Lightweight Models for Multimodal Sequential Data

Soumya Sourav
The University of Texas at Dallas
sxs180011@utdallas.edu

Dr. Jessica Ouyang
The University of Texas at Dallas
jessica.ouyang@utdallas.edu

Abstract

Human language encompasses more than just text; it also conveys emotions through tone and gestures. We present a case study of three simple and efficient Transformer-based architectures for predicting sentiment and emotion in multimodal data. The *Late Fusion* model merges unimodal features to create a multimodal feature sequence, the *Round Robin* model iteratively combines bimodal features using cross-modal attention, and the *Hybrid Fusion* model combines trimodal and unimodal features together to form a final feature sequence for predicting sentiment. Our experiments show that our small models are effective and outperform the publicly released versions of much larger, state-of-the-art multimodal sentiment analysis systems.

1 Introduction

Language is composed of three different modalities: text, audio, and video. These three modalities together make it easier for humans to convey emotion and sentiment. Thus, a machine learning model for sentiment analysis needs to learn the features and interactions of all three modalities. For example, a frown in the video can alter the emotion expressed in the text transcript, or audio intensity can help determine if a speaker is getting agitated.

Multimodal learning has recently received a good deal of attention from the natural language processing community [Sun et al., 2016, Chen et al., 2018, Liu et al., 2018, Pham et al., 2019]. The Transformer network [Vaswani et al., 2017], with its self-attention modules, has achieved strong performance in multimodal learning; attention provides a natural way to model the relationship between pairs of modalities.

In this work we investigate three small, lightweight, Transformer-based architectures for multimodal sentiment analysis and emotion recognition. Our first model is an implementation of the *Late Fusion* model commonly used as a baseline system, which assigns individual Transformer blocks to each of the three modalities for feature extraction and then combines these unimodal features to learn cross-modal interactions. The second model is an implementation of the *Round Robin* approach; the model generates bimodal features by using cross-modal attention to combine pairs of modalities, one pair at a time. Our last model is a *Hybrid* of the early and late fusion schemes. This model merges the features extracted using a late fusion pipeline, as well as those from an early fusion pipeline, where the three modalities are concatenated and passed through a single Transformer block for feature extraction;.

We present experiments using these three models on three multimodal datasets: IEMOCAP [Busso et al., 2008], an emotion recognition dataset, and CMU-MOSI [Zadeh et al., 2016] and CMU-MOSEI [Zadeh et al., 2018b], two multimodal sentiment analysis datasets. Our results show that our small models are competitive with state-of-the-art models that use much more complex architectures.

Our main contributions are as follows:

- We present three lightweight architectures for multimodal sentiment analysis that achieve comparable results to much larger, state-of-the-art models.
- We analyze the effect of removing or simplifying components of state-of-the-art multimodal architectures.
- We conduct experiments on small training sets, demonstrating the ability of our lightweight architectures to leverage limited training data and computational resources.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 129–137

2 Related Work

We do not give an exhaustive list of prior work in multimodal sentiment analysis, but focus on recent neural approaches that achieved state-of-the-art performance at their times of publication.

2.1 Recurrent Network Approaches

The Memory Fusion Network (MFN) of Zadeh et al. [2018a] uses a separate LSTM to encode each of the three modalities and then uses attention to model cross-modal interactions for different combinations of modalities. The Recurrent Attended Variation Embedding Network (RAVEN) of Wang et al. [2019] encodes the audio and video features using two recurrent neural networks; these features are combined with the textual input using cross-modal attention in a Gated Modality Mixing Network. The Multi-Attention Recurrent Network (MARN) of Zadeh et al. [2018c] is an LSTM-based architecture that stores representations of each of the three modalities, which are then combined using a multi-attention block. Finally, the Multimodal Cyclic Translation Network (MCTN) of Pham et al. [2019] produces multimodal features by translating one modality into another, learning a joint encoding in that direction, and then back-translating to learn a joint encoding in the other direction.

2.2 Transformer Network Approaches

The Transformer network [Vaswani et al., 2017] has been used widely in neural machine translation [Tubay and Costa-jussà, 2018, Edunov et al., 2018, Xia et al., 2019, Devlin et al., 2019] and has proven effective for sentiment analysis and emotion recognition. However, existing architectures are very dense compared to our three lightweight models.

The Multimodal Transformer (MuLT) of Tsai et al. [2019] modifies the Transformer block to compute cross-modal attention for two modalities at a time. It combines modalities in directed pairs, using a total of six Transformers, whose outputs are then merged into a single multimodal representation. Unlike other works, MuLT is able to handle cases where the three modalities are not aligned at the word level; it learns soft alignments via the cross-modal attention weights for each pair of modalities. The model works well in the unaligned case, and in the aligned case, it gives state of the art performance the *Happy* emotion in IEMOCAP.

The Factorized Multimodal Transformer (FMT) of Zadeh et al. [2019] introduces Factorized Multimodal Self-Attention (FSM) modules, which compute self-attention over unimodal, bimodal, and trimodal inputs in parallel. FMT gives state of the art performance in the word-aligned case on CMU-MOSI and on the *Sad*, *Angry*, and *Neutral* emotions in IEMOCAP. We use FMT, along with the word-aligned version of MuLT, as baselines for comparison in our experiments.

2.3 Canonical Correlation Approach

The Interaction Canonical Correlation Network (ICCN) [Sun et al., 2020] implements Deep Canonical Correlation Analysis (DCCA) [Andrew et al., 2013] to extract bimodal features from the outer product matrix of a pair of modalities. Sun et al. use two pairs, text with audio and text with video; these "text-based audio" and "text-based video" features are concatenated with purely textual features to form a multimodal embedding for sentiment analysis. ICCN gives state-of-the-art performance on CMU-MOSEI and on the *Sad* emotion in IEMOCAP.

3 Models

3.1 Input Alignment

We use T, A, and V, to represent the three modalities: text, audio, and video, respectively. Following the notation in [Tsai et al., 2019] and [Zadeh et al., 2019], we denote the input as

$$X_{T,A,V} = \{x_T, x_A, x_V\}$$

where

$$x_i = [x_{t,i}] \text{ for } i \in [T, A, V] \text{ and } t \in [1, \tau]$$

and τ is the length of the input sentence.

Each of the three modalities has its own low-level features, such as the Mel spectrogram for audio or facial landmarks for video. These features are extracted at different sampling rates — one set of features per word or character for text, per millisecond for audio, and per frame for video — and thus the input sequences for the three modalities are often different. A five-thousand-millisecond audio sequence, for example, may be only a three-word sequence from a textual perspective and a 50-frame sequence from a video perspective.

We align the audio and video to the text using the timestamps provided in the text transcripts. The set of audio or video samples that correspond to a

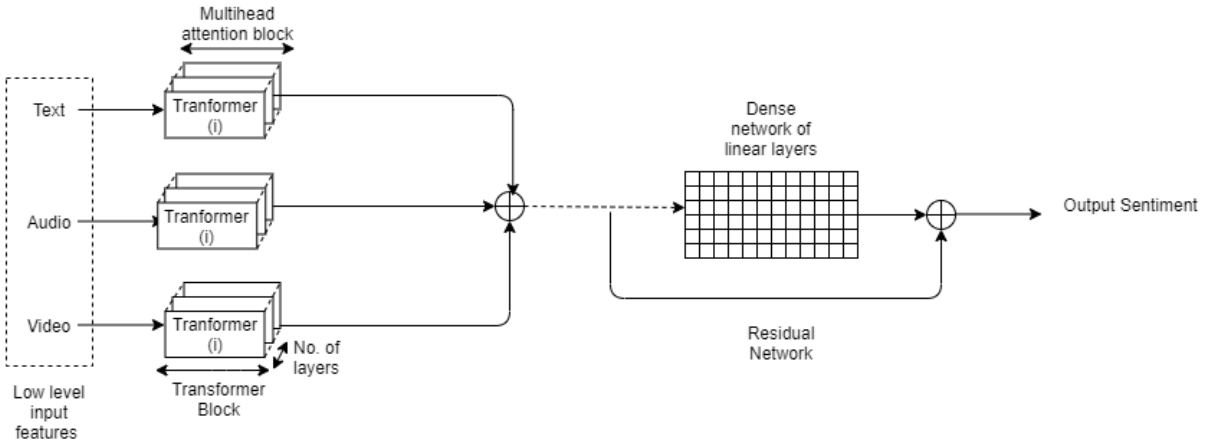

Figure 1: Architecture of our *Late Fusion* model. Unimodal Transformers process each modality separately; the outputs of these Transformers are summed and passed through a residual network of linear layers to produce the final prediction.

word in the transcript are combined using a series of 1D convolutional layers:

$$\bar{X}_{\{T,A,V\}} = conv1D\left(X_{\{T,A,V\}}\right) \in R^d$$

where d is a common feature dimension size. This procedure ensures that the input sequence length is the same across modalities.

3.2 Transformer Blocks

Our three lightweight architectures are comprised of Transformer blocks [Vaswani et al., 2017], which are non-recurrent neural networks that can process sequential data. It consists of alternating attention and linear layers. The attention block of a Transformer uses multi-head attention, where each head computes scaled dot product attention:

$$\text{attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
$$\text{head}_i = \text{attn}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$
$$\text{multi}(Q,K,V) = [\text{head}_1;\ldots;\text{head}_h]W^O$$

where Q, K, V represent the query, key and value; d_k is the key dimension size; W_i^Q, W_i^K, W_i^V are learned projection matrices for head i; and W^O is a learned projection matrix for the attention block.

In addition, Vaswani et al. note that positional encodings must be added to Transformer input because there is no sequential information present in the Transformer itself:

$$PE_{(\text{pos},2i)} = sin(\text{pos}/10000^{2i/d_{model}})$$
$$PE_{(\text{pos},2i+1)} = cos(\text{pos}/10000^{2i/d_{model}})$$
$$\hat{X} = \bar{X} + PE$$

3.3 Three Multimodal Architectures

Figure 1 shows our *Late Fusion* architecture. Three unimodal Transformers learn high-level features from the low-level input features of each modality. The outputs of these unimodal Transformers are then merged together using a simple summation, rather than the merge layer used in previous work [Tsai et al., 2019], and passed to a residual network of linear layers [Xie et al., 2017] for sentiment prediction.

Figure 2 shows our *Round Robin* architecture, which is a simplification of MuTL [Tsai et al., 2019]. Three cross-modal Transformers learn bimodal feaatures for ordered pairs of modalities, where the query is one modality and the key/value is the other. We use only three pairs — text query and audio key/value, audio query and video key/value, and video query and text key/value — with bimodal information flowing in only one direction; in contrast, MuLT uses six pairs of cross-modal Transformers, with information flowing in both directions. MuLT also uses three Transformers, one for each modality, to merge the two pairs sharing that modality as key/value; our pairwise features are simply concatenated and passed to the output residual network.

Figure 3 shows our *Hybrid Fusion* architecture, which uses both an early fusion approach that concatenates the inputs and passes them to a single Transformer to learn trimodal features, as well as a late fusion approach that passes each modality through a separate Transformer to learn unimodal features. The trimodal and unimodal features are concatenated together and merged using a layer of Gated Recurrent Units [Liang et al., 2018].

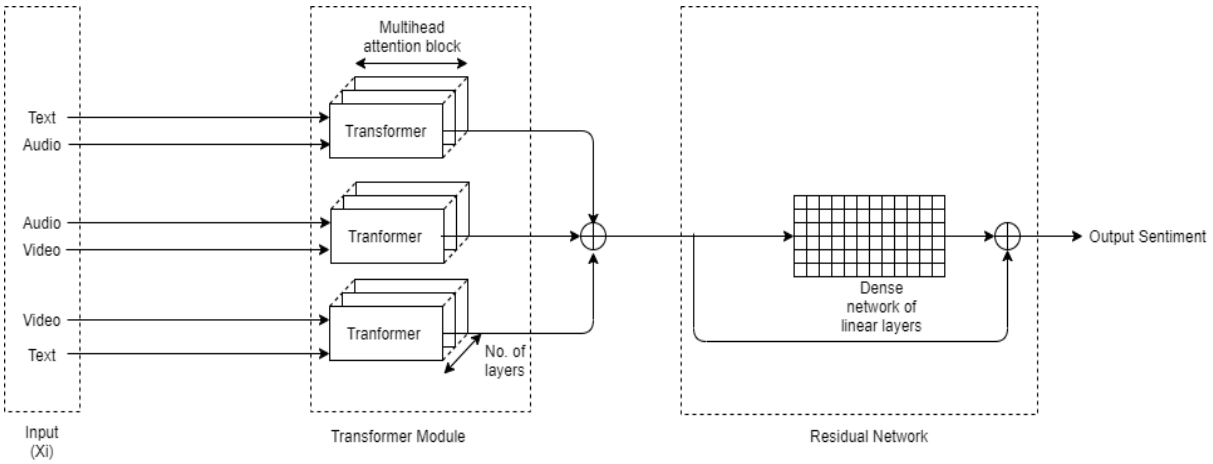

Figure 2: Architecture of our *Round Robin* model. Modalities are combined in a round-robin fashion via thrree cross-modal Transformers, one for each ordered pair of modalities: $[T, A]$, $[A, V]$, $[V, T]$. The outputs of these cross-modal Transformers are concatenated and passed through a residual network of linear layers to produce the final prediction.

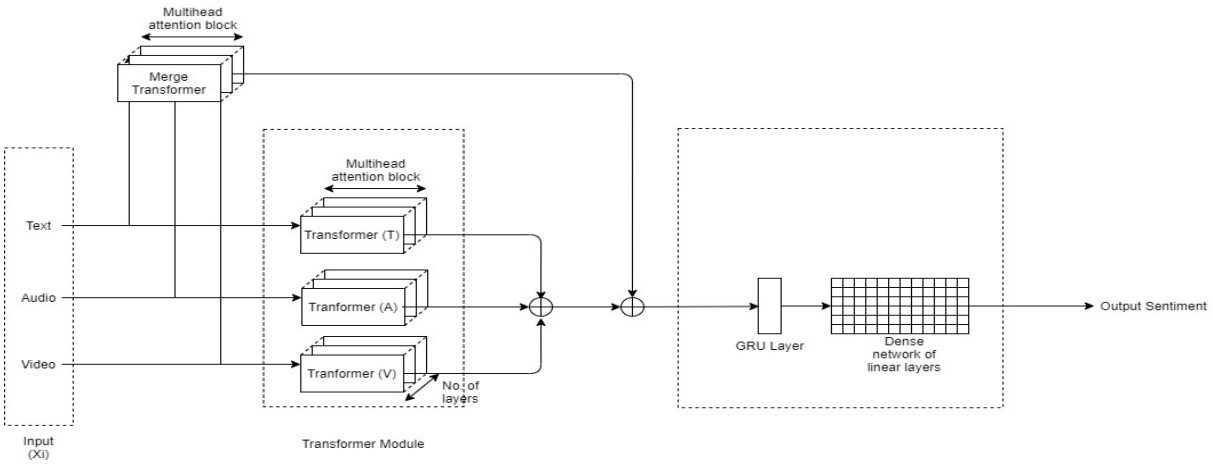

Figure 3: Architecture of our *Hybrid Fusion* model. All three modalities are passed through an early fusion Transformer to produce trimodal features; in parallel, they are individually passed to separate Transformers to produce unimodal features. All features are then concatenated and passed through a GRU and a residual network of linear layers to produce the final prediction.

4 Experiments

We train our models on a single NVIDIA K80 GPU. We tune hyperparameter values for our model using the validation sets provided by our evaluation datasets; we achieve the best validation performance using 8 attention blocks per Transformer, each with 5 attention heads, and a hidden size was set to 40. The dropout rate was set to 0.15; the best learning rate for IEMOCAP was 0.02, while for CMU-MOSI and CMU-MOSEI it was 0.01, with batch sizes of 32, 128, and 40, respectively.

4.1 Datasets

IEMOCAP [Busso et al., 2008] consists of video recordings of 151 conversation sessions (dialogues), totaling around 6k verbal interactions. This dataset is intended for multilabel emotion classification; we evaluate on the four labeled emotions (*Happy*, *Sad*, *Angry*, and *Neutral*) used in previous work [Wang et al., 2019]; also following previous work, we report binary accuracy and F1 score as the evaluation metrics on this dataset.

CMU-MOSI [Zadeh et al., 2016] is a sentiment analysis dataset of 2199 short monologues labeled in the range $[-3, 3]$, with -3 being strongly negative and $+3$ being strongly positive. Following previous work, we report seven-class and binary accuracy, F1 score, mean absolute error, and correlation with human judgments.

CMU-MOSEI [Zadeh et al., 2018b] is a sentiment and emotion analysis dataset of 23K movie reviews from YouTube. As with CMU-MOSI, it is labeled in the range of $[-3, 3]$, and its evaluation metrics are the same as in CMU-MOSI.

4.2 Features and Alignment

Text Features: For word-level textual features we use the pretrained, 300-dimensional, Common Crawl GloVe embeddings [Pennington et al., 2014].

Audio Features: We use Neurospeech [Orozco-Arroyave et al., 2018] to extract 74 low-level audio

Model	Happy		Sad		Angry		Neutral	
Metric	BA	F1	BA	F1	BA	F1	BA	F1
MARN [Zadeh et al., 2018b]	86.7	83.6	82.0	81.2	84.6	84.2	66.8	65.9
MFN [Zadeh et al., 2018a]	86.5	84.0	83.5	82.1	85.0	83.7	69.6	69.2
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
ICCN [Sun et al., 2020]	87.4	84.7	**88.6**	**88.0**	86.3	85.9	69.7	68.5
MuLT [Tsai et al., 2019]	**90.7**	**88.6**	86.7	86.0	87.4	87.0	**72.4**	70.7
MuLT*	84.7	83.5	84.5	84.1	84.9	84.7	70.4	70.7
FMT [Zadeh et al., 2019]	**88.8**	**87.2**	**88.0**	**87.7**	**89.7**	**89.5**	**74.0**	**73.8**
FMT*	85.6	85.1	84.1	83.8	87.9	88.2	70.6	70.4
Late Fusion	87.7	86.8	87.3	86.8	87.9	87.0	72.0	71.5
Round Robin	87.5	84.9	85.2	84.4	87.4	87.5	70.0	69.4
Hybrid Fusion	88.0	86.0	86.9	86.2	89.0	89.0	71.0	71.5

Table 1: Emotion recognition results on IEMOCAP. The metrics are binary (one vs all) accuracy and the F1 score for each of the four emotions. * indicates results from open source code. **Bold** indicates scores higher than that of our model.

Model	ACC_7	ACC_2	F1	MAE	Corr
MARN	-	77.1	77.0	0.97	0.63
MFN	-	77.4	77.3	0.97	0.63
RAVEN	33.2	78.0	76.6	0.92	0.69
MCTN	35.6	79.3	79.1	0.91	0.68
ICCN	39.0	83.1	**83.0**	**0.86**	**0.71**
MuLT	40.0	83.0	**82.8**	**0.87**	**0.70**
MuLT*	30.7	77.5	76.9	1.04	0.66
FMT	-	83.5	**83.5**	**0.84**	**0.74**
FMT*	-	78.3	77.8	**0.91**	**0.70**
Late Fusion	40.2	**83.6**	80.0	0.92	0.69
Round Robin	39.3	78.1	76.7	0.96	0.68
Hybrid Fusion	**40.6**	82.1	79.9	0.94	0.69

Table 2: Sentiment analysis results on CMU-MOSI. ACC_7 was not reported by some baselines. The metrics are seven-way and binary accuracy, F1 score, mean absolute error, and correlation with human judgments. All metrics are better when higher, except for mean absolute error.

Model	ACC_7	ACC_2	F1	MAE	Corr
MFN	45.0	76.9	77.0	0.71	0.54
RAVEN	50.0	79.1	79.5	0.61	0.66
MCTN	49.6	79.8	80.6	0.61	0.67
ICCN	51.6	**84.2**	**84.2**	**0.57**	**0.71**
MuLT	51.8	**82.5**	**82.3**	**0.58**	**0.70**
MuLT*	48.9	80.7	**80.9**	0.63	0.65
Late Fusion	**52.3**	80.7	80.7	0.61	0.69
Round Robin	51.4	80.6	79.9	0.62	0.66
Hybrid Fusion	51.9	80.6	80.5	0.61	0.68

Table 3: Sentiment analysis results on CMU-MOSEI. The metrics used are the same as in Table 2.

features, including Mel-frequency cepstral coefficients and transformations thereof, as well as harmonic, percussive, and glottal source parameters. We also use COVERAP [Degottex et al., 2014] to extract pitch tracking and voiced/unvoiced sloping parameters, peak slope parameters, and maximum dispersion quotients.

Video Features: We extract 35 facial units using Facet [iMotions, 2017], as well as 35 facial action units and 30 facial landmark and gaze features using OpenFace [Baltrusaitis et al., 2018].

4.3 Baseline Models

We compare our results with the state-of-the-art Multimodal Transformer (MuLT)[1] [Tsai et al., 2019] and Factorized Multimodal Transformer (FMT) [Zadeh et al., 2019], as well as Memory Fusion Network (MFN) [Zadeh et al., 2018a], Recurrent Attended Variation Embedding Network (RAVEN) [Wang et al., 2019], Multi-Attention Recurrent Network (MARN) [Zadeh et al., 2018c], and Multimodal Cyclic Translation Network (MCTN) [Pham et al., 2019]. These systems are described in Section 2; all attained state of the art on at least one of the evaluation datasets at their times of publication, and all use a similar feature set to our work.

5 Results and Discussion

We present the results of our model compared to the reported results of our baseline models in Tables 1, 2, and 3. The best-performing MuLT and FMT models are extremely dense, with around 15 and 77 million parameters, respectively. In contrast, our models have between 7-9 million trainable parameters, depending on the architecture; despite using about half as many parameters as MuLT, we see that our models produce comparable results.

We perform fairly well on IEMOCAP, which has around 2717 training samples; we achieve scores around 1-2% below the best-performing model, FMT. On the tiny CMU-MOSI dataset, which has just 1284 training samples, our *Hybrid Fusion* and

[1] We use the aligned version of MuLT for fair comparison with models that obligatorily use word alignments.

	IEMOCAP		CMU-MOSI		CMU-MOSEI	
Model	Time (min)	Mem use (%)	Time (min)	Mem. use (%)	Time (min)	Mem. use (%)
MuLT*	8.1	24.2	7.4	20.0	58.6	59.1
FMT*	55.2	28.0	151.1	26.5	-	-
Late Fusion	1.8	22.1	1.3	22.0	14.9	48.0
Round Robin	1.7	22.7	2.7	19.0	15.8	45.0
Hybrid Fusion	2.5	23.0	1.4	20.0	21.3	54.0

Table 4: Comparison of training time and memory use among MuLT*, FMT*, and our models.

Model	Happy		Sad		Angry		Neutral	
Metric	BA	F1	BA	F1	BA	F1	BA	F1
MuLT*	82.6	81.5	79.4	**80.7**	78.3	78.9	60.1	60.7
FMT*	82.1	81.2	80.2	**80.9**	80.0	**81.7**	60.5	60.2
Late Fusion	84.1	**82.4**	80.3	76.5	**81.0**	79.4	61.6	61.2
Round Robin	85.2	81.2	79.9	77.2	79.0	76.6	63.2	58.1
Hybrid Fusion	**85.5**	80.7	**80.8**	79.9	**81.0**	80.8	**64.7**	**63.5**

Table 5: Results on the reduced IEMOCAP dataset of 1284 training samples. The metrics used are the same as in Table 1.

Model	Happy		Sad		Angry		Neutral	
Metric	BA	F1	BA	F1	BA	F1	BA	F1
Unimodal [T]	86.4	84.0	82.7	78.5	81.6	78.3	67.9	65.9
Unimodal [A]	85.9	79.0	82.2	81.5	85.9	85.9	62.8	60.5
Unimodal [V]	85.1	81.0	79.1	70.4	75.6	74.1	58.8	56.3
Bimodal [T,A]	84.5	82.6	84.8	84.1	85.8	86.1	68.9	67.2
Bimodal [T,V]	85.3	85.1	80.1	80.7	84.2	83.5	66.4	65.4
Bimodal [V,A]	86.8	82.9	81.4	77.9	86.4	86.1	62.5	62.6
Late Fusion [T,A,V]	87.7	86.8	87.3	86.8	87.9	87.0	72.0	71.5

Table 6: Ablation results on IEMOCAP for our *Late Fusion* model.

Model	Happy		Sad		Angry		Neutral	
Metric	BA	F1	BA	F1	BA	F1	BA	F1
Bimodal [T,A]	85.2	82.9	82.9	83.9	86.2	86.4	70.2	69.5
Bimodal [T,V]	86.4	83.9	79.3	77.4	81.4	81.4	65.1	65.0
Bimodal [V,A]	86.4	82.5	79.6	78.6	85.6	85.2	63.1	62.7
Round Robin [T,A,V]	87.5	84.9	85.2	87.4	87.5	86.8	70.0	69.4

Table 7: Ablation results on IEMOCAP for our *Round Robin* model.

Late Fusion models give state of the art results on seven-way and binary accuracy, respectively.

The CMU-MOSEI dataset is much larger than IEMOCAP and CMU-MOSI, with close to 16265 training samples. Our models perform the weakest on this dataset, falling short of the state of the art models by around 2-3%, suggesting that our models may be too small to learn the entire distribution. Neither MARN [Zadeh et al., 2018c] nor FMT [Zadeh et al., 2019] reports results on CMU-MOSEI, so they are omitted from Table 3.

We also experiment with the open source code available for MuLT and FMT (denoted by *). Using the hyperparameter settings provided[2], we were nevertheless unable to match those systems' reported performance, possibly due to differences resulting from random initialization. In training MuLT* and FMT*, we observe that the models are overfitting, with a mean difference of 15-20% between the train and test accuracy; in contrast, the largest train-test accuracy difference among our three models is only about 10%. The smaller number of parameters in our model reduces the risk of overfitting on smaller datasets, while still achieving good performance on larger datasets.

5.1 Analysis of Lightweight Architectures

We compare the training time and memory footprint of our models with MuLT* and FMT* in Table 4[3]. All models are trained on a single NVIDIA K80 GPU with 24GB of memory. We train for 30 epochs on IEMOCAP, 100 on CMU-MOSI and 40

[2]Batch size for FMT* is not given; we use 20, the default.

[3]FMT* does not provide hyperparameter settings for CMU-MOSEI, so those results are omitted.

Model	Happy		Sad		Angry		Neutral	
Metric	BA	F1	BA	F1	BA	F1	BA	F1
Bimodal early [T,A]	84.8	83.1	82.8	81.3	85.1	86.2	68.6	68.7
Bimodal early [T,V]	86.2	83.6	80.3	80.7	85.6	84.8	67.8	67.7
Bimodal early [V,A]	83.9	86.2	84.1	84.2	84.8	85.1	70.0	68.5
Hybrid Fusion [T,A,V]	88.0	86.0	86.9	86.2	89.0	89.0	71.0	71.5
Bimodal late [T,A]	87.0	85.1	85.0	84.9	86.7	86.9	70.3	68.8
Bimodal late [T,V]	86.2	83.8	83.7	83.5	85.6	85.8	67.8	66.9
Bimodal late [V,A]	85.7	83.2	81.1	82.0	86.8	86.9	69.9	67.6

Table 8: Ablation results on IEMOCAP for our *Hybrid Fusion* model: bimodal early fusion with trimodal late fusion (top) and trimodal early fusion with bimodal late fusion (bottom).

Model	Happy		Sad		Angry		Neutral	
Metric	BA	F1	BA	F1	BA	F1	BA	F1
MuLT	**90.7**	**88.6**	**86.7**	**86.0**	**87.4**	87.0	**72.4**	**70.7**
MuLT*	84.7	83.5	84.5	84.1	84.85	84.7	70.4	**70.7**
Round Robin [$T \rightarrow A \rightarrow V$]	87.5	84.9	85.2	84.4	**87.4**	**87.5**	70.0	69.4
Round Robin [$V \rightarrow A \rightarrow T$]	83.0	81.8	82.2	83.7	85.9	82.7	68.2	68.3

Table 9: Results on IEMOCAP for our *Round Robin* model, comparing information flow in each direction, alongside MuLT and MuLT*, which capture information flowing in both directions.

on CMU-MOSEI (the number of epochs needed for MuLT to converge, as reported by Tsai et al. [2019]).

On the smallest dataset, CMU-MOSI, training MuLT* took just over seven minutes, while FMT* took 2.5 hours. Our models train in under three minutes and outperform both MuLT* and FMT*, and this difference in training speed holds for CMU-MOSI and CMU-MOSEI as well. Thus our model, available in the supplementary materials[4], is the fastest and best-performing multimodal sentiment system currently available for public use.

We also conduct experiments on a substantially reduced IEMOCAP training subset of 1284 samples, matching the size of CMU-MOSI, which we create by randomly sampling from the full IEMOCAP training set. Table 5 shows the results of our models, as well as MuLT* and FMT*, retrained on this smaller IEMOCAP training set, and evaluated on the full IEMOCAP test set. We see that our models, with their smaller numbers of parameters, are better able to learn from limited training data than are state-of-the-art models with double or more the number of trainable parameters.

5.2 Analysis of Architecture Components

We perform ablation experiments on our models using the IEMOCAP dataset; ablation results for CMU-MOSI and CMU-MOSEI are omitted due to space constraints, but exhibit similar trends. Table 6 presents the results of modality ablation on the simplest *Late Fusion* model; it clearly shows that unimodal and bimodal models are unable to match the performance of a full multimodal model. This demonstrates the importance of considering all modalities when analyzing spoken language, since some of the emotions or sentiment may be dependent more on the audio or the visual actions of the speaker, rather than the text.

Examining the unimodal results, we see that the Text modality is the most informative for predicting *Happy, Sad*, and *Neutral*, while Audio is the most informative for *Angry*. However, the bimodal results do not always match the unimodal results. The best-performing bimodal model for *Happy* is [V,A], despite Video being the worst-performing single modality, and [T,A] is the worst-performing bimodal model, despite both Text and Audio outperforming Video individually. Considering the other three emotions, we see that the best bimodal model varies between [T,A] and [V,A], with [T,V] generally performing the worst.

Table 7 shows the results of modality ablation on the *Round Robin* model; as the architecture does not support unimodal experiments, only bimodal results are shown. Comparing Table 6 to Table 7, we see that the cross-modal Transformers of the full *Round Robin* model are outperformed by the full *Late Fusion* model. However, the relative performance among modality pairs is consistent across Tables 6 and 7.

Finally, Table 8 shows the results of modality ablation on the *Hybrid Fusion* model, where we

[4]We will release it online after the anonymity period.

compare the relative contributions of the early fusion and late fusion halves of the architecture. The top of the table shows the results of reducing the early fusion half to only two modalities while retaining all three modalities in the late fusion half, and the bottom shows the results of reducing the late fusion half to two modalities while retaining all three in the early fusion half; in both sets of experiments, the overall model has access to all three modalities, but only through either the early fusion path or the late fusion path.

Surprisingly, although standalone early fusion models are outperformed by standalone late fusion models [Tsai et al., 2019], we find that a hybrid model containing a full, trimodal early fusion half is more robust to modality ablation in its late fusion half than a model with a full late fusion half is to an ablated early fusion half. Our results in this experiment also show greater variability among modality pairs. The [T,A] combination, which gave the best performance in the *Late Fusion* and *Round Robin* experiments, remains the strongest modality pair for the full early fusion, bimodal late fusion model. In contrast, for the bimodal early fusion, full late fusion model, [T,A] is outperformed by one of the two Video-based modality pairs, [T,V] or [V,A], on each of the four emotions, suggesting that the performance gap of early versus late fusion differs across modalities.

5.2.1 Order of Modalities in Round Robin

The effect of direction on our *Round Robin* model is shown in Table 9; this experiment shows the impact of the direction of information flow across modalities within the model. Comparing our results to those of MuLT and MuLT*, we see that capturing information flow in one direction, text to audio to video and back to text, is enough for a model to give good predictions, without requiring the additional overhead of handling both directions. We can also see that the direction does matter; the performance of the Round Robin model with information flowing in the opposite direction, from video to audio to text and back to video, is relatively poor. These results suggest that the interactions between pairs of modalities are directed.

6 Conclusion

We have presented three lightweight architectures for multimodal sentiment analysis and emotion recognition. The *Late Fusion* model merges unimodal features, the *Round Robin* model iteratively combines bimodal features, and the *Hybrid Early-Late Fusion* model combines early-fusion trimodal and late-fusion unimodal features. Our proposed models are much smaller in size compared to existing state-of-the-art models; they are able to attain new state-of-the-art scores on the CMU-MOSI and CMU-MOSEI datasets on two metrics, while remaining competitive on the others. Further, our experiments analyzing the relative contribution of modalities and architecture components in our models suggest new directions for developing multimodal systems. We hope that our simple architectures for sentiment and emotion detection, currently the fastest and best-performing publicly available system, as well as the insights revealed in our experimental results, can be useful for further research in the field.

References

G. Andrew, R. Arora, J. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Michael Schuster, Zhi-Feng Chen, Yonghui Wu, and Macduff Hughes. The best of both worlds: Combining recent advances in neural machine translation. In *ACL*, 2018.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep — a collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, 2018.

iMotions. Facial expression analysis, 2017. URL https://rb.gy/hkrcc4.

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

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018.

Juan R. Orozco-Arroyave, Juan Camilo Vasquez-Correa, Jesus Francisco Vargas-Bonilla, Raman Arora, Najim Dehak, Phani S. Nidadavolu, Heidi Christensen, Frank Rudzicz, Maria Yancheva, Hamid R. Chinaei, Alyssa Vann, Nikolai Vogler, Tobias Bocklet, Milos Cernak, Julius Hannink, and Elmar Noth. Neurospeech: An open-source software for parkinson's speech analysis. *Digital Signal Process.*, 77:207–221, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

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

Felix Sun, David F. Harwath, and James R. Glass. Look, listen, and decode: Multimodal speech recognition with images. *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 573–578, 2016.

Zhongkai Sun, P. Sarma, W. Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *ArXiv*, abs/1911.05544, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558–6569, 2019. version 2.

Brian Tubay and Marta R. Costa-jussà. Neural machine translation with the transformer and multi-source romance languages for the biomedical wmt 2018 task. In *WMT*, 2018.

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

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 33 1:7216–7223, 2019.

Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. Tied transformers: Neural machine translation with shared encoder and decoder. In *AAAI*, 2019.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

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

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018a. version 1.

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

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2018:5642–5649, 2018c. version 2.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. *ArXiv*, abs/1911.09826, 2019.