What makes the difference? An Empirical Comparison of Fusion Strategies for Multimodal Language Analysis

Dimitris Gkoumas[a,*], Qiuchi Li[b], Christina Lioma[c], Dawei Son[a]

[a] *The Open University, United Kingdom*
[b] *University of Padua, Italy*
[c] *University of Copenhagen, Denmark*

Abstract

Understanding human language is an emerging interdisciplinary field bringing together artificial intelligence, natural language processing, and cognitive science. It goes beyond linguistic modality, by effectively combining non-verbal behaviour (i.e., visual, acoustic) which is crucial for inferring speaker intent. Being a rapidly growing area of research, a range of models of multimodal language analysis has been introduced within the last two years. In this paper, we present a large-scale empirical comparison of eleven state-of-the-art (SOTA) modality fusion approaches to find out which aspects could be effectively used to solve the problem of multimodal language analysis. An important feature of our study is the critical and experimental analysis of the SOTA approaches. In particular, we replicate diverse complex neural networks, utilizing attention, memory, and recurrent components. We propose a methodology to investigate both their effectiveness and efficiency in two mul-

*Corresponding author
Email addresses: dimitris.gkoumas@open.ac.uk (Dimitris Gkoumas), qiuchili@dei.unipd.it (Qiuchi Li), c.lioma@di.ku.dk (Christina Lioma), dawei.song@open.ac.uk (Dawei Son)

timodal tasks: a) video sentiment analysis and b) emotion recognition. We evaluate all approaches on three SOTA benchmark corpora, namely, a) Multimodal Opinion-level Sentiment Intensity (MOSI), b) Multimodal Opinion Sentiment and Emotion Intensity (MOSEI), which is the largest available dataset for video sentiment analysis, and c) Interactive Emotional Dyadic Motion Capture (IEMOCAP). Comprehensive experiments show that the attention mechanism components are the most effective for modelling interactions across different modalities. Besides, utilization of linguistic modality as a pivot modality for nonverbal modalities, incorporation of long-range cross-modal interactions across multimodal sequences, and integration of modality context, are also among the most effective aspects for human multimodal affection recognition tasks.
Keywords: multimodal human language understanding, video sentiment analysis, emotion recognition, reproducibility in multimodal machine learning

---

1. Introduction

Human language is inherently multimodal and is manifested via words (i.e., linguistic modality), gestures (i.e., visual modality), and vocal intonations (i.e., acoustic modality). Consequently, we need to process both verbal (e.g., linguistic utterances) and nonverbal signals (e.g., visual, acoustic utterances) to better understand human language. Verbal signals often vary dynamically in different nonverbal contexts. Even though for humans, comprehending human language is an easy task, this is a non-trivial challenge for machines. Giving machines the capability to effectively understand human

language opens new horizons for human-machine conversation systems [1], tutoring systems [2], and health care [3], to name a few applications.

The challenge of modelling human language lies in coordinating time-variant modalities. At its core, this research area focuses on modelling intramodal and crossmodal dynamics [4]. Intramodal dynamics refer to interactions within a specific modality, independent of other modalities. An example is word interactions in a sentence. Crossmodal dynamics refer to interactions across several modalities, for example, a simultaneous presence of a negative word, with a frown, and a soft voice. Such interactions, occurring at the same time step, are called synchronous crossmodal interactions. Crossmodal interactions might span over a long-range multimodal sequence and are called asynchronous crossmodal interactions. For example, the negative word, with the soft voice at the time step t might interact with the frown at the time step $t + 1$.

Early approaches for learning multimodal representations have widely utilized conventional natural language processing (NLP) techniques in multimodal settings [5, 6, 7, 8]. A recent trend in multimodal embedding learning research is to build more complex models utilizing attention, memory, and recurrent components [9, 10, 11, 12, 13, 14, 15]. Various review papers have surveyed the advancements in multimodal machine learning [16, 17, 18, 19, 20]. In particular, they mostly provide an insightful organization of modality fusion strategies. They also identify broader challenges faced by multimodal representation learning, such as synchronization across different modalities, confidence level, contextual information etc. Though, none of them has conducted a comprehensive empirical study across different state-of-the-art

(STOA) fusion approaches to multimodal language analysis, with aim at providing a critical and experimental analysis. Such an extensive empirical evaluation would be useful to find out which aspects in the STOA approaches are the most effective to solve the problem of multimodal language analysis. This paper aims to fill the gap. In particular, we replicate and evaluate the most recent SOTA fusion approaches for modelling human language on three widely used benchmark corpora for multimodal sentiment and emotion analysis [21, 22, 23] and investigate the following Research Questions (RQ).

- **RQ1** How effective are the current machine learning based multimodal fusion strategies for the sentiment analysis and emotion recognition tasks?
- **RQ2** How efficient are the SOTA multimodal fusion strategies, and how could the effectiveness affects efficiency, in the context of the multimodal sentiment and emotion analysis tasks?
- **RQ3** Which components/aspects in the multimodal language models and fusion strategies are the most effective?

The rest of the paper is organized as follows: Section 2 briefly reviews the related work. Section 3 describes the experiments in detail. The experimental results are shown and discussed in Section 4 and 5 respectively. Finally, Section 5 concludes the paper.

2. Related Work

We provide a review of multimodal representation learning and multimodal time series for video sentiment analysis and emotion recognition.

2.1. Multimodal representation learning

Multimodal representation learning is a research area of great interest due to the huge multimedia (e.g., textual, visual, and acoustic) data available in different contexts. A recent trend in NLP research has been geared towards a variety of multimodal applications, including visual recognition [24], multimodal sentiment analysis [25], visual-acoustic emotion recognition [26], visual question answering [27], and medical image analysis [28].

An early overview of multimodal information retrieval (MMIR) presents briefly the basic concepts of MMIR with emphasis on challenges in MMIR systems, feature extraction, and fusion strategies [29]. A more comprehensive review of various multimodal tasks is given by [16]. In [17], Sun reviews multiple kernel and subspace algorithms for multi-view learning. Recent advances in multimodal machine learning have been reviewed covering various directions of the field, such as representation, translation, alignment, fusion, and co-learning [19, 18].

More recently, research in the affective computing field has attracted the attention of many researchers due to the recent availability of relatively large-scale datasets for video sentiment analysis and emotion recognition tasks [21, 22, 23]. A comprehensive literature review of multimodal affective analysis frameworks is given by Poria et al. [20]. Furthermore, Fatemeh et al. [30] survey emotion body gesture recognition approaches. However, none of the above surveys provides a comprehensive empirical study of the very recent multimodal language fusion strategies for sentiment analysis.

2.2. Multimodal Sentiment Analysis

Learning multimodal language embeddings is based on modelling intramodal and crossmodal dynamics. Early, late, and hybrid fusion strategies have been utilized to model such dynamics. Early fusion approaches integrate features after being extracted [31]. Late fusion approaches build up diverse classifiers for each modality and then aggregate their decisions by voting [32], averaging [33], weighted sum [34] or a trainable model [35, 36]. The hybrid approach combines outputs from early fusion and individual unimodal predictors. Early work has pushed some progress towards multimodal language embedding learning [37, 38]. A range of neural approaches, such as RNNs [39], LSTMs [40], and CNNs [41], have been used for learning language-based multimodal embeddings by fusing either input features per timestamp or unimodal output hidden units [5, 6, 7, 8].

Recent advances in deep learning have led to more sophisticated approaches for modelling temporal intramodal and crossmodal interactions across unimodal sequences. Early advancements of this field utilized tensor-based fusion approaches for entangling [42] and disentangling [43, 44] multimodal representations. Those approaches fuse unimodal features at the utterance level [42, 43, 44], word-level [45], or in a hierarchical manner [46]. Considering human language contains time-series and thus requiring fusing time-varying signals, a recent trend is to exploit LSTMs and RNNs to fuse unimodal representations at the feature level [12, 47]. Amongst those approaches, some of them use hybrid memory components, constructed from the hidden units of each modality at the previous timestamp and fed as an additional input of the next timestamp [12, 13]. Inspired by successful trends

in NLP, some approaches introduced encoder-decoder structures in sequence-to-sequence learning by translating a target modality to a source modality [14, 15, 48], reinforcement learning [49], fuzzy logic[50], and simple but strong baselines [51]. Recently, attention mechanisms have been exploited for aligning different modalities, resulting in better-performed modality fusion approaches [9, 10, 11].

In this work, we align nonverbal features with words before training. That is, we model crossmodal interactions on aligned timestamps (i.e., synchronous crossmodal interactions) without considering long-range contingencies across different modalities (i.e., asynchronous crossmodal interactions). Though, recently a few approaches have been proposed to model long-range crossmodal interactions across multimodal sequences [9, 11, 47]. However, working on unaligned features is a non-trivial task. A fair comparison between word-aligned sequences and unaligned multimodal time series shows a decreased performance for unaligned multimodal streams [9].

Finally, it is worth noting that there exist other approaches considering contextual information from surrounding utterances, thus aiding the sentiment analysis and emotion recognition tasks. Current work utilizes supervised NLP approaches for modeling contextual interactions among utterances, including recurrent neural networks [6, 52], memory networks [53, 54], sequence-to-sequence networks [55], graph neural networks [56], and quantum-inspired networks [57]. Nevertheless, these approaches are beyond the scope of this paper since they consider modality fusion as a simple concatenation of unimodal features.

3. Methodology

This section details the methodology we take for our empirical study of the most recent STOA multimodal language fusion approaches, in the context of video sentiment and emotion analysis tasks. We first formulate the task on which our study is carried out. Sentiment analysis is a binary multimodal classification task inferring either positive or negation emotions. Emotion recognition is a multimodal multilabel classification task inferring one or more emotions, i.e., happy and joyful. Though, both tasks target to capture emotions of video utterance and fall under affective computing field [58].

3.1. Task definition

The goal is to infer the emotion of utterances from video speakers. Each video consists of N sequential utterances $U = (U_1, ..., U_i, ..., U_N)$, where i is the i^{th} utterance. Each utterance U_i is associated with three modalities, linguistic, visual, and acoustic, $U_i = (U_i^l, U_i^v, U_i^a)$, $1 \leq i \leq N$. The corresponding labels for the N segments are denoted as $y = (y_1, ..., y_i, ..., y_N)$, $y_i \in \mathbb{R}$. We apply word-level alignment, where visual and acoustic features are averaged across the time interval of each spoken word. Then, we zero-pad the utterances to obtain time-series data of the same length. After this step, language, visual, and acoustic features have the same length L. For the linguistic modality the U_i utterance is represented by $U_i^l = (l_i^1, ..., l_i^L)$. Similarly for visual and acoustic modalities, it is represented by $U_i^v = (v_i^1, ..., v_i^L)$ and $U_i^a = (a_i^1, ..., a_i^L)$, respectively.

3.2. Datasets

We empirically evaluate the SOTA approaches from the last two years on multimodal sentiment analysis task by using two SOTA benchmark data-sets, namely CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) [21] and the largest available dataset for multimodal sentiment analysis, CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [22]. We also evaluate the approaches on the multimodal emotion recognition task using the IEMOCAP dataset [23]. We compare all approaches on word-aligned multimodal language sequences, leaving the very challenging comparison with unaligned language sequences for future work.

CMU-MOSI is a relatively balanced (1176 positive and 1023 negative utterances) human multimodal sentiment analysis dataset consisting of 2,199 short monologue video clips (each lasting the duration of a sentence). It has 1,284, 229, and 686 utterances in training, validation, and test sets. CMU-MOSEI is a larger scale sentiment and emotion analysis dataset made up of 22,777 movie review video clips from more than 1,000 online Youtube speakers. The training, validation, and test sets comprise of 16,265, 1,869 and 4,643 utterances, respectively. Each sample is labelled by human annotators with a ratio score from -3 (highly negative) to 3 (highly positive) including zero. Hence, the multimodal sentiment analysis task can be formulated as a regression problem.

For MOSI and MOSEI, we use the CMU-Multi-modal Data SDK[1] [22] for feature extraction. Following previous work [9, 13, 47, 42, 43, 59], we convert

[1]https://github.com/A2Zadeh/CMU-MultimodalSDK

video transcripts into 300-dimensional pre-trained Glove word embeddings (glove.840B.300d) [60]. Besides, GloVe embedding is more computationally affordable than other more effective, yet computationally expensive, word embeddings [61, 62]. Facet [2] is used to capture facial muscle movement including per-frame basic and advanced emotions and facial action units. We use VOCAREP [63] to extract low-level acoustic features (e.g., 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients). For MOSI, we extract visual and acoustic features at a frequency of $15Hz$ and 12.5 Hz respectively. For MOSEI, we extract at a frequency of 15 Hz and $20Hz$. To reach the same time alignment across modalities, we apply a word-level alignment. To align visual and acoustic modalities with words we use P2FA [64]. Then, to obtain the aligned timesteps, we perform averaging on the visual and audio features within these time ranges. All sequences in the word-aligned case have length 50. For each word the dimension of the feature vector is set to 300 (linguistic), 20 (visual), and 5 (acoustic) for MOSI, and 300 (linguistic), 35 (visual), and 74 (acoustic) for MOSEI.

For multimodal emotion recognition, we use IEMOCAP. It consists of 151 videos about dyadic interactions, where professional actors are required to perform scripted scenes that elicit specific emotions. It has 2,717, 798, 938 utterances in training, validation, and test sets. Each sample is labelled by human annotators for 4 emotions (neutral, happy, sad, angry). The labels

[2] https://pair-code.github.io/facets/

for every emotion are binary. This allows us to reduce the multiclass learning problem to a problem solvable using binary classifiers. Following a one-vs-all strategy, for each emotion, we train a robust classifier to recognize one emotion from all the others. We follow a similar process to CMU-MOSEI for extracting features from 3 streams. The linguistic, facial and acoustic embeddings are 300-dimensional, 35-dimensional, and 74-dimensional vectors, respectively. All sequences are word-aligned having length 50.

3.3. Evaluation Metrics

For evaluating effectiveness on MOSI and MOSEI, we adopt a series of evaluation performance metrics used in prior work [22, 9, 13, 12]: binary accuracy (i.e., Acc_2 : positive sentiment if $values \geq 0$, and negative sentiment if $values < 0$), 7-class accuracy (i.e., Acc_7 : sentiment score classification in $Z \cap [-3, 3]$), $F1$ score, Mean Absolute Error (MAE) of the score, and the Pearson's correlation ($Corr$) between the model predictions and regression ground truth. For all the metrics, higher values denote better performance, except MAE where lower values denote better performance.

For evaluating effectiveness on IEMOCAP, in contrast to previous work reporting accuracy [9, 47], we report recall and F_1 score for individual emotion classes. We empirically found that accuracy was a misleading measurement for evaluating one-vs-all emotion classifiers. This is because there is a class imbalance. For instance, the ratio of utterances labelled as happy versus the other emotion equals 1/6. Indeed, some classifiers showed high accuracy even they failed to detect many emotion class correctly. To evaluate the overall performance of the SOTA models, we also calculate the weighted recall and F_1 score measurements.

We evaluate efficiency by reporting a) the number of parameters for each approach, b) the training time of learning, i.e., speed-up during inference, and c) the validation set convergence.

3.4. Experiments

To address our research questions, we devise three experiments as follows:

1. **Experiment 1:** We first replicate the SOTA approaches following the same experiment set up, as reported in the original papers. Then, we investigate the performance through a comprehensive critical and experimental analysis.
2. **Experiment 2:** We compare the SOTA approaches in terms of efficiency.
3. **Experiment 3:** We conduct several ablation studies to understand a) the importance of modalities and b) which components contribute most for modelling crossmodal interactions across the three modalities.

3.5. SOTA models

We replicate into a unified framework in PyTorch a variety of sequential attention mechanism, memory, tensor fusion, and translation neural approaches[3]. Most of their authors have made implementations available on github. We replicated from scratch the EF-LSTM, LF-LSTM, RMFN, and MARN models.

Except Multimodal Transformer (MulT) [9], the rest of the modality fusion methods are typically RNN-based deep learning networks. Though,

[3]The code for our models and experiments can be found on Github.

we go beyond a typical one by one comparison and propose a taxonomy in terms of model features, namely: recurrent-based, tensor-based, attention mechanism-based, memory-based, and translation-based networks. This taxonomy will enable researchers to better understand the SOTA field and identify directions for future research.

3.5.1. Recurrent cell-based networks

This category includes modality fusion approaches which mainly utilize recurrent cells for each time step. In this case, the cells get stacked one after the other implementing an efficient stacked RNN.

- **Early-Fusion LSTM (EF-LSTM)** [40] EF-LSTM concatenates linguistic, visual, and acoustic features at each timestamp, and builds an LSTM to construct sentence-level multimodal representation. The last hidden state is taken and sequentially passed to two fully-connected layers to produce the output sentiment.

- **Late-Fusion LSTM (LF-LSTM)** [40]. LF-LSTM builds LSTMs for linguistic, visual, and acoustic inputs separately, and concatenates the last hidden state of the three LSTMs as sentence-level multimodal representation. It is taken and sequentially passed to two fully-connected layers to produce the output sentiment.

- **Recurrent Multistage Fusion Network (RMFN)** [10] RMFN models crossmodal interactions through a divide-and-conquer approach in several stages. Intramodal dynamics are modelled through modality-specific RNNs. For each timestep, the unimodal hidden states of RNNs

are concatenated. Then, the concatenated representation is processed in multiple stages. For each stage, the most important modalities are highlighted through an attention module, and then fused with the previous stage fused representations. In the end, a summary action generates a multimodal joint representation which is fed back into the intramodal RNNs as an additional input for the next timestep.

3.5.2. Tensor-based networks

This group of networks is mainly based on the tensor product of modalities for entangling and disentangling information.

- **Tensor Fusion Network (TFN)** [42] TFN explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures unimodal, bimodal, trimodal interactions across linguistic, visual, and acoustic modalities.

- **Low-rank Multimodal Fusion (LMF)** [43]. LMF adopts the same approach as TFN to model the multimodal representation. After that, it applies a tensor decomposition approach by calculating the inner product of multimodal tensor with a weight tensor. The output is a low-dimension vector which used to make predictions.

3.5.3. Attention mechanism-based networks

These approaches mainly exploit various attention mechanism components to fuse modalities.

- **Multi-Attention Recurrent Network (MARN)** [59]. MARM captures crossmodal dynamics at each timestamp. A multi-attention block

is built to construct a crossmodal representation based on hidden states of the previous timestamp and fed into the inputs of the current timestamp. The crossmodal representation and hidden states of the last timestamp are concatenated to form a multimodal sentence embedding, which is sequentially passed to two fully-connected layers to produce the output sentiment.

- **Multimodal Transformer (MulT)** [9] MulT merges multimodal time-series via a feed-forward fusion process from multiple directional pairwise crossmodal transformers. Each crossmodal transformer is a deep stacking of several crossmodal attention blocks. As a final step, it concatenates the outputs from the crossmodal transformers and passes the multimodal representation through a sequence model to make predictions.

- **Multimodal Uni-Utterance - Bimodal Attention (MMUU-BA)** [10] MMUU-BA encodes linguistic, visual, and acoustic streams through three separate Bi-GRU layers followed by fully-connected dense layers. Then, pairwise-attentions are computed across all possible combinations of modalities, i.e, linguistic-visual, linguistic-acoustic, and visual-acoustic. Finally, individual modalities and bimodal attention pairs are concatenated to create the multimodal representation, used for final classification. MMUU-BA makes predictions by applying a fully-connected layer to each timestamp. In our experiments, since we do not considerate proceeding utterances, we extract the last hidden state only and fit it to a fully-connected layer to make predictions.

- **Recurrent Attended Variation Embedding Network (RAVEN)** [47] RAVEN learns multimodal-shifted word representations conditioned on the visual and acoustic modalities. Concretely, visual and acoustic embeddings interact with each word embedding through an attention gated mechanism to yield a nonverbal visual-acoustic vector. The resulted vector is integrated into the original word embedding to model the intensity of the visual-acoustic influence on the original word. By applying the same method for each word in a sentence, the model outputs a multimodal-shifted word-level representation. The representation is encoded into an LSTM followed by a fully connected layer to produce an output that fits the task. Yet, in our experiments, we consider the last hidden state to construct nonverbal visual-acoustic embeddings since we work on word-level aligned data.

3.5.4. Memory-based networks

This category extends recurrent neural model with a memory component to model modality interactions.

- **Memory Fusion Network** (MFN) [13] MFN is a memory fusion network that builds a multimodal gated memory component, and the memory cell is updated along the evolution of the hidden states of three unimodal LSTMs. The final memory cell is concatenated with the last hidden states of unimodal LSTMs as the multimodal sentence representation, and it is sequentially passed to two fully-connected layers to produce the output sentiment.

3.5.5. Translation-based networks

This category includes neural machine translation approaches for modelling human language by converting a source modality to a target modality.

- **Multimodal Cyclic Translations Network (MCTN)** [14] MCTN is a hierarchical neural machine translation network with a source modality and two target modalities. The first level learns a joint representation by using back-translation. Then, the intermediate representation is translated into the second target modality without back-translation. The multimodal representation is fed into RNN for final classification. For our experiments, the source modality is the linguistic one.

We first fine-tune all models by performing a fifty-times grid search over their parameter pool. We report the final settings in Appendix A. After the fine-tuning process, we train again all the models for 50 epochs, five times. We use Adam optimizer with L1 loss as the loss function for CMU-MOSI and CMU-MOSEI since sentiment analysis is formulated as a regression problem. For IEMOCAP, we use cross-entropy loss since emotion recognition is formulated as classification problem. We report the average performance on the test set for all experiments.

4. Results

4.1. Effectiveness

In Table 1 we see that attention mechanism-based approaches, i.e., MulT, MMUU-BA, and RAVEN, attain the highest binary accuracy (being between

78.2% and 78.7%) on MOSI. MulT reports just 0.1 % higher accuracy than RAVEN. Yet, for Acc_7, Raven reports an increased performance of 34.6% as compared to 33.8% of MMUU-BA and 33.6% of MulT. TFN achieves the highest accuracy of 34.9% for Acc_7 . Raven and MMUU-BA report the highest correlation ($Corr$). Despite the low accuracy, MCTN attains the lowest mean absolute error. This might imply that MCTN needs more epochs to converge (we found in [14] that MCTN has been trained for 200 epochs). Overall, RAVEN is the most effective approach on MOSI. T-tests did not reveal a significant difference in binary accuracy across all approaches.

There is a discrepancy between the empirical results from our experiments and the reported ones in literature. Specifically, we empirically found lower accuracy for all the SOTA approaches, except RAVEN which attained an increased accuracy of 78.6% compared to 78% in [47]. A possible reason for the discrepancy between literature and our empirical results may be because different versions of the MOSI dataset have been used in the published works. Those versions consists of different feature dimensions and sequence lengths. Another possible explanation for this might be the fine-tuning parameters, which are rarely reported in current work, making reproducibility a particularly difficult task. Currently, MulT is the SOTA approach in literature reporting an increased binary accuracy of 83.0% compared to 78.7% in our experiments on MOSI. Note that, for MulT we use the same datasets, implementation, and configuration settings as described in [9].

In Table 2 we present the results for multimodal sentiment analysis on MOSEI. All approaches attain an improved performance compared to that one on MOSI dataset. We suspect this is because MOSEI is much larger

Table 1: Comparative analysis across the SOTA approaches on MOSI.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
EF-LSTM [40]	32.7	75.8	75.6	1.000	0.630
LF-LSTM [40]	32.7	76.2	76.2	0.987	0.624
RMFN [10]	32.3	76.8	76.4	0.980	0.626
TFN [42]	**34.9**	75.6	75.5	1.009	0.605
LMF [43]	30.5	75.3	75.2	1.018	0.605
MARN [59]	31.8	76.4	76.2	0.984	0.625
MulT [9]	33.6	**78.7**	78.4	0.964	0.662
MMUU-BA [10]	33.8	**78.2**	78.1	0.947	**0.675**
RAVEN [47]	**34.6**	**78.6**	78.6	0.948	**0.674**
MFN [13]	31.9	76.2	75.8	0.988	0.622
MCTN [14]	32.3	76.2	76.2	**0.903**	0.630

dataset. MMUU-BA attains an increased binary accuracy of 80.7% compared to 80.2% of RAVEN and MulT. MMUU-BA also reports the highest accuracy for Acc_7 and the highest correlation ($Corr$ in Table 2) compared to all other approaches. In general, we found that attention mechanism-based fusion strategies, i.e., MMUU-BA, MulT, and RAVEN, significantly outperform the other approaches. Yet there is no significant difference across MMUU-BA, MulT, and RAVEN in terms of binary performance.

MOSEI is a recently published dataset. We can only compare the empirical results from our experiments to the reported ones in literature for RAVEN, MulT and MMUU-BA. In literature, MulT reports the best binary performance, attaining an increased binary accuracy of 82.5% compared to

Table 2: Comparative analysis across the SOTA approaches on MOSEI.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
EF-LSTM [40]	45.7	78.2	77.1	0.687	0.573
LF-LSTM [40]	47.1	79.2	78.5	0.655	0.614
TFN [42]	47.3	79.3	78.2	0.657	0.618
LMF [43]	47.6	78.2	77.6	0.660	0.623
MARN [59]	47.7	79.3	77.8	0.646	0.629
MulT [9]	46.6	**80.2**	79.8	0.657	0.661
MMUU-BA [10]	**48.4**	**80.7**	80.2	**0.627**	**0.672**
RAVEN [47]	47.8	**80.2**	79.8	0.636	0.654
MFN [13]	47.4	79.9	79.1	0.646	0.626

80.2% in our experiments even though we used the same experimental settings as in [9]. In contrast, MMUU-BA reports an increased binary accuracy of 80.7% compared to 79.8% in literature. In [47], authors do not conduct experiments on MOSEI. Yet, in [9], for RAVEN, authors report a decreased accuracy of 79.1% compared to 80.2% (see Table 2). We could not run experiments for RMFN and for MCTN on MOSEI. RMFN was computationally too expensive and MCTN could not support MOSEI.

Following previous work [65], the binary performance across different modality fusion approaches is compared for the MOSI and MOSEI tasks, as shown in Figure 1. Each line style corresponds to the taxonomy of the SOTA approaches. According to the Figure 1, all approaches improve on the MOSEI task. In addition, MulT and Raven yield similar performance for both MOSI and MOSEI tasks. That is, they show similar learning be-

haviour. Though, MMUU-BA shows a positive trend with a sharper rise in the performance for MOSEI task than MulT and RAVEN approaches.

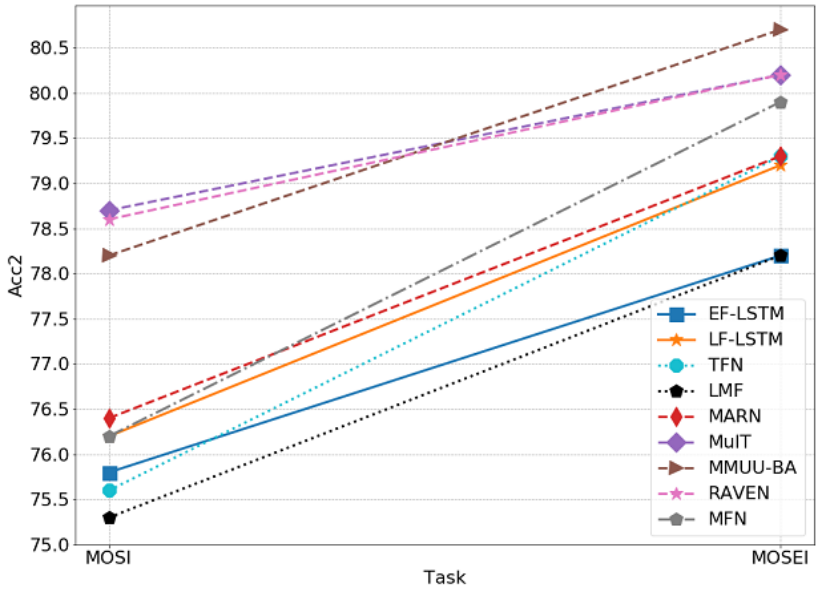

Figure 1: Accuracy comparison across different modality fusion approaches for MOSI and MOSEI tasks.

We present the results for the emotion recognition task in Table 3. In contrast to sentiment analysis tasks, instead of accuracy, we calculate the class-wise recall to find out how many emotions detected correctly from total emotions for each emotion class. We also calculate the weighed recall for each modality fusion method. The results show that happy emotion class is the most challenging for all approaches, whilst the angry class is the most straightforward. Attention mechanism approaches, e.g., MulT and MMUU-BA, are the most effective for the emotion recognition task. In particularly, MMUU-BA achieves the highest recall for happy and sad classes, whilst MulT recalls the most neutral utterances (see Table 3). Though, EF-LSTM has the

highest sensitivity for the angry class. Overall, MulT is the most effective approach for the emotion recognition task yielding an increased weighted recall of 60.2% as compared to 58.7% of the next best approach, i.e., MMUU-BA. We can not directly compare our results with those in literature, since binary accuracy is used as a prime performance measurement. Though, in [9], MulT is also the SOTA for IEMOCAP task.

Table 3: Comparative analysis across the SOTA approaches on IEMOCAP dataset.

	Neutral		Happy		Sad		Angry		Weighted	
Approach	*Recall*	$F1$	*Recall*	$F1$	*Recall*	$F1$	*Recall*	$F1$	*Recall*	$F1$
EF-LSTM [40]	57.3	61.2	20.7	30.8	57.7	**62.0**	**80.7**	71.7	57.8	59.5
LF-LSTM [40]	58.5	60.0	31.7	40.0	53.7	56.0	66.1	69.6	55.5	58.6
RMFN [10]	56.9	60.3	17.3	25.6	55.4	57.3	65.5	70.8	53.2	57.2
TFN [42]	60.0	61.9	19.3	28.0	53.4	57.3	76.4	**72.9**	56.7	58.7
LMF [43]	46.6	54.7	34.5	40.6	49.8	54.3	80.1	**72.9**	53.6	57.0
MARN [59]	55.1	59.6	27.1	35.1	57.2	57.4	70.4	71.2	55.2	58.4
MulT [9]	**64.9**	**64.2**	19.9	29.6	56.8	58.5	79.3	70.9	**60.2**	59.7
MMUU-BA [10]	57.0	60.0	**35.6**	41.8	**58.2**	61.2	75.5	71.9	58.7	**60.5**
RAVEN [47]	33.6	42.6	0.7	1.4	14.5	23.2	21.4	32.7	22.0	30.3
MFN [13]	49.4	55.6	35.1	**42.1**	56.2	55.5	64.5	67.3	52.4	56.5

Overall, we see that all approaches attain a lower binary performance compared to the reported one in literature, except RAVEN, achieving a higher performance on both MOSEI and MOSI, and MMUU-BA achieving a higher accuracy on MOSEI. RAVEN is the most effective model for the MOSI task, MMUU-BA for MOSEI, and MulT for IEMOCAP. That is, attention mechanism-based approaches are the most effective for human multimodal affection recognition tasks. MulT is a robust competitive model, but in contrast to the literature, we found that it does not attain the highest

performance on sentiment analysis task. Yet, without considering efficiency, we noticed that MulT, MMUU-BA, and RAVEN are the most appropriate models for sentiment analysis, whilst MMUU-BA and MulT the most appropriate ones for emotion recognition. While RAVEN showed outstanding performance for the sentiment analysis tasks, it yields the lowest performance for the emotion recognition task.

Error Analysis. We conduct an error analysis on the above experiments. Figure 2 shows the percent error[4] per sentiment class on MOSI. Each line style corresponds to the taxonomy of the SOTA approaches. Despite MOSI is a relatively balanced dataset, consisting of 1176 positive and 1023 negative utterances, all fusion modality approaches yield a higher percent error for the positive sentiment class compared to the negative sentiment class (see Figure 2). In particular, most approaches show a twice higher percent error for the positive sentiment class compared to the negative sentiment class. We also noticed that attention-mechanism-based approaches, e.g., MMUU-BA, MulT, and RAVEN, achieve the lowest percent error for the positive sentiment class. Though, tensor-based modality fusion approaches, e.g., TFN and LMF, are more effective in terms of performance for the negative sentiment class. It is worth noting that RAVEN, achieving the lowest percent error for the positive class, yields the highest percent error for the negative class.

Figure 3 depicts the percent error per sentiment class on MOSEI. In contrast to MOSI, all approaches achieve a low percent error for the positive

[4]We define as percent error within a class, the difference between the estimated number and the actual number when compared to the actual number expressed in percent format.

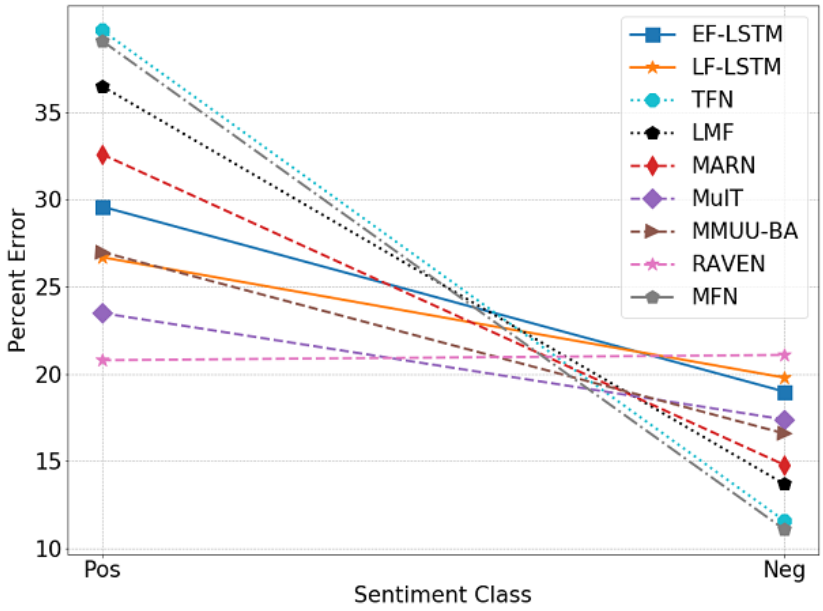

Figure 2: Percent error per sentiment class on MOSI.

sentiment class, whereas they struggle with negative utterances. We suspect this is because MOSEI is an unbalanced dataset. That is, it consists of 11544 positive and 4721 negative utterances. The results show that once we got enough data, there is no any significant difference among different fusion modality approaches in terms of performance (see positive class in Figure 3).

Figure 4 shows the percent error for each emotion on IEMOCAP. The results show that the percent error is high, i.e., greater than 64%, for the Happy emotion class. We suppose that this is due to the small number of samples. Specifically, the Happy emotion class has only 135 samples compared to 383, 193, and 227 of the Neutral, Sad, and Angry emotion class in the test set. This implies that the performance for each emotion class is analogous to the number of samples for each class. Though, some approaches, such as

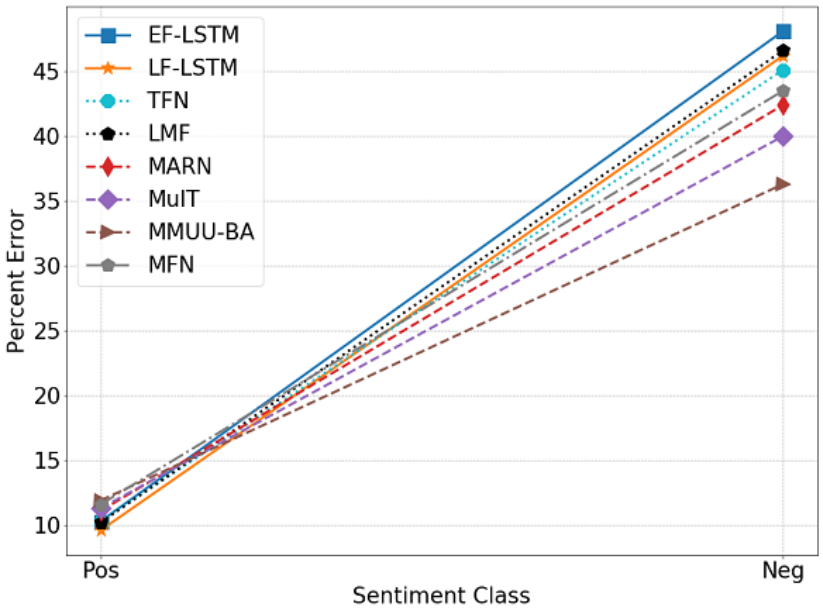

Figure 3: Percent error per sentiment class on MOSEI.

MMUU-BA and MulT, are more effective than others, such as RAVEN and MFN. That is, there is a considerable variance in percent error across different modality fusion approaches.

We further carry out the following analysis on test outputs of MOSI. We group the outputs of all the samples in the test dataset. The first group (i.e., easy) is 49 cases where all methods predict correctly; the second group (i.e., medium) is 21 cases where half the methods predict correctly; the third (i.e., hard) is 18 cases where 9 out of 11 methods predict correctly; and the fourth (i.e., very hard) is 15 cases where all methods predict wrongly. We expose four samples for each group in Table 4.

Out of 686 utterances, 49 ones, that is 7.1%, are predicted correctly by all approaches. These are usually sentences consisting of highly sentiment

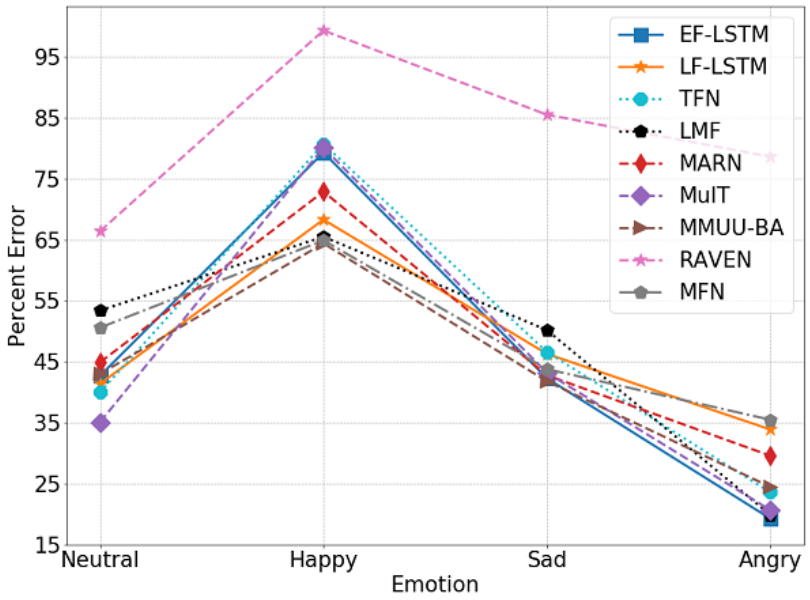

Figure 4: Percentage error per emotion class on IEMOCAP.

words such as "horrible", "love", etc (see Table 4, Easy category). Only 21 utterances, 3.1%, predicted correctly by half of the approaches. All those utterances are either neutral or positive. For example, one possible reason that approaches fail to make a correct prediction for utterances "*But it does have some adult humour*" and "*It actually surprised me*" (see Table 4, Medium category) is due to missing content. 18 utterances, that is 2.6%, can not be correctly predicted by 9 out of 11 approaches, even though utterances include highly sentiment words like "pretty girl", "laughing", etc (see Table 4, Hard category). Finally, all approaches can not predict 15 utterances, that is 2.2%. Utterances like "*Everything that happened in Shrek 1,2, and 3 are wiped away*" and "*A lot of people don't like Scream 2*" (see Table 4, Very Hard category) are mainly dominated by highly negative words, but the

Table 4: Error Cases across all approaches on MOSI

Category	Case	Label
Easy (100%)	This movie was horrible.	Neg.
	I had no idea why I even saw this movie.	Neg.
	This movie seemed um a little long.	Neg.
	You will really love this movie if you are 8.	Pos.
Medium (50%)	But it does have some adult humour.	Pos.
	He is a pretty average guy.	Pos.
	The two women in this movie are particularly good looking.	Pos.
	It actually surprised me.	Pos.
Hard (20%)	They are back to you having two killers thankfully.	Pos.
	She is a really pretty girl.	Pos.
	It had me laughing out loud.	Pos.
	Not a bad idea for a sequel.	Pos.
Very Hard (0%)	Who I don't usually like.	Pos.
	I did like Transformers 2 even though a lot of people didn't like that.	Pos.
	A lot of people don't like Scream 2.	Pos.
	Everything that happened in Shrek 1,2, and 3 are wiped away.	Pos.

overall sentiment is positive. It is worth mentioning that all the error cases of medium, hard, and very hard group are positive sentiment utterances only.

To our knowledge, this is a novel finding.

4.2. Efficiency

In experiment 2, we report the model sizes (i.e., parameters), the training time of learning, and the validation set convergence. We illustrate the validation set convergence across all competitive approaches on MOSI, MOSEI and IEMOCAP in Figure 5, Figure 6, and Figure 7 respectively. For MOSI, we empirically find that MMUU-BA converges faster to better results at training compared to other approaches (see Figure 5). RAVEN shows a more stabilized mean absolute error (MAE) at training compared to MulT, but it is still higher compared to MMUU-BA. In general, all approaches converge quite fast, up to 10 epochs. We assume that this is due to the small data size. We observe that MCTN needs much more than 50 epochs to converge.

For MOSEI, we observe that EF-LSTM, LF-LSTM, TFN, LMF, and MARN increase the MAE after 5 epochs (see Figure 6). A possible explanation for this might be due to overfitting since MOSEI is a large dataset. MulT and RAVEN show a pretty destabilized MAE at training. Despite RAVEN being among the most robust approaches on MOSEI in terms of binary accuracy, it achieves the highest MAE among all approaches (see Figure 6). Finally, we empirically find that MMUU-BE converges faster to better results attaining the lowest MAE.

For IEMOCAP, most of the approaches increase the cross-entropy loss after the 5^{th} epoch (see Figure 7). Only RAVEN and MulT attain a low and stabilized cross-entropy loss. Specifically, MulT, reporting the best recall performance for the "Neutral" class, attains the lowest cross-entropy loss. EF-LSTM, achieving an improved performance as compared to other complex

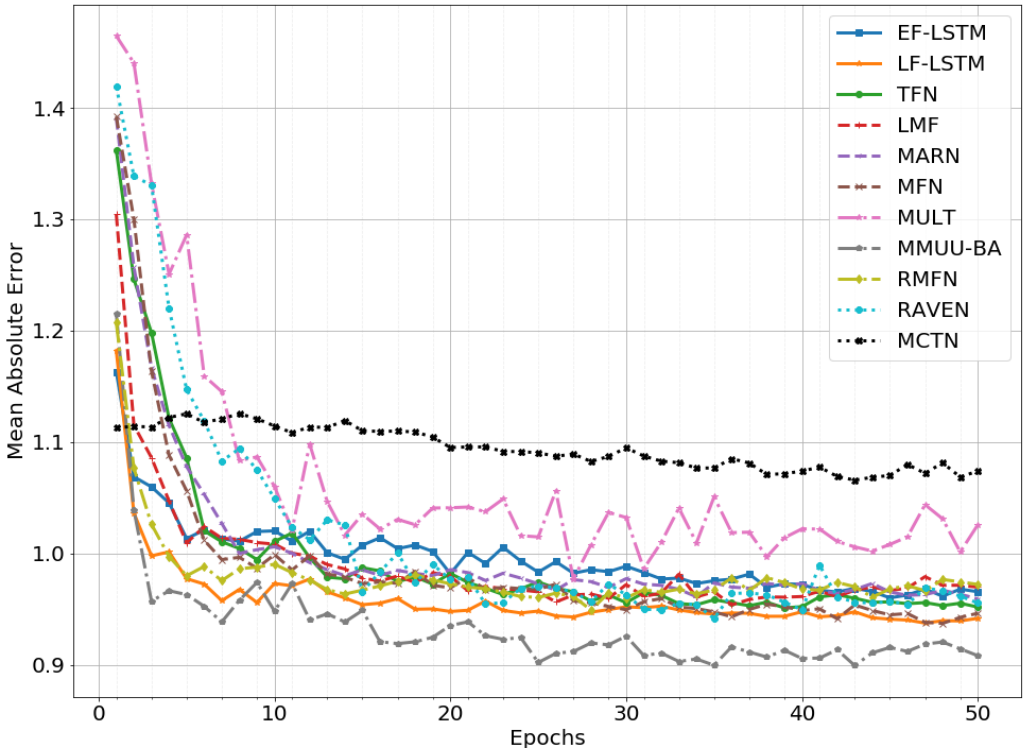

Figure 5: Validation set convergence across the SOTA approaches on MOSI task

competitive approaches, shows a fair and stabilized loss at training until 25^{th} epoch.

We investigate the complexity of models by presenting the number of parameters and training time in minutes for MOSI, MOSEI, and IEMOCAP in Table 5. We observe that approaches integrating LSTMCell components, such as LF-LSTM, MARN, and RMFN, are not able to speedup. For LSTMCell, being a variant of LSTM, Pytorch can not currently maintain the same speed. Despite the low performance, tensor-based approaches attain significant speedup during inference. For MOSI, MMUU-BA is faster than RAVEN, even though the latter has fewer parameters. We attribute this slowdown

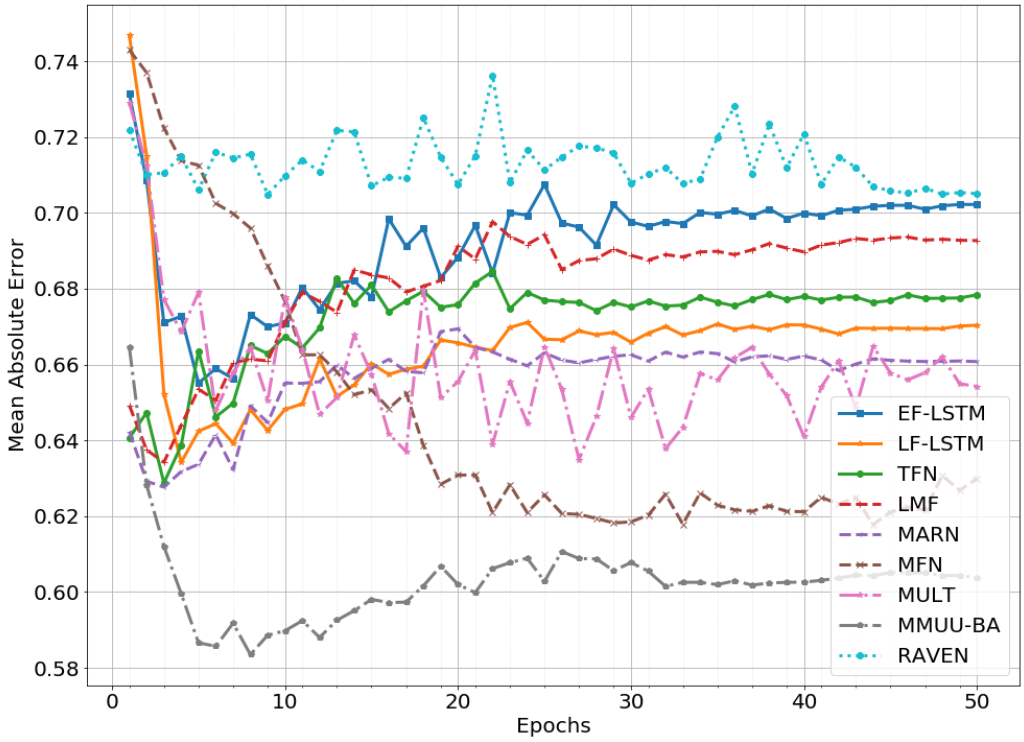

Figure 6: Validation set convergence across the SOTA approaches on MOSEI task

to the LSTMCell component of RAVEN. MulT, being a more complicated model, requires a longer time (i.e., 17.6 minutes) compared to MMUU-BA and RAVEN (i.e., 0.64 and 3.71 minutes respectively). We observed similar behaviour for MOSEI. Even though MOSEI is a relatively large dataset compared to MOSI, some models have fewer parameters on MOSEI compared to MOSI. A possible explanation for this might be because of seting up different configuration settings after the fine-tuning process. For IEMOCAP, EF-LSTM is not only effective but also an efficient approach attaining a significant (26 times) speedup over its counterpart (i.e., MulT) in terms of performance.

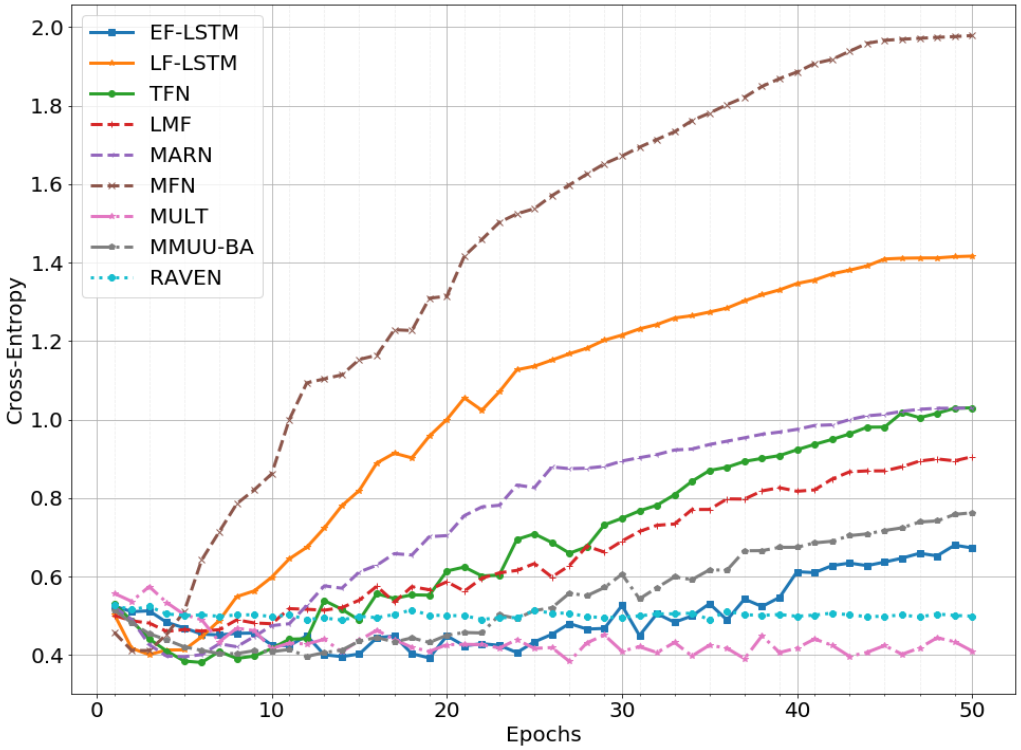

Figure 7: Validation set convergence across the SOTA approaches on IEMOCAP task

4.3. Ablation studies

To address the third research question, we design various ablation studies to analyse a) the importance of modalities and b) important components for learning crossmodal interactions. We conduct all ablation studies on MOSI.

4.3.1. Importance of modalities

To understand the importance of modalities in multimodal tasks, we conduct ablation studies on TFN, which inherently models unimodal, bimodal, and trimodal interactions, and MulT, which attains high accuracy both on sentiment analysis and emotion recognition tasks. For TFN, we test the

Table 5: Complexity of models. Training time and learning parameters on MOSI, MOSEI, and IEMOCAP tasks

	MOSI		MOSEI		IEMOCAP	
Approach	Mins.	Params.	Mins.	Params.	Mins.	Params.
EF-LSTM [40]	0.43	177,329	6.59	217,457	1.40	206,152
LF-LSTM [40]	3.14	1,155,109	54.47	5,111,485	3.59	946,756
RMFN [10]	57.42	1,950,805	-	-	20.85	1,732,884
TFN [42]	0.51	14,707,911	1.87	6,804,859	0.53	23,198,398
LMF [43]	0.43	1,144,493	2.00	5,079,473	1.12	962,116
MARN [59]	69.5	1,350,389	268.20	5,442,313	4.6	1,362,116
MulT [9]	17.6	1,071,211	31.20	874,651	36.89	1,074,998
MMUU-BA [10]	0.64	2,424,965	7.07	2,576,165	0.79	2,605,484
RAVEN [47]	3.71	171,433	23.87	159,213	3.00	173,680
MFN [13]	1.88	1,513,221	18.56	415,521	5.13	1,325,508
MCTN [14]	15.64	147,100	-	-	-	-

TFN approach with unimodal, bimodal, and trimodal subtensors. Table 6 shows the results of ablation studies. We observed that language is the most informative modality being a pivot for visual and acoustic modalities. The unimodal visual, acoustic subnetworks and the bimodal visual-acoustic subnetwork attain a pretty low accuracy compared to those ones integrating the linguistic modality. Specifically, combining language with visual or acoustic is generally better than combining the visual and acoustic modalities. In contrast to [42], we found that the language-based subnetwork performs similarly to the trimodal tensor network in terms of the binary accuracy.

That is, our experiments showed that tensor-based fusion is not an effective approach for modelling crossmodal interaction across three modalities.

Table 6: Comparison of TFN with its subtensor variants on MOSI.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
TFN_l	31.3	**75.7**	75.6	1.017	**0.756**
TFN_v	17.3	53.2	50.5	1.465	0.125
TFN_a	15.2	56.6	54.4	1.425	0.181
$\text{TFN}_{l,v}$	30.3	75.1	75.0	1.013	0.610
$\text{TFN}_{l,a}$	31.1	75.9	75.9	**1.012**	0.624
$\text{TFN}_{v,a}$	15.4	56.9	55.5	1.414	0.178
$\text{TFN}_{w/oc}$	35.7	75.1	74.9	1.024	0.605
$\text{TFN}_{l,v,a}$ [42]	**34.9**	75.6	75.5	1.009	0.605

For MulT, we first consider the performance for linguistic, visual, acoustic only transformers. We found a binary accuracy of 79.5% of the language transformer compared to 77.4% in literature [9]. The language transformer significantly outperforms the visual- and acoustic-only transformers (see Table 7).

We also study the importance of individual crossmodal transformers according to the target modality (i.e., $L, V \rightarrow A$, $V, A \rightarrow L$, and $L, A \rightarrow V$). Among the three crossmodal transformers, the one where acoustic is the target modality works best. This result is consistent with [14], but in contrast with [9], reporting that presenting language as a target modality leads to best performance. The experiments show that there is no need to consider multiple directional pairwise crossmodal transformers. Specifically, when we

consider acoustic as a target modality yields an increased accuracy of 79.6% compared to 78.7% of MulT. Though, there is no any statistical difference in performance among the three crossmodal transformers and the multiple directional pairwise crossmodal transformer (i.e., MulT).

Table 7: Comparison of MulT with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
MulT_l	**34.3**	**79.5**	79.2	**0.939**	0.662
MulT_v	20.9	59.7	58.3	1.401	0.154
MulT_a	18.75	60.5	60.1	1.348	0.211
$\text{MulT}_{v,a\rightarrow l}$	31.3	76.7	76.5	1.037	0.604
$\text{MulT}_{l,a\rightarrow v}$	32.6	78.9	78.7	0.993	**0.787**
$\text{MulT}_{l,v\rightarrow a}$	33.6	**79.6**	79.4	0.996	0.663
MulT_{H_5}	31.9	79.0	78.8	1.014	0.662
$\text{MulT}_{H_{10}}$	33.5	79.0	79.0	0.995	0.667
MulT [9]	33.6	78.7	78.4	0.964	0.662

4.3.2. Important Modules for Crossmodal Interactions

To understand the influence of individual components for modelling crossmodal interactions, we perform comprehensive ablation analysis over the SOTA approaches on MOSI. First, we study the importance of extra dimensions with value 1 of $\text{TFN}_{l,v,a}$ [42], which models unimodal and bimodal dynamics, besides trimodal ones. We found that the TFN version without constant ($\text{TFN}_{w/oc}$ in Table 6) reports a decreased accuracy of 75.1% compared to 75.6% of TFN. Though, for Acc_7, the model improves from 34.9% to 35.7% when comparing $\text{TFN}_{l,v,a}$ to $\text{TFN}_{w/oc}$.

For MulT, we consider the number of heads in crossmodal attention module. We experiment with 5 and 10 heads (MulT_{H_5} and $\text{MulT}_{H_{10}}$ in Table 7 respectively). We did not observe any difference in terms of binary accuracy. Though, for Acc_7, the more heads yield an increased performance of 33.5% compared to 31.9% (see Table 7).

In [59], authors claim that for each timestamp, there might exist multiple crossmodal interactions. We experiment with three variants of MARN to investigate the number of attentions needed to extract all crossmodal dynamics. Specifically, we try one, five, and ten attentions. In contrast to [59], our experiments show that the MARN with only one attention slightly outperforms the models with multiple attentions in terms of binary accuracy (see Table 8). Yet, the MARN with five attentions outperforms the other two variants, for Acc_7. We also remove the multi-attention block (MAB) from MARN. Specifically, we replace the MAB with a fully-connected layer and remove the softmax function. We observe that there is no any effect on binary accuracy (see Table 8) whilst for Acc_7, the difference is marginal.

Table 8: Comparison of MARN with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\text{MARN}_{K=1}$	30.9	**76.9**	76.7	0.983	**0.629**
$\text{MARN}_{K=5}$	31.5	76.1	76.0	1.001	0.616
$\text{MARN}_{K=10}$	30.9	76.4	76.2	1.012	0.621
$\text{MARN}_{w/oMAB}$	**32.4**	76.4	76.2	**0.979**	0.622
MARN [59]	31.8	76.4	76.2	0.984	0.625

For MMUU-BA, we analyze the attention module to understand its learn-

ing behaviour. We experiment with two other variants of MMUU-BA (see Table 9). The architecture of these variants differs with respect to the attention computation module. Particularly, in MMUU-UA, we compute one-directional attention, e.g., from linguistic to visual modality only. In MMUU-SA, we only compute self-attention within modalities. We found that one-directional attention results in an increased binary accuracy of 78.8% compared to 78.2% of the proposed framework. Both MMUU-UA and MMUU-BA attain the same performance, for Acc_7 (see Table 9). For the self-attention approach, we found that it is less effective than the one-directional crossmodal attention, but more effective than the bi-directional crossmodal attention.

Table 9: Comparison of MMUU with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
MMUU-UA	**33.8**	**78.8**	78.6	**0.925**	0.680
MMUU-SA	32.0	78.6	78.5	0.950	**0.688**
MMUU-BA [10]	**33.8**	78.2	78.1	0.947	0.675

For MFN, first, we investigate if crossmodal interactions can happen over multiple time instances. Specifically, we experiment with a variant of MFN by shrinking the context from time t and $t-1$ to only the current timestamp t in the memory component. We found that $\text{MFN}_{w/o\Delta}$ (see Table 10) significantly underperforms the MFN approach. This implies that we should not model crossmodal interactions on aligned time steps, but consider long-range crossmodal contingencies across a multimodal sequence. Second, we evaluate the importance of spatial-temporal crossmodal interactions through time

by removing all memory components. The results show the effectiveness of memory components on the proposed approach. Both outcomes agree with the reported experiments in [13].

Table 10: Comparison of MFN with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\text{MFN}_{w/o\Delta}$	31.5	73.8	73.8	1.042	0.584
$\text{MFN}_{w/oMemory}$	31.6	75.0	74.8	1.011	0.598
MFN [13]	**31.9**	**76.2**	75.8	**0.988**	**0.662**

For RAVEN, we have already removed the Nonverbal Subnetworks [47] as mentioned in 3.5 Section. This modification results in an increased binary accuracy of 78.6% compared to 78.0% in [47] on MOSI. We also investigate the temporal interactions between the nonverbal "subword" units with language utterances. Specifically, we remove the shift component, which learns to dynamically shift the text representation by integrating the nonverbal vector. Practically, visual and acoustic representations are concatenated with the word embeddings before being fed to downstream networks. We found that integrating the nonverbal context with words is beneficial for understanding human language. Specifically, Raven shows a significantly increased binary performance of 78.6% compared to 75.6% of $\text{RAVEN}_{w/oShift}$.

For RMFN, decomposing the fusion problem into multiple stages, we experiment with the number of stages needed for modelling crossmodal dynamics. Specifically, we experiment with one, three, and six stages. Our experiments show that RMFN attains a similar performance even we apply one or six stages for fusing information.

Table 11: Comparison of RAVEN with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
RAVEN$_{w/oShift}$	31.8	75.6	75.5	1.016	0.615
RAVEN [47]	**34.6**	**78.6**	**78.6**	**0.948**	**0.674**

Table 12: Comparison of RMFN with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
RMFN$_{s=1}$	32.9	75.3	75.2	**0.982**	0.616
RMFN$_{s=3}$	32.5	75.5	75.3	0.991	**0.623**
RMFN$_{s=6}$	**33.1**	**75.6**	75.5	0.991	0.613
RMFN[10]	31.7	75.2	75.1	1.005	0.612

Overall, we found that linguistic modality is a pivot for visual and acoustic modalities. This basic finding is consistent with literature. Yet, the results from ablation studies are not always following findings reported in literature. In particular, we found that:

- fusing multimodal information into multiple levels (e.g., MulT, MARN, and RMFN) does not necessarily result in better binary performance. In some cases, fusing information into multiple levels might achieve slightly better fine-grained accuracy, that is, Acc_7;
- tensor-based approaches underperforms the linguistic modality;
- integrating the temporal (e.g., MFN) or modality (e.g., RAVEN) context over the multimodal fusion process results in a significantly better performance.

5. Discussion

In this paper, we replicate the most recent SOTA models for multimodal language analysis. We evaluate their effectiveness through comprehensive comparative studies, error analyses and series of ablation studies. The efficiency of the models is also compared in terms of three evaluation metrics, namely, parameters, training time, and validation set convergence. The results associated with ablation studies help us find out which components and methodologies contribute most to solve the problem of affective computing.

In terms of effectiveness, the experiments showed that approaches exploiting attention mechanism components improve the model performance for both sentiment analysis and emotion recondition tasks. We speculate that this is because the attention mechanism acts as an implicit multimodal alignment component. Memory networks reached a similar performance as well. On the other side, despite tensor-based approaches got a lower present error for the negative sentiment class on MOSI, in general, they did not attain high performance. Similarly, recurrent cell-based approaches do not achieve a high performance either. Overall, most of the SOTA approaches attain lower performance in a range of 2% to 4.5% compared to the reported one in the literature. We mainly attribute such discrepancies to the fine-tuning process. The different versions of MOSEI and MOSI datasets used in published works could be another reason for most of those cases.

From the efficiency viewpoint, attention mechanism-based approaches are usually more complex and require more training time as compared to the rest of modality fusion approaches. To alleviate that issue, we could consider less fine-grained crossmodal interactions. Indeed, ablation studies showed that

adding more levels of interactions across modalities results in a decreased performance. Recurrent cell-based approaches are extremely computationally expensive. On the other side, memory and tensor networks are more efficient.

Overall, our results demonstrate that attention mechanisms are the most effective component for affective computing tasks despite being computationally expensive. Crucially, our ablation studies showed that crossmodal interactions are not aligned on corresponding time steps, but spread across a multimodal sequence. Though, little effort has been devoted towards this direction. A further study in this direction would be to investigate approaches which exploit crossmodal interactions across a multimodal sequence instead of corresponding timestamps. Finally, multimodal sentiment analysis can benefit from the integration of context. In the future, it would be worth investigating how multimodal sentiment analysis could be enhanced by considering proceeding utterances and existing knowledge bases, which might entail sentiment or emotional knowledge.

One limitation of our study is that we use a simple approach to align modalities. Following previous work, we average visual and acoustic modalities throughout word intervals since advancing the SOTA is not the aim of this work. Yet, further investigation is needed in this direction to find out if other alignment approaches could enhance the relatively poor performance of the non-verbal modalities. In term of the implementation, we noticed that the LSTMCell component cannot speed up. That made approaches which primely utilize recurrent cell components less efficient.

6. Conclusion

We have replicated and proposed a large-scale empirical comparison among SOTA approaches for multimodal human language analysis. We thoroughly investigated both their effectiveness and efficiency on two human multimodal affection recognition tasks and found out important components in multimodal language models. The results showed that attention mechanism approaches are the most effective for both sentiment analysis and emotion recognition tasks, even though they are not computationally cheap. Besides, components being able to capture crossmodal interactions across different timestamps, integrate context and utilize linguistic modality as a pivot for the non-verbal ones achieved improved performance. It is worth mentioning that positive sentiment utterances are the most challenging cases for all modality fusion approaches. To our knowledge, this is a novel finding. In the future, we are going to focus on conversational video sentiment analysis tasks in that the utterance context has been proved to be beneficial for understanding human language.

ACKNOWLEDGEMENT

This work supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721321.

References

[1] W. Wu, Z. Guo, X. Zhou, H. Wu, X. Zhang, R. Lian, H. Wang, Proactive human-machine conversation with explicit conversation goal, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3794–3804. doi:10.18653/v1/P19-1369.
URL https://www.aclweb.org/anthology/P19-1369

[2] P. Pham, J. Wang, Predicting learners' emotions in mobile mooc learning via a multimodal intelligent tutor, in: R. Nkambou, R. Azevedo, J. Vassileva (Eds.), Intelligent Tutoring Systems, Springer International Publishing, Cham, 2018, pp. 150–159.

[3] A. Prange, M. Niemann, A. Latendorf, A. Steinert, D. Sonntag, Multimodal speech-based dialogue for the mini-mental state examination, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3290607.3299040.
URL https://doi.org/10.1145/3290607.3299040

[4] S. S. Rajagopalan, L.-P. Morency, T. Baltrušaitis, R. Goecke, Extending long short-term memory for multi-view structured learning, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 338–353.

[5] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: 2016

IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 439–448.

[6] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, 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), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 873–883. doi:10.18653/v1/P17-1081.
URL https://www.aclweb.org/anthology/P17-1081

[7] H. Wang, A. Meghawat, L. Morency, E. P. Xing, Select-additive learning: Improving generalization in multimodal sentiment analysis, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 949–954.

[8] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intelligent Systems 31 (2016) 82–88.

[9] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569. doi:10.18653/v1/P19-1656.
URL https://www.aclweb.org/anthology/P19-1656

[10] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, P. Bhattacharyya, Contextual inter-modal attention for multi-modal sentiment analysis, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3454–3466. doi:10.18653/v1/D18-1382.
URL https://www.aclweb.org/anthology/D18-1382

[11] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, I. Marsic, Multimodal affective analysis using hierarchical attention strategy with word-level alignment, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2225–2235. doi:10.18653/v1/P18-1207.
URL https://www.aclweb.org/anthology/P18-1207

[12] P. P. Liang, Z. Liu, A. Bagher Zadeh, L.-P. Morency, Multimodal language analysis with recurrent multistage fusion, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 150–161. doi:10.18653/v1/D18-1014.

[13] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: AAAI Conference on Artificial Intelligence, 2018a, p. 5634–5641.

[14] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic transla-

tions between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6892–6899.

[15] S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, arXiv preprint arXiv:1911.07848.

[16] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia systems 16 (6) (2010) 345–379.

[17] S. Sun, A survey of multi-view machine learning, Neural computing and applications 23 (7-8) (2013) 2031–2038.

[18] D. Ramachandram, G. W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Processing Magazine 34 (6) (2017) 96–108.

[19] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2019) 423–443.

[20] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125.

[21] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos (2016). arXiv:1606.06259.

[22] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246. doi:10.18653/v1/P18-1208.
URL https://www.aclweb.org/anthology/P18-1208

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (4) (2008) 335.

[24] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[25] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 169–176. doi:10.1145/2070481.2070509.
URL https://doi.org/10.1145/2070481.2070509

[26] S. Ghosh, E. Laksana, L.-P. Morency, S. Scherer, Representation learn-

ing for speech emotion recognition., in: Interspeech, 2016, pp. 3603–3607.

[27] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[28] A. P. James, B. V. Dasarathy, Medical image fusion: A survey of the state of the art, Information fusion 19 (2014) 4–19.

[29] M. U. Bokhari, F. Hasan, Multimodal information retrieval: Challenges and future trends, International Journal of Computer Applications 74 (14).

[30] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, IEEE transactions on affective computing.

[31] S. K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Computing Surveys (CSUR) 47 (3) (2015) 1–36.

[32] E. Morvant, A. Habrard, S. Ayache, Majority vote of diverse classifiers for late fusion, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, 2014, pp. 153–162.

[33] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: Proceedings of the 2016

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 160–170.

[34] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, F. Schwenker, Multiple classifier systems for the classification of audio-visual emotional states, in: S. D'Mello, A. Graesser, B. Schuller, J.-C. Martin (Eds.), Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 359–368.

[35] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, F. Schwenker, Multiple classifier systems for the classification of audio-visual emotional states, in: S. D'Mello, A. Graesser, B. Schuller, J.-C. Martin (Eds.), Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 359–368.

[36] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics, in: 2013 IEEE Symposium on Computational Intelligence for Human-like Intelligence (CIHLI), 2013, pp. 108–117.

[37] A. Lazaridou, N. T. Pham, M. Baroni, Combining language and vision with a multimodal skip-gram model, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 153–163.

doi:10.3115/v1/N15-1016.
URL https://www.aclweb.org/anthology/N15-1016

[38] R. Kiros, R. Salakhutdinov, R. S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models., CoRR abs/1411.2539.
URL http://dblp.uni-trier.de/db/journals/corr/corr1411.html#KirosSZ14

[39] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, S. Khudanpur, Recurrent neural network based language model, in: INTERSPEECH, 2010, p. 1045–1048.

[40] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[41] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[42] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1103–1114. doi:10.18653/v1/D17-1115.
URL https://www.aclweb.org/anthology/D17-1115

[43] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion

with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2247–2256. doi:10.18653/v1/P18-1209.

[44] E. J. Barezi, P. Fung, Modality-based factorization for multimodal fusion, in: I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, M. Rei (Eds.), Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019, Association for Computational Linguistics, 2019, pp. 260–269. doi:10.18653/v1/w19-4331.
URL https://doi.org/10.18653/v1/w19-4331

[45] P. P. Liang, Z. Liu, Y.-H. Tsai, Q. Zhao, R. Salakhutdinov, L.-P. Morency, Learning representations from imperfect time series data via tensor rank regularization, in: ACL, 2019.

[46] S. Mai, H. Hu, S. Xing, Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 481–492. doi:10.18653/v1/P19-1046.
URL https://www.aclweb.org/anthology/P19-1046

[47] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, L.-P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal

behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7216–7223.

[48] S. H. Dumpala, I. Sheikh, R. Chakraborty, S. K. Kopparapu, Audio-visual fusion for sentiment classification using cross-modal autoencoder, NIPS, 2019.

[49] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 163–171.

[50] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognition Letters 125 (264-270).

[51] P. P. Liang, Y. C. Lim, Y.-H. H. Tsai, R. Salakhutdinov, L.-P. Morency, Strong and simple baselines for multimodal utterance embeddings, arXiv preprint arXiv:1906.02125.

[52] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6818–6825.

[53] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2122–2132. doi:10.18653/v1/N18-1193.
URL https://www.aclweb.org/anthology/N18-1193

[54] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.

[55] T. Young, V. Pandelea, S. Poria, E. Cambria, Dialogue systems with audio context, Neurocomputing.

[56] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, arXiv preprint arXiv:1908.11540.

[57] Y. Zhang, Q. Li, D. Song, P. Zhang, P. Wang, Quantum-inspired interactive networks for conversational sentiment analysis, in: IJCAI, 2019.

[58] E. Cambria, Affective computing and sentiment analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107.

[59] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: AAAI Conference on Artificial Intelligence, 2018, p. 5642–5649.

[60] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[61] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

[62] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL, 2018.

[63] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep — a collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 960–964.

[64] J. Yuan, M. Liberman, Speaker identification on the scotus corpus, Journal of the Acoustical Society of America 123 (5) (2008) 3878.

[65] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, D.-Y. Li, S.-H. Li, Electroencephalogram emotion recognition based on empirical mode decomposition and optimal feature selection, IEEE Transactions on Cognitive and Developmental Systems 11 (4) (2018) 517–526.

Appendices

A. Fine-tuning Final Settings

Table 13: Hyperparameters of EF-LSTM we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	64	64	16
Initial Learning Rate	0.002	0.002	0.001
LSTM Output	96	128	128
Multimodal Embedding Dimension	64	128	16
Multimodal Embedding Dropout	0.1	0.2	0.1
Gradient Glip	0.4	0.8	0.3

Table 14: Hyperparameters of LF-LSTM we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	64	16	16
Initial Learning Rate	0.005	0.001	0.001
LSTM Outputs	128,16,80	128,64,16	128,64,16
Multimodal Embedding Dimension	32	48	32
Multimodal Embedding Dropout	0.2	0.4	0.2
Gradient Glip	0.4	0.3	0.7

Table 15: Hyperparameters of TFN we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	96	128	128
Initial Learning Rate	0.001	0.002	0.001
Subnetwork Outputs	128,80,80	128,16,32	128,80,60
Subnetwork Dropout Probabilities	0.1,0.1,0.1	0.2,0.2,0.2	0.5,0.5,0.5
Sentiment Subnetwork Outpout	16	96	128
Sentiment Subnetwork Probability	0.4	0.3	0.4
Gradient Glip	0.1	0.1	0.5

Table 16: Hyperparameters of LMF we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	96	128	32
Initial Learning Rate	0.001	0.002	0.001
Rank	4	4	16
Subnetwork Outputs	128,32,80	128,64,32	128,64,32
Subnetwork Dropout Probabilities	0.5,0.5,0.5	0.1,0.1,0.1	0.3,0.3,0.3
Crossmodal Representation	0.2	0.2	0.4
Gradient Glip	0.2	0.2	0.4

Table 17: Hyperparameters of MARN we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	128	16	64
Initial Learning Rate	0.001	0.002	0.001
LSTM Outputs	128,64,80	128,80,80	128,80,32
Attention Blocks	2	2	5
Attention Cell	16	64	32
Compressed dimension	64,32,8	64,40,40	64,16,8
Output cell dimension	16	16	96
Gradient Glip	0.1	0.2	0.7

Table 18: Hyperparameters of MFN we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	128	128	32
Initial Learning Rate	0.001	0.002	0.005
LSTM Outputs	128,80,16	128,80,16	128,64,16
γ_1, γ_2 cell dimensions	128,128	128,128	64,32
Attention cell dimensions	64,32	64,32	256,32
Memory dimension	256	256	256
Output cell dimension	64	64	128
Gradient Glip	0.2	0.2	0.7

Table 19: Hyperparameters of MulT we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	16	128	32
Initial Learning Rate	0.001	0.002	0.005
Transformers Hidden Unit Size	40	40	40
Crossmodal Blocks	4	4	4
Crossmodal Attention Heads	8	10	10
Temporal Convolution Kernel Size	3/3/3	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 Glip	0.2	0.2	0.7

Table 20: Hyperparameters of MMUU-BA we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	64	64	64
Initial Learning Rate	0.005	0.002	0.001
RNN dropouts	0.15,0.15,0.15	0.1,0.1,0.1	0.7,0.7,0.7
GRU dropouts	0.1,0.1,0.1	0.3,0.3,0.3	0.15,0.15,0.15
FC dropouts	0.15,0.15,0.15	0.8,0.8,0.8	0.15,0.15,0.15
Output cell dimensions	32	32	64
Output dropout	0.15	0.3	0.1
Gradient Glip	0.3	0.9	0.5

Table 21: Hyperparameters of RMFN we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	64	128	16
Initial Learning Rate	0.005	0.002	0.002
Shift Weight	0.8	0.7	0.1
LSTM layers	3	1	1
Cell Output	50	40	30
Gradient Glip	0.7	1	0.1

Table 22: Hyperparameters of RMFN we use for the various tasks.

	MOSI	MOSEI	IEMOCAP
Batch Size	64	128	16
Initial Learning Rate	0.005	0.002	0.002
Shift Weight	0.8	0.7	0.1
LSTM layers	3	1	1
Cell Output	50	40	30
Gradient Glip	0.7	1	0.1