligned case have length 50. The process remains the same across all the datasets. On the other hand, for the unaligned version, we keep the original audio and visual features as extracted, without any word-segmented alignment or manual subsampling. As a result, the lengths of each modality vary significantly, where audio and vision sequences may contain up to $> 1,000$ time steps. We elaborate on the three tasks below.

CMU-MOSI & MOSEI. CMU-MOSI (Zadeh et al., 2016) is a human multimodal sentiment analysis dataset consisting of 2,199 short monologue video clips (each lasting the duration of a sentence). Acoustic and visual features of CMU-MOSI are extracted at a sampling rate of 12.5 and 15 Hz, respectively (while textual data are segmented per word and expressed as discrete word embeddings). Meanwhile, CMU-MOSEI (Zadeh et al., 2018b) is a sentiment and emotion analysis dataset made up of 23,454 movie review video clips taken from YouTube (about $10\times$ the size of CMU-MOSI). The unaligned CMU-MOSEI sequences are extracted at a sampling rate of 20 Hz for acoustic and 15 Hz for vision signals.

For both CMU-MOSI and CMU-MOSEI, each sample is labeled by human annotators with a sentiment score from -3 (strongly negative) to 3 (strongly positive). We evaluate the model performances using various metrics, in agreement with those employed in prior works: 7-class accuracy (i.e., Acc_7: sentiment score classification in $\mathbb{Z} \cap [-3, 3]$), binary accuracy (i.e., Acc_2: positive/negative sentiments), F1 score, mean absolute error (MAE) of the score, and the correlation of the model's prediction with human. Both tasks are frequently used to benchmark models' ability to fuse multimodal (sentiment) information (Poria et al., 2017; Zadeh et al., 2018a; Liang et al., 2018; Tsai et al., 2019; Pham et al., 2019; Wang et al., 2019).

IEMOCAP. IEMOCAP (Busso et al., 2008) consists of 10K videos for human emotion analysis. As suggested by Wang et al. (2019), 4 emotions (happy, sad, angry and neutral) were selected for emotion recognition. Unlike CMU-MOSI and CMU-MOSEI, this is a multilabel task (e.g., a person can be sad and angry simultaneously). Its multimodal streams consider fixed sampling rate on audio (12.5 Hz) and vision (15 Hz) signals. We follow (Poria et al., 2017; Wang et al., 2019; Tsai et al., 2019) to report the binary classification accuracy and the F1 score of the predictions.

4.2 Baselines

We choose Early Fusion LSTM (EF-LSTM) and Late Fusion LSTM (LF-LSTM) as baseline models, as well as Recurrent Attended Variation Embedding Network (RAVEN) (Wang et al., 2019) and Multimodal Cyclic Translation Network (MCTN) (Pham et al., 2019), that achieved SOTA results on various word-aligned human multimodal language tasks. To compare the models comprehensively, we adapt the *connectionist temporal classification* (CTC) (Graves et al., 2006) method to the prior approaches (e.g., EF-LSTM, MCTN, RAVEN) that cannot be applied directly to the unaligned setting. Specifically, these models train to optimize the CTC alignment

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**

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

objective and the human multimodal objective simultaneously. We leave more detailed treatment of the CTC module to Appendix. For fair comparisons, we control the number of parameters of all models to be approximately the same. The hyperparameters are reported in Appendix. [1]

4.3 Quantitative Analysis

Word-Aligned Experiments. We first evaluate MulT on the *word-aligned sequences*— the "home turf" of prior approaches modeling human multimodal language (Sheikh et al., 2018; Tsai et al., 2019; Pham et al., 2019; Wang et al., 2019). The upper part of the Table 1, 2, and 3 show the results of MulT and baseline approaches on the word-aligned task. With similar model sizes (around 200K parameters), MulT outperforms the other competitive approaches on different metrics on all tasks, with the exception of the "sad" class results on IEMOCAP.

Unaligned Experiments. Next, we evaluate MulT on the same set of datasets in the unaligned setting. Note that MulT can be directly applied to unaligned multimodal stream, while the baseline models (except for LF-LSTM) require the need of additional alignment module (e.g., CTC module).

The results are shown in the bottom part of Table 1, 2, and 3. On the three benchmark datasets, MulT improves upon the prior methods (some with CTC) by 10%-15% on most attributes. Empirically, we find that MulT converges faster to better results at training when compared to other competitive approaches (see Figure 5). In addition, while we note that in general there is a performance drop on all models when we shift from the word-aligned to unaligned multimodal time-series, the impact MulT takes is much smaller than the other approaches. We hypothesize such performance drop occurs because the asynchronous (and much longer) data streams introduce more difficulty in recognizing important features and computing the appropriate attention.

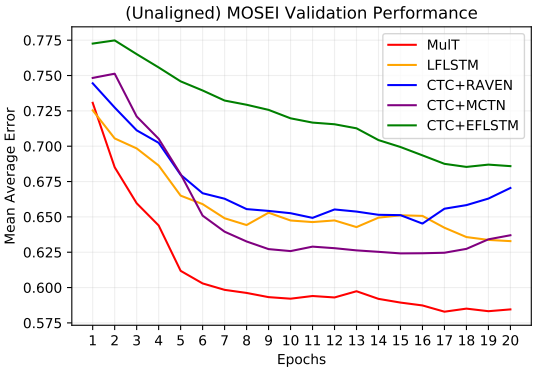


Figure 5: Validation set convergence of MulT when compared to other baselines on the unaligned CMU-MOSEI task.

Ablation Study. To further study the influence of the individual components in MulT, we perform comprehensive ablation analysis using the unaligned version of CMU-MOSEI. The results are shown in Table 4.

First, we consider the performance for only

[1] All experiments are conducted on 1 GTX-1080Ti GPU. The code for our model and experiments can be found in https://github.com/yaohungt/Multimodal-Transformer

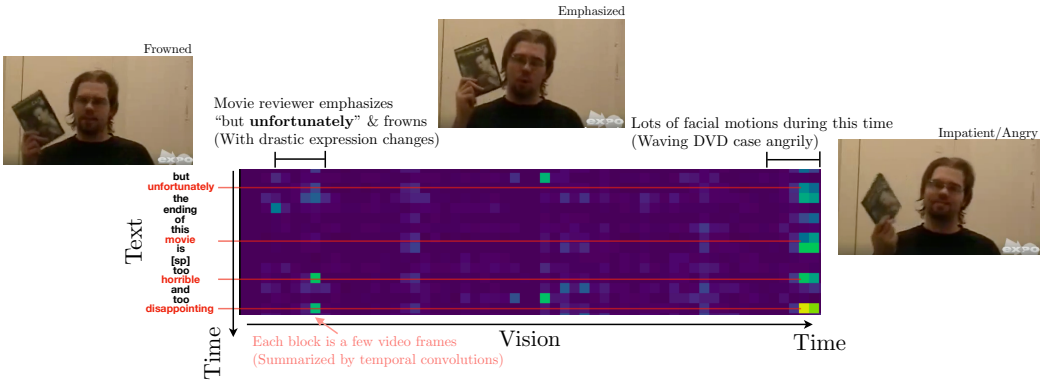


Figure 6: Visualization of sample crossmodal attention weights from layer 3 of $[V \rightarrow L]$ crossmodal transformer on CMU-MOSEI. We found that the crossmodal attention has learned to correlate certain meaningful words (e.g., "movie", "disappointing") with segments of stronger visual signals (typically 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.

Description	(Unaligned) CMU-MOSEI Sentiment				
	Acc_7^h	Acc_2^h	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**

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

using unimodal transformers (i.e., language, audio or vision only). We find that the language transformer outperforms the other two by a large margin. For example, for the Acc_2^h metric, the model improves from 65.6 to 77.4 when comparing audio only to language only unimodal transformer. This fact aligns with the observations in prior work (Pham et al., 2019), where the authors found that a good language network could already achieve good performance at inference time.

Second, we consider 1) a late-fusion transformer that feature-wise concatenates the last elements of three self-attention transformers; and 2) an early-fusion self-attention transformer that takes in a temporal concatenation of three asynchronous sequences $[\hat{X}_L, \hat{X}_V, \hat{X}_A] \in \mathbb{R}^{(T_L+T_V+T_A)\times d_q}$ (see Section 3.2). Empirically, we find that both EF- and LF-Transformer (which fuse multimodal signals) outperform unimodal transformers.

Finally, we study the importance of individual crossmodal transformers according to the target modalities (i.e., using $[V, A \rightarrow L]$, $[L, A \rightarrow V]$, or $[L, V \rightarrow A]$ network). As shown in Table 4, we find crossmodal attention modules consistently improve over the late- and early-fusion transformer models in most metrics on unaligned CMU-MOSEI. In particular, among the three crossmodal transformers, the one where language(L) is the target modality works best. We also additionally study the effect of adapting intermediate-level instead of the low-level features from source modality in crossmodal attention blocks (similar to the NMT encoder-decoder architecture but without self-attention; see Section 3.1). While MulT leveraging intermediate-level features still outperform models in other ablative settings, we empirically find adapting from low-level features works best. The ablations suggest that crossmodal attention concretely benefits MulT with better representation learning.

4.4 Qualitative Analysis

To understand how crossmodal attention works while modeling unaligned multimodal data, we empirically inspect what kind of signals MulT picks up by visualizing the attention activations. Figure 6 shows an example of a section of the crossmodal attention matrix on layer 3 of the $V \rightarrow$

L network of MulT (the original matrix has dimension $T_L \times T_V$; the figure shows the attention corresponding to approximately a 6-sec short window of that matrix). We find that crossmodal attention has learned to attend to meaningful signals across the two modalities. For example, stronger attention is given to the intersection of words that tend to suggest emotions (e.g., "movie", "disappointing") and drastic facial expression changes in the video (start and end of the above vision sequence). This observation advocates one of the aforementioned advantage of MulT over conventional alignment (see Section 3.3): crossmodal attention enables MulT to directly capture potentially long-range signals, including those off-diagonals on the attention matrix.

5 Discussion

In the paper, we propose Multimodal Transformer (MulT) for analyzing human multimodal language. At the heart of MulT is the crossmodal attention mechanism, which provides a latent crossmodal adaptation that fuses multimodal information by directly attending to low-level features in other modalities. Whereas prior approaches focused primarily on the aligned multimodal streams, MulT serves as a strong baseline capable of capturing long-range contingencies, regardless of the alignment assumption. Empirically, we show that MulT exhibits the best performance when compared to prior methods.

We believe the results of MulT on unaligned human multimodal language sequences suggest many exciting possibilities for its future applications (e.g., Visual Question Answering tasks, where the input signals is a mixture of static and time-evolving signals). We hope the emergence of MulT could encourage further explorations on tasks where alignment used to be considered necessary, but where crossmodal attention might be an equally (if not more) competitive alternative.

Acknowledgements

This work was supported in part by DARPA HR00111990016, AFRL FA8750-18-C-0014, NSF IIS1763562 #1750439 #1722822, Apple, Google focused award, and Samsung. We would also like to acknowledge NVIDIA's GPU support.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations (ICLR)*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *ACL*.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2018. Transformer-xl: Language modeling with longer-term dependency.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *ICASSP*. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. 2019. Audio-visual fusion for sentiment classification using cross-modal autoencoder. *NIPS*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

Kathleen R Gibson, Kathleen Rita Gibson, and Tim Ingold. 1994. *Tools, language and cognition in human evolution*. Cambridge University Press.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal affective analysis using hierarchical attention strategy

with word-level alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

iMotions. 2017. Facial expression analysis.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *EMNLP*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Prismatic Inc, Steven J. Bethard, and David Mcclosky. 2014. The stanford corenlp natural language processing toolkit. In *In ACL, System Demonstrations*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Poczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. *AAAI*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations.

Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. 2018. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 35–39.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038. Association for Computational Linguistics.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. *ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *AAAI*.

Lei Yu, Jan Buys, and Phil Blunsom. 2016. Online segment to segment neural transduction. *arXiv preprint arXiv:1609.08194*.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*.

A Positional Embedding

A purely attention-based transformer network is *order-invariant*. In other words, permuting the order of an input sequence does not change transformer's behavior or alter its output. One solution to address this weakness is by embedding the positional information into the hidden units (Vaswani et al., 2017).

Following (Vaswani et al., 2017), we encode the positional information of a sequence of length T via the sin and cos functions with frequencies dictated by the feature index. In particular, we define the positional embedding (PE) of a sequence $X \in \mathbb{R}^{T \times d}$ (where T is length) as a matrix where:

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

for $i = 1, \ldots, T$ and $j = 0, \lfloor \frac{d}{2} \rfloor$. Therefore, each feature dimension (i.e., column) of PE are positional values that exhibit a sinusoidal pattern. Once computed, the positional embedding is added directly to the sequence so that $X + \text{PE}$ encodes the elements' position information at every time step.

B Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) (Graves et al., 2006) was first proposed for unsupervised Speech to Text alignment. Particularly, CTC is often combined with the output of recurrent neural network, which enables the model to train end-to-end and simultaneously infer speech-text alignment without supervision. For the ease of explanation, suppose the CTC module now are aiming at aligning an audio signal sequence $[a_1, a_2, a_3, a_4, a_5, a_6]$ with length 6 to a textual sequence "I am really really happy" with length 5. In this example, we refer to audio as the source and texts as target signal, noting that the sequence lengths may be different between the source to target; we also see that the output sequence may have repetitive element (i.e., "really"). The CTC (Graves et al., 2006) module we use comprises two components: alignment predictor and the CTC loss.

First, the alignment predictor is often chosen as a recurrent networks such as LSTM, which performs on the source sequence then outputs the possibility of being the unique words in the target sequence as well as a empty word (i.e., x). In our example, for each individual audio signal, the alignment predictor provides a vector of length 5 regarding the probability being aligned to [x, 'I', 'am', 'really', 'happy'].

Next, the CTC loss considers the negative log-likelihood loss from only the proper alignment for the alignment predictor outputs. The proper alignment, in our example, can be results such as

i) [x, 'I', 'am', 'really', 'really', 'happy'];

ii) ['I', 'am', x, 'really', 'really', 'happy'];

iii) ['I', 'am', 'really', 'really', 'really', 'happy'];

iv) ['I', 'I', 'am', 'really', 'really', 'happy']

In the meantime, some examples of the suboptimal/failure cases would be

i) [x, x, 'am', 'really', 'really', 'happy'];

ii) ['I', 'am', 'I', 'really', 'really', 'happy'];

iii) ['I', 'am', x, 'really', x, 'happy']

When the CTC loss is minimized, it implies the source signals are properly aligned to target signals.

To sum up, in the experiments that adopting the CTC module, we train the alignment predictor while minimizing the CTC loss. Then, excluding the probability of blank words, we multiply the probability outputs from the alignment predictor to source signals. The source signal is hence resulting in a pseudo-aligned target singal. In our example, the audio signal is then transforming to a audio signal $[a'_1, a'_2, a'_3, a'_4, a'_5]$ with sequence length 5, which is pseudo-aligned to ['I', 'am', 'really', 'really', 'happy'].

C Hyperparameters

Table 5 shows the settings of the various MulTs that we train on human multimodal language tasks. As previously mentioned, the models are contained at roughly the same sizes as in prior works for the purpose of fair comparison. For hyperparameters such as the dropout rate and number of heads in crossmodal attention module, we perform a basic grid search. We decay the learning rate by a factor of 10 when the validation performance plateaus.

	CMU-MOSEI	CMU-MOSI	IEMOCAP
Batch Size	16	128	32
Initial Learning Rate	1e-3	1e-3	2e-3
Optimizer	Adam	Adam	Adam
Transformers Hidden Unit Size d	40	40	40
# of Crossmodal Blocks D	4	4	4
# of Crossmodal Attention Heads	8	10	10
Temporal Convolution Kernel Size ($L/V/A$)	(1 or 3)/3/3	(1 or 3)/3/3	3/3/5
Textual Embedding Dropout	0.3	0.2	0.3
Crossmodal Attention Block Dropout	0.1	0.2	0.25
Output Dropout	0.1	0.1	0.1
Gradient Clip	1.0	0.8	0.8
# of Epochs	20	100	30

Table 5: Hyperparameters of Multimodal Transformer (MulT) we use for the various tasks. The "# of Crossmodal Blocks" and "# of Crossmodal Attention Heads" are for each transformer.

D Features

The features for multimodal datasets are extracted as follows:

- **Language.** We convert video transcripts into pre-trained Glove word embeddings (glove.840B.300d) (Pennington et al., 2014). The embedding is a 300 dimensional vector.

- **Vision.** We use Facet (iMotions, 2017) to indicate 35 facial action units, which records facial muscle movement (Ekman et al., 1980; Ekman, 1992) for representing per-frame basic and advanced emotions.

- **Audio.** We use COVAREP (Degottex et al., 2014) for extracting low level acoustic features. The feature includes 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. Dimension of the feature is 74.