Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis

Dimitris Gkoumas [a,*], Qiuchi Li [b], Christina Lioma [c], Yijun Yu [a], Dawei Song [a,d,*]

[a] The Open University, United Kingdom
[b] University of Padua, Italy
[c] University of Copenhagen, Denmark
[d] Beijing Institute of Technology, China

ARTICLE INFO

Keywords:
Multimodal human language understanding
Video sentiment analysis
Emotion recognition
Reproducibility in multimodal machine learning

ABSTRACT

Multimodal video sentiment analysis is a rapidly growing area. It combines verbal (i.e., linguistic) and non-verbal modalities (i.e., visual, acoustic) to predict the sentiment of utterances. A recent trend has been geared towards different modality fusion models utilizing various attention, memory and recurrent components. However, there lacks a systematic investigation on how these different components contribute to solving the problem as well as their limitations. This paper aims to fill the gap, marking the following key innovations. We present the first large-scale and comprehensive empirical comparison of eleven state-of-the-art (SOTA) modality fusion approaches in two video sentiment analysis tasks, with three SOTA benchmark corpora. An in-depth analysis of the results shows that the attention mechanisms are the most effective for modelling crossmodal interactions, yet they are computationally expensive. Second, additional levels of crossmodal interaction decrease performance. Third, positive sentiment utterances are the most challenging cases for all approaches. Finally, integrating context and utilizing the linguistic modality as a pivot for non-verbal modalities improve performance. We expect that the findings would provide helpful insights and guidance to the development of more effective modality fusion models.

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 understand human language effectively 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, 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 [5–8]. A recent trend in multimodal embedding learning research is to build more complex models utilizing attention, memory, and recurrent components [9–15]. Various review papers have surveyed the advancements in multimodal machine learning [16–20]. 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

[*] Corresponding authors.
E-mail addresses: dimitris.gkoumas@open.ac.uk (D. Gkoumas), qiuchili@dei.unipd.it (Q. Li), c.lioma@di.ku.dk (C. Lioma), yijun.yu@open.ac.uk (Y. Yu), dawei.song@open.ac.uk (D. Song).

https://doi.org/10.1016/j.inffus.2020.09.005
Received 30 November 2019; Received in revised form 2 September 2020; Accepted 13 September 2020
Available online 18 September 2020

information, etc. However, none of them has conducted a comprehensive empirical study across different state-of-the-art (SOTA) fusion approaches to multimodal language analysis, intending to provide critical and experimental analysis. Such an extensive empirical evaluation would be useful to find out which aspects in the SOTA approaches are the most effective in solving 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 [21–23], 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 affect 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 Sections 4 and 5, respectively. Finally, Section 6 concludes the paper.

2. Related work

In this section, 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 proliferation of multimedia data (e.g., textual, visual, and acoustic) available in various 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) briefly presents 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 [18,19].

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 [21–23,30]. A comprehensive literature review of multimodal affective analysis frameworks is given by Poria et al. [20]. Besides, Fatemeh et al. [31] survey strategies for emotion recognition from body gestures. However, none of the above surveys provides a comprehensive empirical study of the very new 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 [32]. Late fusion approaches build up diverse classifiers for each modality and then aggregate their decisions by voting [33], averaging [34], weighted sum [35] or a trainable model [36–38]. Hybrid strategies, combining outputs from early fusion and individual unimodal predictions, outperform simple feature-level or decision-level approaches [37]. Early work has pushed some progress towards multimodal language embedding learning [39,40]. A range of neural approaches, such as Recurrent Neural Networks (RNNs) [41], Long Short-Term Memory (LSTM) neural networks [42], and Convolutional Neural Networks (CNNs) [43], have been used to learn language-based multimodal embeddings by fusing either input features per timestamp or unimodal output hidden units [5–8].

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 [44] and disentangling [45,46] multimodal representations. Those approaches fuse unimodal features at the utterance level [44–46], word level [47], or in a hierarchical manner [48]. Recently, Mai et al. [49] exploited the tensor-based strategy to fuse segmented unimodal information for capturing local interactions. Then, the local tensors fed a bidirectional skip connection LSTM to learn global interactions.

Considering human language contains time series and thus requires fusing time-varying signals, a recent trend is to exploit LSTMs and RNNs to fuse unimodal representations at the feature level [12,50]. 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]. In [51], Beard et al. proposed a recursive attention-based memory network for constructing contextualized multimodal embeddings. In contrast to typical RNN, the consecutive cells of the proposed recursive recurrent neural network share the same input.

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 [14,15,52]. In contrast to the previous translation strategies, Wang et al. [53] adopted a parallel translation approach by fusing linguistic with acoustic features and linguistic with visual features independently to eliminate noise interference between modalities. Other fusion strategies incorporated reinforcement learning [54], fuzzy logic [55], bilinear pooling [56], deep canonical correlation analysis [57], hierarchical fusion strategies [58,59], and simple but strong baselines [60].

Recently, attention mechanisms have been exploited to align different modalities, resulting in better-performing modality fusion approaches [9–11,61]. Besides multimodal fusion, attention mechanisms have also been exploited for visual and acoustic feature extraction, yielding improved performance [56]. [9] was considered the SOTA feature-level fusion approach for utterance-level video sentiment analysis for CMU-MOSI, CMU-MOSEI, and IEMOCAP tasks. However, very recently [61] achieved higher performance than [9] on CMU-MOSI and IEMOCAP tasks.

Most of the above strategies are black-box approaches, which come with the price of lacking interpretability. Having said that, some holistic frameworks endowed models with inherent interpretability by separating crossmodal interactions. For instance, the contributions to the prediction from each modality and the interactions between modalities, i.e., bi-modal and tri-modal interactions, have been investigated through an interpretable multimodal fusion framework [62]. Hazarika et al. [63] exploited two subspaces, a joint subspace and a modality-specific subspace, to capture uni-modal and tri-modal interactions. In [61], authors applied seven distinct self-attention mechanisms to the factorized multimodal representation, capturing all possible uni-modal, bi-modal, and tri-modal interactions, simultaneously.

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). Recently a few approaches have been proposed to model long-range crossmodal interactions across multimodal sequences [9,11,50]. 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 are other approaches that consider contextual information from surrounding utterances, thus aiding the sentiment analysis and emotion recognition tasks. Current work utilizes supervised NLP approaches to model contextual interactions among utterances, including recurrent neural networks [6,64], memory networks [65,66], sequence-to-sequence networks [67], graph neural networks [68], and quantum-inspired networks [69]. 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 used for our empirical study of the most recent SOTA multimodal language fusion approaches, in the context of video sentiment and emotion analysis tasks. We first formulated the task on which our study was carried out. Sentiment analysis was a binary multimodal classification task inferring either positive or negation emotions. Emotion recognition was a multimodal multilabel classification task inferring one or more emotions, e.g., happy and joyful. However, both tasks aim to capture emotions of video utterance and fall under affective computing field [70].

3.1. Task definition

The goal is to infer the emotion of utterances from video speakers. Each video consists of N sequential utterances $U = (U_1, \dots, U_i, \dots, U_N)$, where i is the ith utterance. Each utterance U_i is associated with three modalities, namely, 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 $y = (y_1, \dots, y_i, \dots, y_N)$, $y_i \in \mathbb{R}$. 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 $U_i^l = (l_i^1, \dots, l_i^L)$. Similarly for visual and acoustic modalities, it is represented by $U_i^v = (v_i^1, \dots, v_i^L)$ and $U_i^a = (a_i^1, \dots, a_i^L)$, respectively.

3.2. Datasets

We empirically evaluated the SOTA approaches from the last two years on multimodal sentiment analysis tasks by using two SOTA benchmark datasets, 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 evaluated the approaches to the multimodal emotion recognition task using the IEMOCAP dataset [23]. We compared all approaches to 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 2199 short monologue video clips (each lasting the duration of a sentence). It has 1284, 229, and 686 utterances in training, validation, and test sets, respectively. CMU-MOSEI is a larger scale sentiment and emotion analysis dataset made up of 22,777 movie review video clips from more than 1000 online Youtube speakers. The training, validation, and test sets are comprised of 16,265, 1869 and 4643 utterances, respectively. Human annotators labelled each sample 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 used the CMU-Multi-modal Data SDK[1] [22] for feature extraction. Following previous work [9,13,44,45,50,71], we converted video transcripts into 300-dimensional pre-trained Glove word embeddings (glove.840B.300d) [72]. Besides, GloVe embedding is more computationally affordable than other more effective, yet computationally expensive, word embeddings [73,74]. Facet[2] is used to capture facial muscle movement, including per-frame basic and advanced emotions and facial action units. We used VOCAREP [75] 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 extracted visual and acoustic features at a frequency of 15 Hz and 12.5 Hz respectively. For MOSEI, we extracted at a frequency of 15 Hz and 20 Hz. To reach the same time alignment across modalities, we applied a word-level alignment. To align visual and acoustic modalities with words, we used P2FA [76]. Then, to obtain the aligned timesteps, we averaged the visual and audio features within these time ranges. All sequences in the word-aligned case had length 50. For each word the dimension of the feature vector was 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 used IEMOCAP. It consists of 151 videos about dyadic interactions, where professional actors are required to perform scripted scenes that elicit specific emotions. It has 2717, 798, and 938 utterances in training, validation, and test sets, respectively. Human annotators labelled each sample for four emotions (neutral, happy, sad, angry). The labels for every emotion are binary. That allowed us to reduce the multiclass learning problem to a problem solvable using binary classifiers. Following a one-vs-all strategy, for each emotion, we trained a robust classifier to recognize one emotion from all the others. We followed a similar process to the sentiment analysis datasets to extract 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

To evaluate the effectiveness on MOSI and MOSEI tasks, we adopted a series of evaluation performance metrics used in prior work [9,12,13,22]: 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.

To evaluate the effectiveness on IEMOCAP, in contrast to previous work reporting accuracy [9,50], we reported 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. That 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 though they failed to distinguish the emotion of the class from all the others correctly. To evaluate the overall performance of the SOTA models, we also calculate the weighted recall and weighted F_1 score measurements.

We evaluated efficiency by reporting: the number of parameters for each approach, the training time of learning, i.e., speed-up during inference, and the validation set convergence.

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

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

3.4. Experiments

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

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

3.5. SOTA models

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

Except for the Multimodal Transformer (MulT) [9], the rest of the modality fusion methods are typically RNN-based deep learning networks. However, we went beyond a typical one-to-one comparison and proposed 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 understand the SOTA field better 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 efficiently stacked RNN.

- **Early-Fusion LSTM (EF-LSTM)** [42] 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)** [42]. 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. The concatenated hidden states are 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 using 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)** [44] TFN explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures unimodal, bimodal, and trimodal interactions across linguistic, visual, and acoustic modalities.
- **Low-rank Multimodal Fusion (LMF)** [45]. 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 the multimodal tensor with a weight tensor. The output is a low-dimension vector, which is 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)** [71]. 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 did not consider proceeding utterances, we extracted the last hidden state only and fit it to a fully connected layer to make predictions.
- **Recurrent Attended Variation Embedding Network (RAVEN)** [50] 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 resulting 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 considered the last hidden state to construct nonverbal visual–acoustic embeddings since we worked on word-level aligned data.

3.5.4. Memory-based networks

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

[3] The code for our models and experiments is available on Github.

- **Memory Fusion Network** (MFN) [13] MFN is a memory fusion network that builds a multimodal gated memory component. The memory cell is updated along with the evolution of the hidden states of three unimodal LSTMs. The last memory cell is concatenated with the last hidden states of unimodal LSTMs to construct the multimodal sentence representation. Then, the multimodal representation 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 modality.

We first fine-tuned all models by performing a fifty-times random grid search on the hyperparameters. We reported the final settings in Appendix A. After the fine-tuning process, we trained all the models again for 50 epochs, five times. We used the 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 used cross-entropy loss since emotion recognition is formulated as a multilabel classification problem. We reported 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, namely, MulT, MMUU-BA, and RAVEN, exhibit the highest binary accuracy (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% for MMUU-BA and 33.6% for MulT. TFN attained the highest accuracy of 34.9% for Acc_7. Raven and MMUU-BA report the highest correlation ($Corr$). Despite the low accuracy, MCTN exhibits the lowest mean absolute error. That might imply that MCTN needs more epochs to converge (we found in [14] that MCTN had been trained for 200 epochs). Overall, RAVEN was the most effective approach on MOSI task. 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 [50]. A possible reason for the discrepancy between literature and our empirical results may be that different versions of the MOSI dataset had been used in the published works. Those versions consist 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 tricky task. In the literature, MulT is regarded as the SOTA approach among the 11 investigated approaches, reporting an increased binary accuracy of 83.0% as compared to 78.7% in our experiments on MOSI. Note that for MulT we used 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 attained an improved performance compared to that of the MOSI dataset. We suspect this is because MOSEI is a much larger dataset. MMUU-BA attained an increased binary accuracy of 80.7% compared to 80.2% for 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, namely, 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.

Table 1
Comparative analysis across the SOTA approaches on MOSI.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
EF-LSTM [42]	32.7	75.8	75.6	1.000	0.630
LF-LSTM [42]	32.7	76.2	76.2	0.987	0.624
RMFN [10]	32.3	76.8	76.4	0.980	0.626
TFN [44]	**34.9**	75.6	75.5	1.009	0.605
LMF [45]	30.5	75.3	75.2	1.018	0.605
MARN [71]	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 [50]	**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

Table 2
Comparative analysis across the SOTA approaches on MOSEI.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
EF-LSTM [42]	45.7	78.2	77.1	0.687	0.573
LF-LSTM [42]	47.1	79.2	78.5	0.655	0.614
TFN [44]	47.3	79.3	78.2	0.657	0.618
LMF [45]	47.6	78.2	77.6	0.660	0.623
MARN [71]	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 [50]	47.8	**80.2**	79.8	0.636	0.654
MFN [13]	47.4	79.9	79.1	0.646	0.626

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 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 [50], authors did not conduct experiments on MOSEI. Yet, in [9], for RAVEN, authors reported a decreased accuracy of 79.1% compared to our 80.2% (see Table 2). We could not run experiments for RMFN or MCTN on MOSEI. RMFN was computationally too expensive, and MCTN could not support MOSEI.

Following previous work [77], the binary performance across different modality fusion approaches was compared for the MOSI and MOSEI tasks, as shown in Fig. 1. Each line style corresponds to the taxonomy of the SOTA approaches. According to Fig. 1, all approaches improve on the MOSEI task. Besides, MulT and Raven yield similar performance for both MOSI and MOSEI tasks. That is, they show similar learning behaviour. However, MMUU-BA shows a positive trend with a sharper rise in performance for the MOSEI task than the MulT and RAVEN approaches.

We present the results for the emotion recognition task in Table 3. In contrast to sentiment analysis tasks, which calculate accuracy, we calculated the class-wise recall to find out how many emotions were detected correctly out of the total number of emotions for each emotion class. We also calculated the weighed recall for each modality fusion method. The results show that the happy emotion class is the most challenging for all approaches, while the angry class is the most straightforward. Attention mechanism approaches, e.g., MulT and

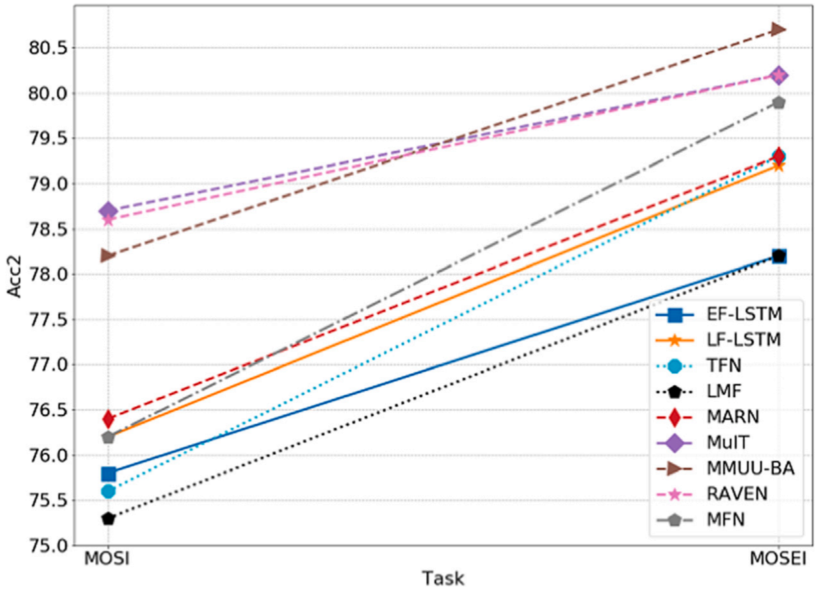

Fig. 1. Accuracy comparison across different modality fusion approaches for MOSI and MOSEI tasks.

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

Approach	Neutral		Happy		Sad		Angry		Weighted	
	Recall	*F1*	*Recall*	*F1*	*Recall*	*F1*	*Recall*	*F1*	*Recall*	*F1*
EF-LSTM [42]	57.3	61.2	20.7	30.8	57.7	**62.0**	**80.7**	71.7	57.8	59.5
LF-LSTM [42]	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 [44]	60.0	61.9	19.3	28.0	53.4	57.3	76.4	**72.9**	56.7	58.7
LMF [45]	46.6	54.7	34.5	40.6	49.8	54.3	80.1	**72.9**	53.6	57.0
MARN [71]	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 [50]	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

MMUU-BA, are the most effective for the emotion recognition task. In particular, MMUU-BA achieves the highest recall for happy and sad classes, while MulT recalls the most neutral utterances (see Table 3). However, 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 cannot directly compare our results with those in literature since binary accuracy is used as a prime performance measurement. However, in [9], MulT is also the SOTA for the IEMOCAP task.

Overall, we see that all approaches attained a lower binary performance compared to the reported ones in literature, except RAVEN, which achieved a higher performance on both MOSEI and MOSI, and MMUU-BA, which achieved 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 did not attain the highest performances on sentiment analysis tasks. Yet, without considering efficiency, we noticed that MulT, MMUU-BA, and RAVEN are the most appropriate models for sentiment analysis, while MMUU-BA and MulT are the most appropriate ones for emotion recognition. While RAVEN shows outstanding performance for the sentiment analysis tasks, it yields the lowest performance for the emotion recognition task.

Error analysis. We conducted an error analysis on the above experiments. Fig. 2 shows the percent error[4] per sentiment class on MOSI. Each line style corresponds to the taxonomy of the SOTA approaches. Despite the fact that 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 Fig. 2). In particular, most approaches show a percent error that is twice as high 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. However, 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.

Fig. 3 depicts the percent error per sentiment class on MOSEI. In contrast to MOSI, all approaches achieve a low percent error for the positive sentiment class, whereas they struggle with negative utterances. We suspect this is because MOSEI is an unbalanced dataset. That is, it consists of 11 544 positive and 4721 negative utterances. The results show that once we collected enough data, there was no

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

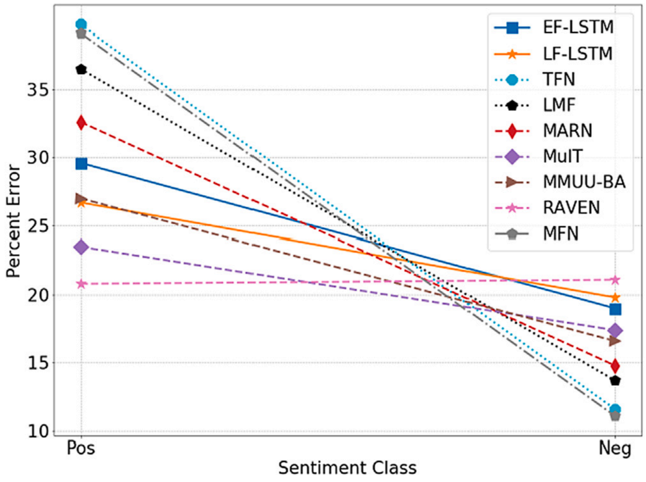

Fig. 2. Percent error per sentiment class on MOSI.

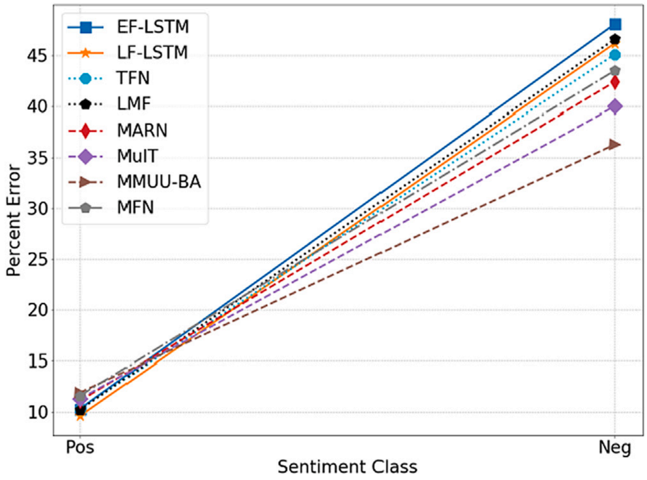

Fig. 3. Percent error per sentiment class on MOSEI.

significant difference among different fusion modality approaches in terms of performance (see the positive class in Fig. 3).

Fig. 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 in the neutral, sad, and angry emotion classes, respectively, in the test set. That implies that the performance for each emotion class is analogous to the number of samples for each class. However, some approaches, such as 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 then carried out the following analysis on test outputs of MOSI. We grouped the outputs of all the samples in the test dataset. The first group (i.e., easy) consists of 49 cases, where all methods predict correctly; the second group (i.e., medium) consists of 21 cases, where half the methods predict correctly; the third (i.e., hard) consists of 18 cases, where 2 out of 11 methods predict correctly; and the fourth (i.e., very hard) consists of 15 cases, where all methods predict incorrectly. We included four samples for each group in Table 4.

Out of 686 utterances, 49 of them, that were 7.1%, are predicted correctly by all approaches. These are usually sentences consisting of highly sentimental words such as “horrible”, “love” (see Table 4, Easy category). Only 21 utterances, 3.1%, were 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 such as “*But it does have some adult humour*” and “*It actually surprised me*” (see Table 4, Medium category) is due to missing content. Eighteen utterances, i.e., 2.6%, could not be correctly predicted by 9 out of 11 approaches, even though utterances include highly sentimental words like “pretty girl”, “laughing” (see Table 4, Hard category). Finally, no approaches could 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 dominated by highly negative words, but the overall sentiment is positive. It is worth mentioning that all the error cases of the medium, hard, and very hard groups are positive sentiment utterances. To our knowledge, this is a novel finding.

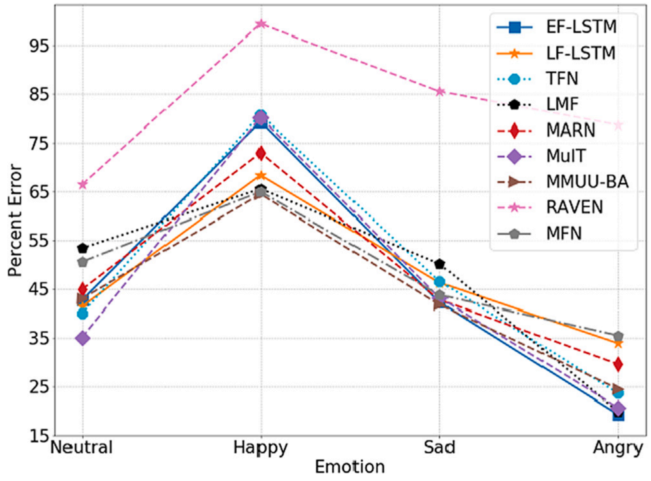

Fig. 4. Percentage error per emotion class on IEMOCAP.

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.

4.2. Efficiency

In experiment 2, we reported the model sizes (i.e., parameters), the training time of learning, and the validation set convergence. We illustrated the validation set convergence across all competitive approaches on MOSI, MOSEI and IEMOCAP in Figs. 5, 6, and 7, respectively. We noticed that all approaches converge in just a few epochs for

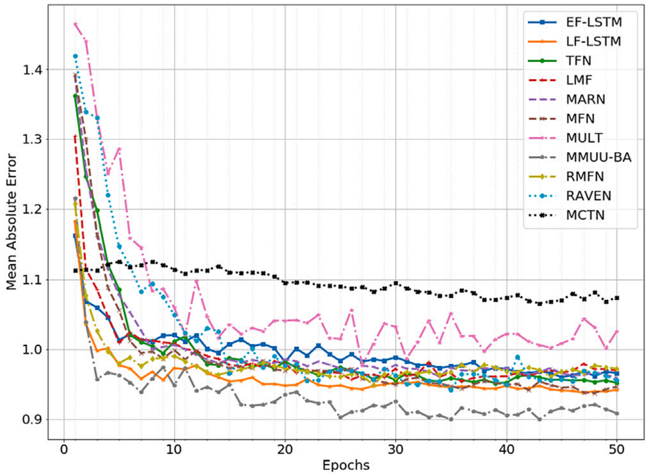

Fig. 5. Validation set convergence across the SOTA approaches on the MOSI task.

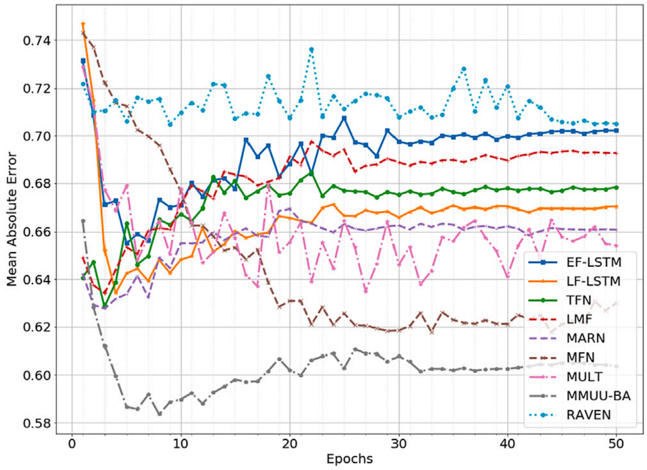

Fig. 6. Validation set convergence across the SOTA approaches on the MOSEI task.

all tasks, i.e., CMU-MOSI, CMU-MOSEI, and IEMOCAP tasks. Overall, we observed that the validation set convergence exhibits different carve trends across different fusion approaches and tasks. At first, all approaches manifest a downtrend. That implies that the learning algorithms seek to minimize the loss function, called optimization. After the optimization process, there were some approaches that the downtrend shifted to an uptrend with a sharp rise (e.g., observe LMF and EF-LSTM convergence in Fig. 6, or MFN and LF-LSTM in Fig. 7). We attribute such a sharp rise to overfitting. Indeed, some approaches are more prone to overfitting than others. Other strategies exhibit a horizontal trend (e.g., the majority of models in Fig. 5, or MULT and MFN in 6) after the optimization process. That means that the optimization algorithm is stuck in a local optimal – a good enough set of weights – or a global optimal — the best set of weights. However, for CMU-MOSI task, the horizontal trends are smooth while for CMU-MOSEI task, they usually manifest a slight negative or positive slope. We speculate that this is due to the high learning rate on CMU-MOSEI.

For MOSI, we empirically found that MMUU-BA converges faster to better results at training compared to other approaches (see Fig. 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 observed that MCTN needs much more than 50 epochs to converge.

For MOSEI, we observed that EF-LSTM, LF-LSTM, TFN, LMF, and MARN increase the MAE after 5 epochs (see Fig. 6). A possible explanation for this might be 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 Fig. 6). Finally, we empirically found 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 5th epoch (see Fig. 7). Only RAVEN and MulT attain a low and stabilized cross-entropy loss. Specifically, MulT, reporting the best recall performance for the “neutral” class, attained the lowest cross-entropy loss. EF-LSTM, achieving an improved performance as compared to other sophisticated competitive approaches, shows a fair and stabilized loss at training until 25th epoch.

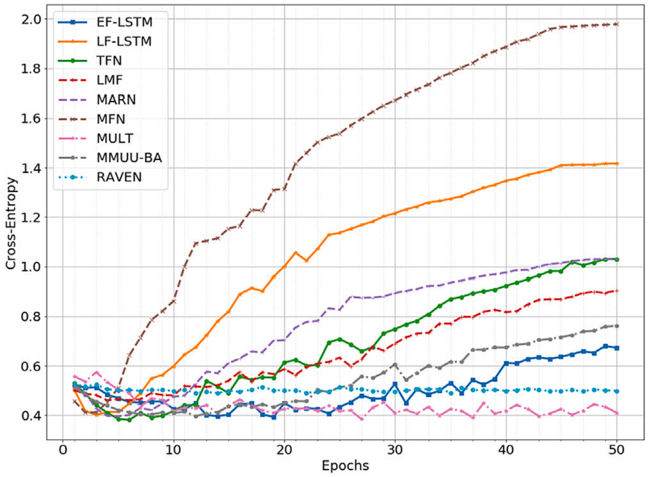

Fig. 7. Validation set convergence across the SOTA approaches on the IEMOCAP task.

Table 5
Comparative analysis across the SOTA approaches on IEMOCAP dataset.

Approach	MOSI		MOSEI		IEMOCAP	
	Mins.	Params.	Mins.	Params.	Mins.	Params.
EF-LSTM [42]	0.43	177,329	6.59	217,457	1.40	206,152
LF-LSTM [42]	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 [44]	0.51	14,707,911	1.87	6,804,859	0.53	23,198,398
LMF [45]	0.43	1,144,493	2.00	5,079,473	1.12	962,116
MARN [71]	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 [50]	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	–	–	–	–

We investigated the complexity of the models by presenting the number of parameters and training times in minutes for MOSI, MOSEI, and IEMOCAP in Table 5. We observed that approaches integrating LSTMCell components, such as LF-LSTM, MARN, and RMFN, are not able to speed up. PyTorch cannot maintain the same speed for LSTMCell, which is a variant of LSTM. Despite the low performances, 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 to the LSTMCell component of RAVEN. MulT, being a more complicated model, requires more time (i.e., 17.6 min) than MMUU-BA and RAVEN (i.e., 0.64 and 3.71 min, 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. This might be because different configuration settings were set up after the

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
$\text{TFN}_{l,v,a}$ [44]	**34.9**	75.6	75.5	1.009	0.605

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

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 [71]	31.8	76.4	76.2	0.984	0.625

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

fine-tuning process. For IEMOCAP, EF-LSTM is not only an effective but also an efficient approach, attaining a more significant speedup (26 times) than its counterpart (i.e., MulT) in terms of performance.

4.3. Ablation studies

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

4.3.1. Importance of modalities

To understand the importance of modalities in multimodal tasks, we conducted ablation studies on TFN, which inherently models unimodal, bimodal, and trimodal interactions, and MulT, which attains high accuracy on both sentiment analysis and emotion recognition tasks. For TFN, we tested the TFN approach with unimodal, bimodal, and trimodal tensors. Table 6 shows the results of the ablation studies. We observed that language is the most informative modality as it is a pivot for visual and acoustic modalities. The unimodal visual and acoustic subnetworks and the bimodal visual–acoustic subnetwork attained fairly low accuracy compared to those integrating the linguistic modality. Specifically, combining language with visual or acoustic modalities is generally better than combining the visual and acoustic modalities. In contrast to [44], 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 the tensor-based fusion is not an effective approach for modelling crossmodal interaction across three modalities.

For MulT, we first considered the performances for linguistic, visual, and acoustic only transformers. We found a binary accuracy of 79.5% for 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 studied 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], which reports that presenting language as a target modality leads to the best performance. The experiments show that there is no need to consider multiple directional pairwise crossmodal transformers. Specifically, when we considered acoustic as a target modality yielded an increased accuracy of 79.6% compared to 78.7% for MulT. However, there is no statistical difference in performance among the three crossmodal transformers and the multiple directional pairwise crossmodal transformer (i.e., MulT).

4.3.2. Important modules for crossmodal interactions

To understand the influence of individual components for modelling crossmodal interactions, we performed comprehensive ablation analysis on the SOTA approaches on MOSI. First, we studied the importance of extra dimensions with value 1 of $\text{TFN}_{l,v,a}$ [44], 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% for TFN. However, 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 considered the number of heads in the crossmodal attention module. We experimented 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. However, for Acc_7, the increased number of heads yielded an increased performance of 33.5% compared to 31.9% (see Table 7).

In [71], authors claim that for each timestamp, there might exist multiple crossmodal interactions. We experimented with three variants of MARN to investigate the number of attentions needed to extract all crossmodal dynamics. Specifically, we tried one, five, and ten attentions. In contrast to [71], 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 removed the multi-attention block (MAB) from MARN. Specifically, we replaced the MAB with a fully connected layer and removed the softmax function. We observed that there is no effect on binary accuracy (see Table 8) while for Acc_7, the difference is marginal.

For MMUU-BA, we analysed the attention module to understand its learning behaviour. We experimented with two other variants of MMUU-BA (see Table 9). The architecture of these variants differs concerning the attention computation module. Particularly, in MMUU-UA, we computed one-directional attention, e.g., from linguistic to visual modality only. In MMUU-SA, we only computed self-attention within modalities. We found that one-directional attention results in an increased binary accuracy of 78.8% compared to 78.2% from the proposed framework. Both MMUU-UA and MMUU-BA attained 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, in terms of the binary performance.

Table 10
Comparison of MFN with other variants of it.

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

Table 11
Comparison of RAVEN with other variants of it.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\mathrm{RAVEN}_{w/oShift}$	31.8	75.6	75.5	1.016	0.615
RAVEN [50]	**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	$Corr$
$\mathrm{RMFN}_{s=1}$	32.9	75.3	75.2	**0.982**	0.616
$\mathrm{RMFN}_{s=3}$	32.5	75.5	75.3	0.991	**0.623**
$\mathrm{RMFN}_{s=6}$	**33.1**	**75.6**	75.5	0.991	0.613
RMFN [10]	31.7	75.2	75.1	1.005	0.612

For MFN, first, we investigated if crossmodal interactions can happen over multiple time instances. Specifically, we experimented 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 $\mathrm{MFN}_{w/o\Delta}$ (see Table 10) significantly underperforms the MFN approach. That implies that we should not model crossmodal interactions on aligned time steps, but consider long-range crossmodal contingencies across a multimodal sequence. Second, we evaluated the importance of spatial–temporal crossmodal interactions over 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].

For RAVEN, we have already removed the Nonverbal Subnetworks [50] as mentioned in Section 3.5. This modification results in an increased binary accuracy of 78.6% compared to 78.0% in [50] on MOSI. We also investigated the temporal interactions between the nonverbal "subword" units with language utterances. Specifically, we removed the shift component, which learns dynamically to shift the text representation by integrating the nonverbal vector. 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 (see Table 11). Specifically, RAVEN shows a significantly increased binary performance of 78.6% compared to 75.6% for $\mathrm{RAVEN}_{w/oShift}$.

For RMFN, we decomposed the fusion problem into multiple stages, we experimented with the number of stages needed for modelling crossmodal dynamics. Specifically, we experimented with one, three, and six stages. Our experiments show that RMFN attained a similar performance whether we apply one or six stages to fuse information (see Table 12).

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 do not always follow 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 underperform 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 on key findings

In this paper, we replicated the most recent SOTA models for multimodal language analysis. We evaluated their effectiveness through comprehensive comparative studies, error analyses and series of ablation studies. The efficiency of the models was also compared in terms of three evaluation metrics, namely, parameters, training time, and validation set convergence. The results associated with ablation studies helped us determine which components and methodologies contribute most to solving 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 hand, despite tensor-based approaches getting 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 attained lower performances in the 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 the MOSEI and MOSI datasets used in published works could be another reason for most of those cases.

From an efficiency viewpoint, attention mechanism-based approaches are usually more complex and require more training time than the rest of the modality fusion approaches. To alleviate that issue, we could consider less fine-grained crossmodal interactions. Indeed, our ablation studies show that adding more levels of interactions across modalities results in a decreased performance. Recurrent cell-based approaches are extremely computationally expensive. On the other hand, memory and tensor networks are more efficient.

Table 13 summarizes the key findings on how different components contribute to solving the problem of affective video content analysis. Overall, the results demonstrate that attention mechanism are the most effective approaches despite being computationally expensive. During the training process, they manifest a stabilized and fast convergence, and they cope with both skewed and balanced datasets. However, autoencoder approaches are more suitable if we work with missing or noisy data. The ablation studies show that crossmodal interactions are not aligned on corresponding time steps but spread across a multimodal sequence. Finally, video sentiment analysis could benefit from the integration of context. However, all approaches struggle with positive sentiment utterances.

These key findings are drawn from experiments over three most widely used standard benchmark datasets in the literature, and data imbalance has been regarded as a vital issue influencing the model performance. The linguistic modality is the most informative compared to visual and acoustic modalities. We attribute that difference to the use of word embedding trained on large corpora, and not to noise issues related to the datasets. All three datasets were carefully collected, pre-processed and annotated by a world-leading group in this area, and the noise within the datasets is minimized. Indeed, there is a need for investigating new approaches for training visual and acoustic embeddings. However, such an investigation is beyond the scope of this paper. Thus, we believe that our results over three high-quality and well-established large-scale benchmark datasets can sufficiently support the conclusions.

In the future, it would be worth investigating how multimodal sentiment analysis could benefit from considering proceeding utterances and existing knowledge bases, which might entail sentiment or emotional knowledge. Little effort has also been devoted towards crossmodal interactions across a multimodal sequence instead of corresponding timestamps. One limitation of our study is that we used a simple approach to align modalities. Following previous work, we

Table 13
Summary of key findings. The first column lists the investigated key components, the second column summarizes which models are using which component, and the third column shows how different components contribute differently to solving the problem of multimodal language analysis.

Component	Model	Contribution
Basic recurrent structures	EF-LSTM [42], LF-LSTM [42]	(1) Computationally cheap. (2) Outperform a few SOTA approaches.
Tensor operator	TFN [44], LMF [45]	(1) Low error for the negative class on MOSI. (2) Computationally cheap.
Attention mechanism	RMFN [10], MARN [71], MulT [9], MMUU-BA [10], RAVEN [50]	(1) State-of-the-art performance on both tasks. (2) Relatively fast convergence. (3) Stabilized learning behaviour. (4) Cope with skewed and balanced datasets.
Memory cell	MFN [13]	Capture non-aligned crossmodal interactions.
Autoencoder	MCTN [14]	(1) Tackle with perturbations and missing data. (2) Fewer learning parameters.

averaged visual and acoustic modalities throughout word intervals since advancing the SOTA was not the aim of this work. Yet, further investigation is needed in this direction to determine if other alignment approaches could enhance the relatively poor performance of the non-verbal modalities. In terms of the implementation, we noticed that the LSTMCell component could not speed up. That made approaches which primarily utilize recurrent cell components less efficient.

6. Conclusions

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 determined 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 that are able to capture crossmodal interactions across different timestamps, integrate context, and utilize linguistic modality as a pivot for the non-verbal modalities 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. We expect that the findings would provide helpful insights to the development of more effective modality fusion models. In the future, we are going to focus on conversational video sentiment analysis tasks because the utterance context has proven to be beneficial for understanding human language.

CRediT authorship contribution statement

Dimitris Gkoumas: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Qiuchi Li:** Software. **Christina Lioma:** Writing - review & editing, Supervision. **Yijun Yu:** Writing - review & editing, Supervision. **Dawei Song:** Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study is supported by the Quantum Information Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321, and Natural Science Foundation of China (grant No.: U1636203).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.inffus.2020.09.005.

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: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 3794–3804, http://dx.doi.org/10.18653/v1/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 - 14th International Conference, ITS 2018, Montreal, QC, Canada, June 11-15, 2018, Proceedings, in: Lecture Notes in Computer Science, vol. 10858, Springer, 2018, pp. 150–159, http://dx.doi.org/10.1007/978-3-319-91464-0_15.

[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, in: CHI EA '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–8, http://dx.doi.org/10.1145/3290607.3299040.

[4] S.S. Rajagopalan, L. Morency, T. Baltrusaitis, 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 - 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VII, in: Lecture Notes in Computer Science, vol. 9911, Springer, 2016, pp. 338–353, http://dx.doi.org/10.1007/978-3-319-46478-7_21.

[5] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z. Zhou, X. Wu (Eds.), IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, IEEE Computer Society, 2016, pp. 439–448, http://dx.doi.org/10.1109/ICDM.2016.0055.

[6] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L. Morency, Context-dependent sentiment analysis in user-generated videos, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 873–883, http://dx.doi.org/10.18653/v1/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, Hong Kong, China, July 10-14, 2017, IEEE Computer Society, 2017, pp. 949–954, http://dx.doi.org/10.1109/ICME.2017.8019301.

[8] A. Zadeh, R. Zellers, E. Pincus, L. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88, http://dx.doi.org/10.1109/MIS.2016.94.

[9] Y.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 6558–6569, http://dx.doi.org/10.18653/v1/p19-1656.

[10] D. Ghosal, M.S. Akhtar, D.S. Chauhan, S. Poria, A. Ekbal, P. Bhattacharyya, Contextual inter-modal attention for multi-modal sentiment analysis, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 3454–3466, http://dx.doi.org/10.18653/v1/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: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2225–2235, http://dx.doi.org/10.18653/v1/P18-1207, URL https://www.aclweb.org/anthology/P18-1207/.

[12] P.P. Liang, Z. Liu, A. Zadeh, L. Morency, Multimodal language analysis with recurrent multistage fusion, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 150–161, http://dx.doi.org/10.18653/v1/d18-1014.

[13] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L. Morency, Memory fusion network for multi-view sequential learning, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5634–5641, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17341.

[14] H. Pham, P.P. Liang, T. Manzini, L. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 6892–6899, http://dx.doi.org/10.1609/aaai.v33i01.33016892.

[15] S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 164–172, URL https://aaai.org/ojs/index.php/AAAI/article/view/5347.

[16] P.K. Atrey, M.A. Hossain, A. El-Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Syst. 16 (6) (2010) 345–379, http://dx.doi.org/10.1007/s00530-010-0182-0.

[17] S. Sun, A survey of multi-view machine learning, Neural Comput. Appl. 23 (7–8) (2013) 2031–2038, http://dx.doi.org/10.1007/s00521-013-1362-6.

[18] D. Ramachandram, G.W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108, http://dx.doi.org/10.1109/MSP.2017.2738401.

[19] T. Baltrusaitis, C. Ahuja, L. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[20] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125, http://dx.doi.org/10.1016/j.inffus.2017.02.003.

[21] A. Zadeh, R. Zellers, E. Pincus, L. Morency, MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, arXiv:1606.06259. URL http://arxiv.org/abs/1606.06259.

[22] A. Zadeh, P.P. Liang, S. Poria, E. Cambria, L. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2236–2246, http://dx.doi.org/10.18653/v1/P18-1208, URL https://www.aclweb.org/anthology/P18-1208/.

[23] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359, http://dx.doi.org/10.1007/s10579-008-9076-6.

[24] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 2625–2634, http://dx.doi.org/10.1109/CVPR.2015.7298878.

[25] L. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, in: H. Bourlard, T.S. Huang, E. Vidal, D. Gatica-Perez, L. Morency, N. Sebe (Eds.), Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011, ACM, 2011, pp. 169–176, http://dx.doi.org/10.1145/2070481.2070509.

[26] S. Ghosh, E. Laksana, L. Morency, S. Scherer, Representation learning for speech emotion recognition, in: N. Morgan (Ed.), Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016, ISCA, 2016, pp. 3603–3607, http://dx.doi.org/10.21437/Interspeech.2016-692.

[27] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: Visual question answering, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 2425–2433, http://dx.doi.org/10.1109/ICCV.2015.279.

[28] A.P. James, B.V. Dasarathy, Medical image fusion: A survey of the state of the art, Inf. Fusion 19 (2014) 4–19, http://dx.doi.org/10.1016/j.inffus.2013.12.002.

[29] M.U. Bokhari, F. Hasan, Multimodal information retrieval: Challenges and future trends, Int. J. Comput. Appl. 74 (14) (2013) 9–12.

[30] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 527–536, http://dx.doi.org/10.18653/v1/p19-1050.

[31] F. Noroozi, C. Corneanu, D. Kamiska, T. Sapiski, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, IEEE Trans. Affect. Comput. PP (2018) 1–20, http://dx.doi.org/10.1109/TAFFC.2018.2874986.

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

[33] E. Morvant, A. Habrard, S. Ayache, Majority vote of diverse classifiers for late fusion, in: P. Fränti, G. Brown, M. Loog, F. Escolano, M. Pelillo (Eds.), Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings, in: Lecture Notes in Computer Science, vol. 8621, Springer, 2014, pp. 153–162, http://dx.doi.org/10.1007/978-3-662-44415-3_16.

[34] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: K. Knight, A. Nenkova, O. Rambow (Eds.), NAACL HLT 2016, the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, The Association for Computational Linguistics, 2016, pp. 160–170, http://dx.doi.org/10.18653/v1/n16-1020.

[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.K. D'Mello, A.C. Graesser, B.W. Schuller, J. Martin (Eds.), Affective Computing and Intelligent Interaction - Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 6975, Springer, 2011, pp. 359–368, http://dx.doi.org/10.1007/978-3-642-24571-8_47.

[36] E. Cambria, N. Howard, J.Y. 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, Singapore, April 16-19, 2013, IEEE, 2013, pp. 108–117, http://dx.doi.org/10.1109/CIHLI.2013.6613272.

[37] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: E. Lank, A. Vinciarelli, E.E. Hoggan, S. Subramanian, S.A. Brewster (Eds.), Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13 - 17, 2017, ACM, 2017, pp. 569–576, http://dx.doi.org/10.1145/3136755.3143011.

[38] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, L. Morency, Deep multimodal fusion for persuasiveness prediction, in: Y.I. Nakano, E. André, T. Nishida, L. Morency, C. Busso, C. Pelachaud (Eds.), Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016, ACM, 2016, pp. 284–288, http://dx.doi.org/10.1145/2993148.2993176.

[39] A. Lazaridou, N.T. Pham, M. Baroni, Combining language and vision with a multimodal skip-gram model, in: R. Mihalcea, J.Y. Chai, A. Sarkar (Eds.), NAACL HLT 2015, the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, The Association for Computational Linguistics, 2015, pp. 153–163, http://dx.doi.org/10.3115/v1/n15-1016.

[40] R. Kiros, R. Salakhutdinov, R.S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, 2014, CoRR abs/1411.2539. arXiv:1411.2539. URL http://arxiv.org/abs/1411.2539.

[41] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, S. Khudanpur, Recurrent neural network based language model, in: T. Kobayashi, K. Hirose, S. Nakamura (Eds.), INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, ISCA, 2010, pp. 1045–1048, URL http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html.

[42] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[43] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing

Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a Meeting Held December 3-6, 2012, Lake Tahoe, Nevada, United States, 2012, pp. 1106–1114.

[44] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics, 2017, pp. 1103–1114, http://dx.doi.org/10.18653/v1/d17-1115.

[45] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2247–2256, http://dx.doi.org/10.18653/v1/P18-1209, URL https://www.aclweb.org/anthology/P18-1209/.

[46] 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, http://dx.doi.org/10.18653/v1/w19-4331.

[47] P.P. Liang, Z. Liu, Y.H. Tsai, Q. Zhao, R. Salakhutdinov, L. Morency, Learning representations from imperfect time series data via tensor rank regularization, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 1569–1576, http://dx.doi.org/10.18653/v1/p19-1152.

[48] S. Mai, H. Hu, S. Xing, Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 481–492, http://dx.doi.org/10.18653/v1/p19-1046.

[49] S. Mai, S. Xing, H. Hu, Locally confined modality fusion network with a global perspective for multimodal human affective computing, IEEE Trans. Multimedia 22 (1) (2020) 122–137, http://dx.doi.org/10.1109/TMM.2019.2925966.

[50] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 7216–7223, http://dx.doi.org/10.1609/aaai.v33i01.33017216.

[51] R. Beard, R. Das, R.W.M. Ng, P.G.K. Gopalakrishnan, L. Eerens, P. Swietojanski, O. Miksik, Multi-modal sequence fusion via recursive attention for emotion recognition, in: A. Korhonen, I. Titov (Eds.), Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018, Association for Computational Linguistics, 2018, pp. 251–259, http://dx.doi.org/10.18653/v1/k18-1025.

[52] S.H. Dumpala, I. Sheikh, R. Chakraborty, S.K. Kopparapu, Audio-visual fusion for sentiment classification using cross-modal autoencoder, in: Visually Grounded Interaction and Language (ViGIL), Vol. NIPS 2018, 2019.

[53] Z. Wang, Z. Wan, X. Wan, Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis, in: Y. Huang, I. King, T. Liu, M. van Steen (Eds.), WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, ACM / IW3C2, 2020, pp. 2514–2520, http://dx.doi.org/10.1145/3366423.3380000.

[54] M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, L. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: E. Lank, A. Vinciarelli, E.E. Hoggan, S. Subramanian, S.A. Brewster (Eds.), Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13 - 17, 2017, ACM, 2017, pp. 163–171, http://dx.doi.org/10.1145/3136755.3136801.

[55] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270, http://dx.doi.org/10.1016/j.patrec.2019.04.024.

[56] Y. Zhang, Z. Wang, J. Du, Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition, in: International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, IEEE, 2019, pp. 1–8, http://dx.doi.org/10.1109/IJCNN.2019.8851942.

[57] Z. Sun, P.K. Sarma, W.A. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 8992–8999, URL https://aaai.org/ojs/index.php/AAAI/article/view/6431.

[58] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, I. Marsic, Multimodal affective analysis using hierarchical attention strategy with word-level alignment, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2225–2235, http://dx.doi.org/10.18653/v1/P18-1207, URL https://www.aclweb.org/anthology/P18-1207/.

[59] E. Georgiou, C. Papaioannou, A. Potamianos, Deep hierarchical fusion with application in sentiment analysis, in: G. Kubin, Z. Kacic (Eds.), Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, ISCA, 2019, pp. 1646–1650, http://dx.doi.org/10.21437/Interspeech.2019-3243.

[60] P.P. Liang, Y.C. Lim, Y.H. Tsai, R. Salakhutdinov, L. Morency, Strong and simple baselines for multimodal utterance embeddings, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 2599–2609, http://dx.doi.org/10.18653/v1/n19-1267.

[61] A. Zadeh, C. Mao, K. Shi, Y. Zhang, P.P. Liang, S. Poria, L.-P. Morency, Factorized multimodal transformer for multimodal sequential learning, 2019, arXiv preprint arXiv:1911.09826.

[62] Y.-H.H. Tsai, M.Q. Ma, M. Yang, R. Salakhutdinov, L.-P. Morency, Interpretable multimodal routing for human multimodal language, 2020, arXiv preprint arXiv:2004.14198.

[63] D. Hazarika, R. Zimmermann, S. Poria, MISA: Modality-invariant and-specific representations for multimodal sentiment analysis, 2020, arXiv preprint arXiv:2005.03545.

[64] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A.F. Gelbukh, E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 6818–6825, http://dx.doi.org/10.1609/aaai.v33i01.33016818.

[65] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: M.A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2122–2132, http://dx.doi.org/10.18653/v1/n18-1193.

[66] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: interactive conversational memory network for multimodal emotion detection, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 2594–2604, http://dx.doi.org/10.18653/v1/d18-1280.

[67] T. Young, V. Pandelea, S. Poria, E. Cambria, Dialogue systems with audio context, Neurocomputing 388 (2020) 102–109, http://dx.doi.org/10.1016/j.neucom.2019.12.126.

[68] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A.F. Gelbukh, DialogueGCN: A graph convolutional neural network for emotion recognition in conversation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 154–164, http://dx.doi.org/10.18653/v1/D19-1015.

[69] Y. Zhang, Q. Li, D. Song, P. Zhang, P. Wang, Quantum-inspired interactive networks for conversational sentiment analysis, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 5436–5442, http://dx.doi.org/10.24963/ijcai.2019/755.

[70] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107, http://dx.doi.org/10.1109/MIS.2016.31.

[71] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L. Morency, Multi-attention recurrent network for human communication comprehension, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5642–5649, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17390.

[72] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/d14-1162.

[73] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/n19-1423.

[74] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M.A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2227–2237, http://dx.doi.org/10.18653/v1/n18-1202.

[75] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP - a collaborative voice analysis repository for speech technologies, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, IEEE, 2014, pp. 960–964, http://dx.doi.org/10.1109/ICASSP.2014.6853739.

[76] J. Yuan, M. Liberman, Speaker identification on the SCOTUS corpus, J. Acoust. Soc. Am. 123 (5) (2008) 3878.

[77] Z. Liu, Q. Xie, M. Wu, W. Cao, D. Li, S. Li, Electroencephalogram emotion recognition based on empirical mode decomposition and optimal feature selection, IEEE Trans. Cogn. Dev. Syst. 11 (4) (2019) 517–526, http://dx.doi.org/10.1109/TCDS.2018.2868121.