Word-Aware Modality Stimulation for Multimodal Fusion

Shuhei Tateishi[1], Makoto Nakatsuji[2], Yasuhito Ohsugi[1]
NTT Docomo, Inc., NTT Research Institute
syuuhei.tateishi.tc@nttdocomo.com, makoto.nakatsuji@ntt.com, yasuhito.oosugi.fg@nttdocomo.com

Abstract

Multimodal learning is generally expected to make more accurate predictions than text-only analysis. Here, although various methods for fusing multimodal inputs have been proposed for sentiment analysis tasks, we found that they may be inhibiting their fusion methods, which are based on attention-based language models, from learning non-verbal modalities, because non-verbal ones are isolated from the linguistic semantics and contexts and do not include them, meaning that they are unsuitable for applying attention to text modalities during the fusion phase. To address this issue, we propose Word-aware Modality Stimulation Fusion (WA-MSF) for facilitating integration of non-verbal modalities with the text modality. The Modality Stimulation Unit layer (MSU-layer) is the core concept of WA-MSF; it integrates language contexts and semantics into non-verbal modalities, thereby instilling linguistic essence into these modalities. Moreover, WA-MSF uses aMLP in the fusion phase in order to utilize spatial and temporal representations of non-verbal modalities more effectively than transformer fusion. In our experiments, WA-MSF set a new state-of-the-art level of performance on sentiment prediction tasks.

Keywords: Multimodal, Sentiment Analysis

1. Introduction

In the field of machine learning, it is thought that multiple source inputs improve the accuracy of the output relative to that of a single one. This belief comes by analogy to the cognitive function of human beings, where multiple senses help us to make more accurate decisions. The corresponding methodologies are called multimodal machine learning (or multimodal fusion), because they utilize multiple sources as input. They have been incorporated in real-world services, e.g. to enable chatbots to provide more suitable responses that are aligned with the user's emotions or sentiments.

Sentiment analysis is one of the main streams of studies on multimodal fusion. "Multimodal" here refers to aspects of human expression in communication, such as what words are said (linguistic), what tone is used (auditory), and what gestures or behaviors are displayed (visual). Several methods have been proposed for fusing these modalities to improve the accuracy of sentiment prediction, for example, by performing tensor multiplication of language, audio, and visual modality feature vectors (Zadeh et al., 2017).

The attention-based language model, exemplified by BERT (Devlin et al., 2019) in self-supervised-learning language modeling, has achieved dramatically higher accuracy scores compared with previous multimodal models. Since its advent, multimodal method research has focused on how to utilize attention-based language models by applying it to language modality vectorization as well as other non-verbal modality vectorizations so as to achieve accuracies greater than that of a single-language-modality prediction. For instance, Yang et al. (2020) demonstrated a methodology where language features from BERT and audio features from COVAREP (Degottex et al., 2014) are initially combined using source-target attention. Subsequently, the resulting combined output is incorporated in BERT's self-attention mechanism for sentiment prediction.

We have conducted a repeatability test to compare the performance of single-language-modality prediction model (i.e. BERT and its family) with that of modern transformer-based multimodal sentiment analysis models (Tsai et al., 2019; Yang et al., 2020; Rahman et al., 2020; Arjmand et al., 2021; Hu et al., 2022; Han et al., 2021; Guo et al., 2022). In particular, although the designers of those models claim that their methods outperform BERT and its family, we found that *almost all of the models failed to score higher than BERT* (See the Evaluation section). Although some of the newer models scored higher than BERT on some of the metrics of the test, none of the models exceeded the accuracy of BERT on all of the evaluation metrics (e.g. Acc^7, MAE).

We think those results are due to differences in the evaluation procedures; the previous studies conducted only a few (e.g. up to 5) iterations of BERT learning as a benchmark, whereas we found many iterations may be needed for it to achieve much higher accuracy scores in sentiment analysis tasks. In this context, we suspect that non-verbal modalities may actually hinder the learning of attention-based language models in existing multimodal learning approaches. This is because the non-verbal modalities, such as acoustic and visual, are usually temporal acoustic or visual embedding sequences and do not have language essences. Thus, they do not fit well the attention-based lan-

LREC-COLING 2024, pages 17664–17674
20-25 May, 2024.

guage models that are invented for understanding languages. As a result, audio/visual embedding streams tend to be noise when current transformer-based multi-modal fusion models fuse them with the language modality. If this hypothesis is true, it is a severe problem for multimodal fusion tasks that must be resolved.

To cope with this problem, we devised a new multimodal fusion method called Word-aware Modality Stimulation Fusion (WA-MSF). The main idea of this method is *the introduction of a new layer, named the Modality Stimulation Unit layer (MSU-layer)*. Within this layer, word semantics and contextual positions are incorporated into non-verbal modalities, facilitating infusion of semantic and contextual details from the textual modality before fusion. The MSU-layer calibrates nonverbal modalities within the self-attention structure based on the textual modality and accomplishes seamless integration of textual and non-textual modalities during their encoding by BERT or its variants. As a result, these modalities stimulated by language embeddings can be integrated more effectively with textual modalities.

Our method also contains an another aspect. That is *the approach to multimodal fusion*. The above MSU-layer establishes affinity between the textual modality and the nonverbal modality. However, in the context of multimodal fusion, it is necessary to adopt strategies that facilitate incorporation of spatial and temporal representations from the non-linguistic modality. With this perspective in mind, we incorporated aMLP in our fusion process. aMLP is a variant of gMLP (Liu et al., 2021), which has its strength in the extraction and analysis of spatial representations in the nonverbal modality and includes a small-scale attention mechanism for language information affinity.

We found that our method significantly improved the accuracy of sentiment analysis compared with that of BERT and other state-of-the-art methods.

This paper's main contributions are:

1. We raise a concern regarding BERT and other recent transformer language models for multimodal sentiment analysis wherein they face difficulties when integrating features from non-verbal modalities, because of their limited grasp of language semantics and context within diverse modality inputs.

2. Our proposal introduces the MSU-layer, which enhances multimodal fusion by incorporating language context into non-verbal modalities as a pre-fusion step. Furthermore, we employ aMLP as the fusion method after the MSU-layer to efficiently manage spatial and temporal representations within nonverbal modalities.

3. WA-MSF achieved state-of-the-art (SOTA) levels of performance on the multimodal sentiment analysis task, with top-1 scores in Mean Absolute Error (MAE) and correlation coefficient (Corr) on the CMU-MOSI and CMU-MOSEI datasets [1,2].

2. Related Work

Deep-learning techniques have been used to acquire high-level multimodal features. Bidirectional LSTMs have been employed to capture long-range dependencies from low-level acoustic descriptors and visual features (Eyben et al. (2010), Wöllmer et al. (2010)). Additionally, CNNs have been used to extract both textual and visual features (Poria et al., 2015). More recently, advanced models have emerged that learn representations of human multimodal language. For example, Dumpala et al. (2019) explored the use of cross-modal autoencoders for audio-visual alignment, and the tensor fusion network (TFN) (Zadeh et al., 2017) calculates tensor products across input modalities (language, audio, and video) with concatenated scalar values in order to represent associations between individual and combined modalities.

Several methods of multimodal fusion using transformers have been proposed in the past. The multimodal transformer (MulT) (Tsai et al., 2019) uses transformer layers to fuse multimodal embedding streams, but does not use them to encode language modality embeddings. Low-rank fusion network (LFN) (Sahay et al., 2020) is an ideological successor of (Zadeh et al., 2017) that uses low-rank fusion to reduce the learning parameters. MulT and LFN also aim to fuse multimodal features without aligning non-verbal feature lengths with the verbal one. Cross-modal BERT (CM-BERT) (Yang et al., 2020) encodes the language modality by using BERT and the speech modality by using COVAREP (Degottex et al., 2014). It then applies source-target attention from the language embedding sequence to the speech embedding. MAG-BERT (and its variant MAG-XLNet) (Rahman et al., 2020) uses a multimodal adaptation gate (MAG), a procedure for melding non-verbal modality information into the verbal information among the transformer layers.

More recently, several methods have achieved prominent positions on the sentiment analysis leaderboards for MOSI and MOSEI datasets. Cross hyper-modality fusion network (CHFN, Guo et al. (2022)) utilizes a "multimodal interaction layer"

[1] As of Feburary 2024, our method mark top-1 scores in CMU-MOSI leaderboard (https://paperswithcode.com/sota/multimodal-sentiment-analysis-on-cmu-mosi) and CMU-MOSEI leaderboard (https://paperswithcode.com/sota/multimodal-sentiment-analysis-on-cmu-mosei-1).

[2] Our code will be made available to support future research studies on multimodal fusion.

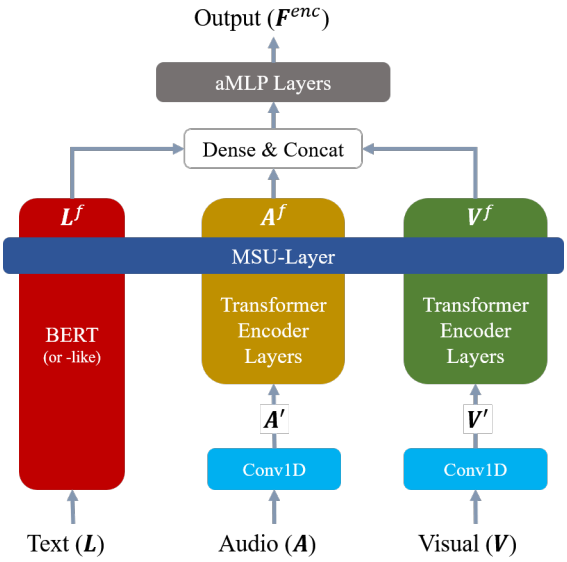

Figure 1: Overview of WA-MSF.

to facilitate alignment among non-verbal modalities and enhance the verbal modality. TEASEL (Arjmand et al., 2021) introduces the speech modality as a dynamic prefix alongside the textual modality, in contrast to conventional language models like RoBERTa (Liu et al., 2019). SPECTRA (Yu et al., 2023) extends the text-based dialog pre-training (utilizing RoBERTa) of the response selection task to speech-text dialog pre-training scenarios and is thereby able to adapt to a broader array of speech-text tasks. MMML (Wu et al., 2023) emphasizes designing fusion techniques aligned with dataset annotation schemas and employs RoBERTa as the text encoder. MultiModalInfoMax (Han et al., 2021) utilizes mutual information (MI) in its training to increase the effectiveness of multimodal fusion by incorporating only information that is highly interrelated between modalities. Finally, UniMSE (Hu et al., 2022) simultaneously learns sentiment and emotion in multimodal utterances using T5 (Raffel et al., 2020).

Here, we introduce a fresh perspective that sets us apart from the above-mentioned methods: pre-learning and setting of language modality information into the non-verbal modality before fusion. This approach aims to facilitate subsequent fusion tasks. Furthermore, we use aMLP as a fusion method known for its strong non-verbal affinity. This combination underlies the superior accuracy of our method compared with existing approaches.

3. Methodology

This section explains our method, WA-MSF.

3.1. Overview

WA-MSF consists of three phases:

1. Apply transformer encoder layers (e.g. BERT or transformer-based) independently to individual modalities (the red, yellow and green boxes in Fig. 1).

2. For each modality, insert an MSU-layer just after a certain number of encoder layers are passed (the navy-blue box in Fig. 1).

3. Concatenate the final encoder outputs of all modalities and input this concatenated result to aMLP for fusion (gray box in Fig. 1).

Each phase is explained below.

3.2. Phase 1: Modality Encoding

This section explains the encoder process of features in each modality through the encoder layers.

3.2.1. Feature Extractions for Individual Modalities

For the language modality, BERT is used for tokenization and featurization. The data source is simply the plain text of the utterances. The text is then tokenized into word IDs and embedded into a feature vector sequence by BERT.

For the audio modality, we utilize feature extraction methods such as COVAREP (Degottex et al., 2014) or wav2vec (Baevski et al., 2020). Notably, wav2vec, being a transformer-based model, seamlessly integrates with our approach.

For the video modality, a facial action unit (AU) extraction methodology such as OpenFace (Baltrusaitis et al., 2016) is used for featurization. AU is a numeric representation of human facial expressions, which is suitable for analyzing sentiment from the visual modality.

3.2.2. Word Alignment for Nonverbal Modalities

As we mentioned in the Introduction, our model uses aMLP in the subsequent processing. Because of it, at this stage, each modality serving as input needs to be uniform token length for the later concatenation process. Nevertheless, the audio and visual modalities undergo unique time slicing during their featurization; given the varying sequence sizes between these modalities and the language modality, direct fusion is not appropriate. To address this issue, we undertake a preprocessing step prior to fusion. This step involves transforming the audio and visual modalities into sequences aligned with words. In the initial state, the audio modality sequence $\boldsymbol{A}$ has the form $l_a \times f_a$, where l_a stands for

the sequence length of the audio and f_a means the number of feature dimensions of the audio. Similarly, the video modality sequence is expressed as $\boldsymbol{V}$, and its form is $l_v \times f_v$. For word alignment, the sequence of each non-verbal modality is transformed from the non-verbal modality sequence $\boldsymbol{M}$ (either $\boldsymbol{A}$ or $\boldsymbol{V}$) into the aligned sequence $\boldsymbol{M}'$ by using a convolutional layer: $\boldsymbol{M}' = \mathrm{Conv1D}(\boldsymbol{M}^{\mathrm{T}})$. In this context, $\mathrm{Conv1D}$ represents a 1-dimensional convolutional layer, and matrices with a superscript T indicate transposed matrices. The described procedure generates modality sequences whose sizes are $l_l \times f_{m_l}$, where l_l corresponds to the length of the linguistic modality sequence $\boldsymbol{L}$ (with dimensions $l_l \times f_l$), and f_{m_l} signifies the dimension size of the non-verbal modality $m \in \{a \text{ or } v\}$. This dimension size is equivalent to $\frac{2f_m}{k}$, where k represents the kernel size of the convolutional layer.

3.2.3. Isolated Transformer Encoding

Next, transformer encoder layers are applied to each modality. Specifically, the BERT layer is utilized for the language modality. In contrast, feature vectors from other modalities require an additional transformer encoder prior to fusion. This is due to the architectural design of the MSU-layer, which is intended to be integrated into transformer encoder layers. Therefore, all modalities need to be encoded using transformer or transformer-like layers before undergoing fusion.

Optimal Positioning of MSU-Layer & Textual Modality Protection As previously mentioned, our model incorporates an MSU-layer immediately after a specific layer of the transformer encoder for each modality. In this paragraph, we elucidate the rationale behind selecting the appropriate layers within the transformer encoders to effectively propagate both word-level and contextual information to other non-verbal modalities. Striking a balance between the layer position is crucial: positioning the MSU-layer too early would mainly convey word-level information, while positioning it too late might sacrifice word-level details in favor of contextual understanding. This decision is underpinned by BERT's learning pattern, which initially captures word-level features in its early layers and subsequently comprehend sentence-level attributes in its later layers (Ethayarajh, 2019). Taking BERT-base as an example, considering its behavior and its progression towards more abstract features after the 9th encoder layer, we assert that the ideal placement for the MSU-layer is around the 9th layer. This is based on the observation made by Ethayarajh (2019) that BERT-base tends to acquire heightened abstraction in its features beyond this layer.

Another aspect of our method is that for the textual modality, the BERT layers before inserting the

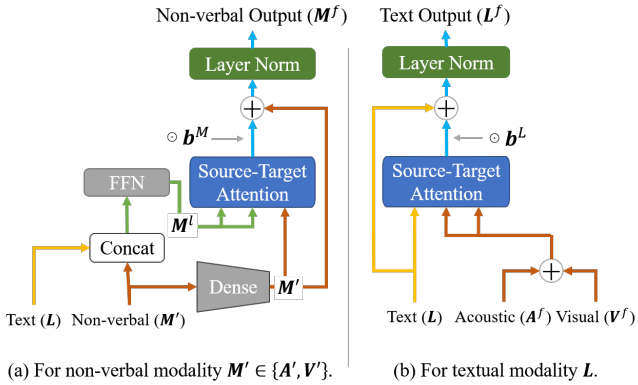

Figure 2: Design of the MSU-Layer.

MSU-layer are frozen. This aims to safeguard the integrity of language modality sequences preceding the MSU-layer against potential contamination from incoming information of non-language modalities. This protective measure prevents any potential interference with BERT's ability to comprehend words and contexts within the language modality. Therefore, the layers before the MSU-layer must be trained separately. We thus adopted a two-step training procedure, as explained below.

3.2.4. Two-step Training

We independently train the language modality encoder before its use in multimodal fusion, as stated in the preceding paragraph.

In the first step, we train the language model using only the text modality from the given dataset and select the best one from the results. In the second step, we train the multimodal fusion by applying the weight we retrieved at the first step to the language model. Furthermore, we freeze the encoder layers up to the point where the MSU-layer is inserted.

3.3. Phase2: MSU-Layer

The MSU-layer, which is vital to WA-MSF, transfers semantic and contextual information from the text modality to the non-verbal ones prior to fusion. Before reaching this layer, the sequence length of each non-verbal modality has been aligned to match l_l, which corresponds to the word length. This alignment is accomplished using Conv1D, as elaborated in section 3.2.2. However, it is essential to recognize that, at this particular stage, the latent non-verbal tokens with a length of l_l are unable to utilize linguistic essences when calculating interrelations between modalities. This layer introduces language semantic and contextual information and then integrates them with non-verbal tokens. Consequently, the non-verbal modality learns relationships among its l_l latent tokens by incorporating

linguistic information (see Fig. 2-(a) also).

First, the MSU-layer initially constructs a "source" tensor $\boldsymbol{M}^l$ (either $\boldsymbol{A}^l$ or $\boldsymbol{V}^l$) for the subsequent source-target attention processing. The process begins by concatenating the non-linguistic modality $\boldsymbol{A}$ (or $\boldsymbol{V}$) and the linguistic modality sequence $\boldsymbol{L}$. Following this, the MSU-layer employs a feed-forward neural network[3] (FFN) layer to reduce the dimensionality of the concatenated sequence from $f_l + f_{a_l}$ (or $f_l + f_{v_l}$) to f_l. This adaptation is necessary because the original size of the language embeddings is meticulously designed to effectively represent its language-specific attributes. $\boldsymbol{M}^l$ is thus computed as $\boldsymbol{M}^l = \mathrm{FFN}([\boldsymbol{L}; \boldsymbol{M'}])$.

Second, a dense layer is employed on each modality sequence; the layer aligns the dimension of the sequence (f_{a_l} for audio and f_{v_l} for visual) with that of the linguistic modality sequence (f_l). This alignment is crucial for computing the source-target attention in the subsequent step. Importantly, these modality sequences have been duplicated prior to forming the "source" tensor. As a result, these sequences remain distinct from the "source" side of the non-linguistic modality. Nonetheless, we will maintain the reference to these processed modality sequences as $\boldsymbol{M'}$, since the original modality sequences are no longer utilized.

Third, source-target attention is applied between the non-verbal modality sequence augmented by the text ($\boldsymbol{M}^l$) as the source and the dimension-modified non-verbal modality sequence ($\boldsymbol{M'}$) as the target. After this process, the MSU-layer provides a fusion-ready non-verbal modality sequence $\boldsymbol{M}^f$, defined as follows:

$$\boldsymbol{M}^f = \mathrm{Norm}(\boldsymbol{b}^M \odot \mathrm{Attn}(\boldsymbol{M}^l, \boldsymbol{M'}) + \boldsymbol{M'}). \quad (1)$$

Here, $\odot$ represents the Hadamard product, Attn denotes an attention layer, and Norm signifies the layer normalization process. Here, we propose that this process should include a residual connection with weight $\boldsymbol{b}^M$. This is for two reasons: we found that (1) a residual connection without any correction destabilizes the learning, and (2) the learning can be stabilized by applying a correction based on the size of the original latent space that the modality had. Therefore, we decided to define $\boldsymbol{b}^M$ as follows:

$$\boldsymbol{b}^M = \min\left(\frac{||\boldsymbol{M'}||_2}{||\mathrm{Attn}(\boldsymbol{M}^l, \boldsymbol{M'})||_2}, \boldsymbol{t}_b\right). \quad (2)$$

$\boldsymbol{t}_b$ is a minimum threshold vector for the coefficient of the residual connection; this is a hyperparameter.

After that, our method transfers the effects of the aforementioned operations to the language modality to enhance the efficiency of the MSU-layer. Concretely, our method aims to back propagate information from the non-verbal modality processed by the MSU-layer to the hidden embeddings of the textual modality. These hidden embeddings come from the encoder layer where the MSU-layer is inserted, which is the first learnable layer just after the frozen layers. The actual procedure is that the audio and visual modalities $\boldsymbol{A}^f$ and $\boldsymbol{V}^f$ are summed up and source-target attention is applied to the language modality in the same way as in equation (1) (see Fig. 2-(b) also). The source is the sum of the non-verbal modalities, and the target is the language modality. After this process, the MSU-layer outputs $\boldsymbol{L}^f$.

[3]The FFN here is similar to the one implemented in the transformer architecture (Vaswani et al., 2017).

3.4. Phase3: Modality Fusion by aMLP

Once all modalities have passed through their dedicated encoder layers including the MSU-layer, they are combined and subjected to another set of aMLP encoder layers (gray box in Fig. 1). The procedure is detailed below.

3.4.1. Modality fusion by concatenation ("Dense & Concat" in Fig. 1)

First, all the modalities are concatenated into one simple fused feature with summarized vectors from the sequences of individual modalities. This summarized vector is expected to play a role like CLS embedding of BERT in fused sequence. For the language modality, the summarized vector is just the classification vector of the first token embedding in the sequence. For the other modalities, maxpooling is applied to their sequence, and the resulting output is retrieved as the summarized vector. This operation retrieves the language vector $\boldsymbol{l}^f$, audio vector $\boldsymbol{a}^f$, and video vector $\boldsymbol{v}^f$. The fused sequence $\boldsymbol{f}$ is constructed through a dense layer as follows: $\boldsymbol{f} = \boldsymbol{w}[\boldsymbol{l}^f; \boldsymbol{a}^f; \boldsymbol{v}^f] + \boldsymbol{b}$.

To reduce the number of parameters, the dimension size of the vector is reduced to n_f by using the above dense layer. n_f is a hyperparameter. Subsequently, the sequences of all of the modalities are concatenated along the sequence direction, and the fused-summarized vector $\boldsymbol{f}$ is designated as the first element. This leads to the computation of a single fused modality sequence $\boldsymbol{F}$ with a shape of $n_f \times (l_l \times 3 + 1)$ as $\boldsymbol{F} := [\boldsymbol{f}, \boldsymbol{l}_1^{n_f}, \ldots, \boldsymbol{l}_{l_l}^{n_f}, \boldsymbol{a}_1^{n_f}, \ldots, \boldsymbol{a}_{l_l}^{n_f}, \boldsymbol{v}_1^{n_f}, \ldots, \boldsymbol{v}_{l_l}^{n_f}]$. In this context, $\boldsymbol{l}^{n_f}$, $\boldsymbol{a}^{n_f}$, and $\boldsymbol{v}^{n_f}$ refer to the individual elements of $\boldsymbol{L}^{n_f}$, $\boldsymbol{A}^{n_f}$, and $\boldsymbol{V}^{n_f}$, respectively. These tensors have been derived from the previously obtained $\boldsymbol{L}^f$, $\boldsymbol{A}^f$, and $\boldsymbol{V}^f$ in the process, with their respective dimension sizes reduced to n_f through dedicated dense layers for the individual modalities.

3.4.2. Encoding Fused Modality Tensor

Then, the fused modality sequence $\boldsymbol{F}$ is encoded by the fusion layers.

Here, we utilize a fusion method called "gMLP with tiny self-attention" (aMLP), which is a variant of the gMLP model (gray box in Fig. 1). The authors of this method stated in their paper that gMLP efficiently captures latent spatial and temporal representations in audio and visual domains. Moreover, aMLP is specifically enhanced for language-related tasks by incorporating a tiny attention mechanism. Hence, we deemed this method to be more suitable for multimodal fusion than methods like the transformer. This is because, when it comes to representing non-linguistic information, leveraging the powerful capabilities of the function representation of MLP is better than introducing the dynamic inductive bias calculated by the attention mechanism.

The input and output structures of aMLP are compatible with the transformer. Thus, the fused sequence, $\boldsymbol{F}$, made from the outputs of BERT and the other modalities directly serves as the input to the aMLP layer.

The encoder outputs the final hidden states $\boldsymbol{F}^{enc}$ with the same shape as $\boldsymbol{F}$. The first embedding of the sequence, $\boldsymbol{f}_1^{enc}$, is retrieved as a representative vector of the fused feature sequence. This vector is fed into an output layer consisting of fully-connected neurons. The resulting output serves as a logit for the sentiment analysis task. To train this logit, we employed a log-cosh loss function (Jadon, 2020).

4. Evaluation

We performed a detailed evaluation of WA-MSF.

4.1. Dataset

The evaluation utilized two multimodal datasets as follows.

CMU-MOSI (Zadeh et al., 2016) is a dataset for multimodal machine-learning tasks compiled by Carnegie Mellon University. It mainly comprises videos of individuals expressing their sentiments about movies and TV dramas, which are then scored as positive or negative on a scale of -3 to 3. We used 1284 sentences for training and 685 sentences for testing.

CMU-MOSEI (Bagher Zadeh et al., 2018) is a similar dataset to CMU-MOSI but has a larger number of utterances. Furthermore, unlike CMU-MOSI, its utterances were randomly chosen from various topics and monologue videos. This dataset includes emotion measures, but we did not utilize them. Instead, we employed sentiments annotated on 16,272 sentences for training and 4,646 sentences for testing.

4.2. Evaluation Strategy and Metrics

For an equitable evaluation, we performed the training and evaluation of our method 100 times and used the maximum scores, average scores, and standard deviation in the comparisons described below. This is because it has been pointed out that the value of the set random seed sometimes affects the accuracy more than the improvement of the model to be verified (Picard, 2023). Therefore, in order to minimize its effect as much as possible, especially in the ablation study, we adopted a policy of performing the verification 100 times consecutively without resetting the random seed and comparing the accuracy by using those three values.

We picked five metrics that are commonly used in quantitative evaluations of multimodal sentiment analysis (Yang et al. (2020); Arjmand et al. (2021)): F-score (F1), binary accuracy (Acc^2), 7-class accuracy (Acc^7), mean absolute error (MAE), and correlation coefficient (Corr).

4.3. Parameter Tuning

We tuned the parameters of our model as follows.

4.3.1. Language Model

We had various options for the language modality model, including BERT, RoBERTA, XLNet, etc. Here, we present results for the "BERT large" model with a 1024-dimensional embedding and 24 encoder layers, which achieved the highest scores among the tested models.

4.3.2. Hyperparameters

We needed to determine some of the parameters for tuning the model itself and for training.

Parameters for tuning the model For the nonverbal modalities, the number of transformer encoder layers was 24, which equaled the number of layers of the BERT model.

For all modalities, the MSU-layer was inserted after the 20th layer, so layers 1 to 20 of BERT were frozen. This is because we predicted that the final four layers of the encoder in the BERT-large model would be utilized for abstract sentence comprehension. We conducted small-scale tests by inserting MSU-layers from the 18th to the 22nd layer and found that the insertion at the 20th layer was the most accurate. The frozen layers used the weights trained from the text-only data and were not affected by other modalities. This was a specific implementation of what is described in section 3.2.3.

The residual connection threshold t_b in the MSU-layer was set to 0.5. It has been observed that

setting a value higher than this significantly reduces the model's accuracy. However, as long as this value is varied within limits lower than this (and higher than 0.0), it will not have any effect on the experimental outcomes.

The dimensions size of the fused vector was 512, which is half the dimension size of the hidden vector of BERT-large. The reason for choosing this number is the need to balance learning speed and performance of the model. In particular, too large a vector risks divergence in learning. Also, the fusion stage had 18 aMLP layers.

We also considered each parameter of Conv1D. For audio, Conv1D used a kernel size k of 20, a stride of 10 and a padding of 5. For video, Conv1D had a kernel size of k equal to 3, a stride of 2, and padding of 1.

The maximum sequence lengths for text, audio, and video were 50, 5000, and 1250, respectively.

Parameters for training The random seed number was initially set to 42 to ensure reproducibility. Randomness was ensured among the evaluation attempts because this random seed was initialized only in the first attempt.

The batch size was 48 for CMU-MOSI and 384 for CMU-MOSEI. These settings were made in consideration of the total data size of each dataset.

RAdam (Liu et al., 2020) was chosen as the optimizer.

The learning rate was 2e-4 and no scheduler was applied. This is because the RAdam optimizer has a self-learning rate-alignment mechanism.

The maximum number of iterations was 50, and an early-stopping mechanism was applied. This is because, in almost all cases, the training saturated after 50 iterations.

4.4. Baseline methods for the evaluation

We prepared several baseline methods for the evaluation. They were selected from the related work and are currently considered SOTA on the open leaderboards.

For CMU-MOSI, CM-BERT, MAG-BERT / XLNet, TEASEL, CHFN, and UniMSE were the baselines.

For CMU-MOSEI, MMIM (Han et al., 2021), MAG-BERT, and UniMSE were the baselines.

4.5. Results

We trained and tested our model using the configuration described above. The results are shown in Table 1 and Table 2. In the subsequent sections of the paper, **bolded** scores in the tables represent the best results, and underlined scores represent the second-best results. The scores of the benchmark methods, except for BERT-large, were obtained from the respective papers, while the score of BERT-large was derived from the highest score observed in our repeatability test.

Table 1: Evaluation Result (CMU-MOSI). $\mathrm{XX_h}$: "higher is better", $\mathrm{XX_l}$: "lower is better".

Method	$\mathrm{F1_h}$	$\mathrm{Acc_h^2}$	$\mathrm{Acc_h^7}$	$\mathrm{MAE_l}$	$\mathrm{Corr_h}$
CM-BERT	84.5	84.5	44.9	0.729	0.791
MAG-BERT	82.5	82.37	43.62	0.727	0.781
MAG-XLNet	85.7	85.6	N/A	0.675	0.821
TEASEL	85	**87.5**	47.52	0.64	0.836
CHFN	86.2	86.4	48.6	0.689	0.809
UniMSE	86.42	86.9	48.68	0.691	0.809
BERT-large	86.04	85.98	50.51	0.636	0.838
Ours (Max)	**86.97**	86.86	**51.82**	**0.623**	**0.842**
Ours (Avg)	85.99	85.96	49.99	**0.629**	0.838

Table 2: Evaluation Result (CMU-MOSEI)

Method	$\mathrm{F1_h}$	$\mathrm{Acc_h^2}$	$\mathrm{Acc_h^7}$	$\mathrm{MAE_l}$	$\mathrm{Corr_h}$
MMIM	85.94	85.97	54.24	0.526	0.772
MAG-BERT	84.5	84.7	N/A	N/A	N/A
UniMSE	**87.46**	**87.50**	54.39	0.523	0.773
BERT-large	N/A	N/A	53.38	0.531	0.775
Ours (Max)	86.09	86.26	**54.63**	**0.515**	**0.785**
Ours (Avg)	85.67	85.80	53.71	**0.520**	**0.782**

First, we can see that BERT-large had high performance especially in the CMU-MOSI evaluation. It had the second-best Acc^7, MAE, and Corr scores; no baseline models performed better than BERT on these three metrics.

On the other hand, our method outperformed the prior SOTA methods on both CMU-MOSI and CMU-MOSEI on the majority of metrics and was at least second-best on almost all metrics. Needless to say, it outscored BERT-large. In particular, it showed significant improvements in terms of Acc^7, MAE, and Corr on the regression task of sentiment analysis. Note that the Acc^7, MAE, and Corr metrics are designed for regression tasks. In the realm of sentiment analysis, regression involves discerning subtle nuances in utterances and emotional context. In this task, the impact of multimodal information on enhancing language-derived data greatly affects the model's accuracy. In essence, these results underscore our model's effective use of multimodal data, leading to impressive sentiment analysis performance.

4.6. Ablation Study

We tested several subsets of our method on the CMU-MOSI dataset to evaluate the effectiveness of the multimodality and the fusion methods. A summary of the results is shown in Table 3.

Table 3: Ablation Study Results - mean and standard deviation (CMU-MOSI)

Method	$F1_h$	Acc_h^2	Acc_h^7	MAE_l	$Corr_h$
Modality combination					
Text (BERT-large)	85.70 ± 0.66	85.67 ± 0.64	47.73 ± 1.52	0.6591 ± 0.0149	0.8270 ± 0.0082
Text + Video	**86.05 ± 0.43**	**86.01 ± 0.41**	49.87 ± 0.78	0.6319 ± 0.0027	0.8363 ± 0.0024
Text + Audio	85.87 ± 0.46	85.85 ± 0.43	49.94 ± 0.69	0.6298 ± 0.0027	0.8368 ± 0.0021
Full	85.99 ± 0.47	85.96 ± 0.44	**49.99 ± 0.74**	**0.6288 ± 0.0027**	**0.8376 ± 0.0019**
Fusion method					
Vanilla	85.54 ± 0.45	85.54 ± 0.43	49.65 ± 0.87	0.6362 ± 0.0039	0.8357 ± 0.0022
+ Transformer	85.89 ± 0.43	85.86 ± 0.41	49.64 ± 0.77	0.6349 ± 0.0027	0.8361 ± 0.0019
+ MSU-Lyr	85.85 ± 0.37	85.82 ± 0.39	49.78 ± 0.77	0.6338 ± 0.0026	0.8358 ± 0.0019
+ aMLP	85.95 ± 0.43	85.92 ± 0.40	49.85 ± 0.78	0.6310 ± 0.0030	0.8370 ± 0.0022
+ MSU-Lyr	**85.99 ± 0.47**	**85.96 ± 0.44**	**49.99 ± 0.74**	**0.6288 ± 0.0027**	**0.8376 ± 0.0019**

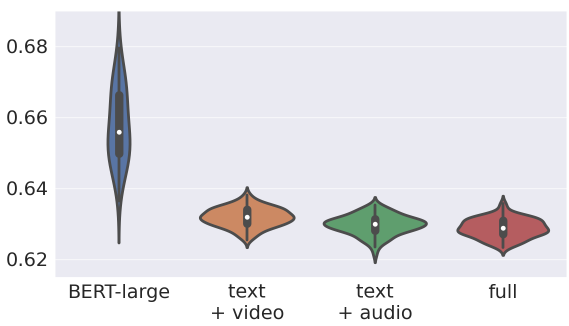

Figure 3: Ablation study, modality (MAE). Bolded black lines in violin plots indicate its quartiles.

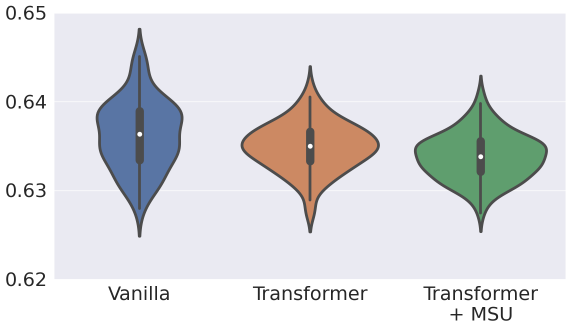

Figure 4: Ablation study, transformer fusion (MAE).

4.6.1. Effectiveness of Multimodality

Initially, we evaluated our method by removing some modalities from its data source. As shown in the top half of Table 3, we selected four patterns of combining modalities, i.e., "text-only", "text and video", "text and audio", and "full modalities" (see Fig. 3). The results indicate that every multimodal combination outperformed BERT-large's text-only approach. For instance, all the modality combinations provided two or more percent higher accuracy in terms of Acc^7. This indicates that our method has sufficient strength in multimodal fusion for the sentiment analysis task. Furthermore, there were only slight differences in the results for the various modality combinations, unlike the big difference for the text-only model. Overall, we found that full modality fusion (text, audio, and video) provided the most stable accuracy and the best performance on the regression task. These findings are proof that full modality fusion is the best in terms of Acc^7, MAE, and corr among all modality combinations on regression-related tasks.

4.6.2. Effectiveness of Fusion Methods

As described above, our method consists of two fusion steps: one involves inserting the MSU-layer among the transformer layers for each modality and the other is the multimodal fusion by aMLP. Here, we conducted an ablation study to assess the actual effectiveness of each step. We prepared five different fusion method combinations:

(1) Fuse all modalities directly after the encoder layer for each modality without any fusion methods (Vanilla),

(2) Use transformer for fusion ("Transformer" in Fig. 4),

(3) Add the MSU-layer before the transformer fusion ("Transformer + MSU" in Fig. 4),

(4) Use aMLP for fusion ("aMLP" in Fig. 5),

(5) Use both the MSU-layer and aMLP layers ("aMLP + MSU" in Fig. 5; this is the same as the full version of our method).

The results are displayed in the bottom half of Table 3. They show that the MSU-layer with transformer fusion has a positive effect on MAE and Acc^7. Furthermore, the combination of the transformer and MSU-layer outperforms the basic "vanilla" fusion. However, inclusion of the MSU-layer led to slightly worse performance in terms of Acc^2, F1, and correlation compared with using only transformer fusion. Therefore, the collaboration between

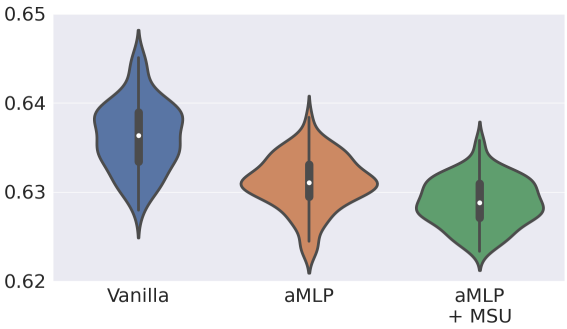

Figure 5: Ablation study, aMLP fusion (MAE).

the transformer and MSU-layer did not exhibit the desired level of synergy.

Another discovery concerns our proposal, aMLP fusion. The aMLP layer showed the highest performance, excelling not only in transformer fusion but also in the fusion combining the transformer and MSU-layer. Furthermore, the aMLP layer had good synergy with the MSU-layer, whereby aMLP-MSU fusion outscored aMLP-only fusion on all metrics. Therefore, both the MSU and aMLP layers effectively improve the accuracy of our model in sentiment analysis tasks. This result supports our hypothesis regarding aMLP.

5. Conclusion

We proposed a new method of multimodal-fused sentiment analysis, called word-Aware-modality-stimulation fusion (WA-MSF). The core idea of our approach revolves around the modality-stimulation-unit layer (MSU-layer) designed to activate linguistic information within non-verbal modalities by referencing the verbal modality sequence prior to the fusion process. We also employed aMLP as a multimodal fusion process, which enables understanding of individual modalities as well as learning after fusion; aMLP is the most applicable fusion method. Our experiments showed strong synergy between the MSU-layer and aMLP in fusion and demonstrated that our approach achieved SOTA accuracy, particularly on sentiment analysis regression tasks. Our research findings will facilitate seamless multimodal fusion and have the potential to accelerate related research in the field of multimodal analysis.

6. Ethical considerations

In this study, when undertaking multi-modal learning, it is necessary to utilize both visual and auditory information of the speakers. This information is closely associated with personal data, and it is ethically imperative to ensure sufficient anonymization within the utilized datasets or within the internal model structures.

The datasets we employed, namely CMU-MOSI and CMU-MOSEI, have effectively implemented anonymization measures for these aspects and have made the data available in a suitable form as open data. Additionally, WA-MSF promptly transforms input data into features, so no personally identifiable information is left within the model parameters. Therefore, the requirements for the ethical perspective mentioned above have been met.

7. References

Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. TEASEL: A transformer-based speech-prefixed language model. *Journal on Computing Research Repository*, abs/2109.05522.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–10.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, and Sunil Kumar Kopparapu. 2019.

Audio-visual fusion for sentiment classification using cross-modal autoencoder. In *Proceedings of 32nd conference on neural information processing systems*, pages 1–4.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65.

Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. 2010. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces, Special Issue on Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots*, 3:7–12.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3394–3402.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851.

Shruti Jadon. 2020. A survey of loss functions for semantic segmentation. In *Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–7.

Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. 2021. Pay attention to mlps. *Journal on Computing Research Repository*, abs/2105.08050.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the 8th International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

David Picard. 2023. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *Journal on Computing Research Repository*, abs/2109.08203.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman. 2020. Low rank fusion based transformers for multimodal sequences. In *Proceedings of the Second Grand-Challenge and Workshop on Multimodal Language*, pages 29–34.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression

using bidirectional lstm modeling. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pages 2362–2365.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2023. Multi-modality multi-loss fusion network.

Kaicheng Yang, Hua Xu, and Kai Gao. 2020. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 521–528.

Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023. Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7900–7913.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *Journal on Computing Research Repository*, abs/1606.06259.

Appendix A. Evaluation environment

The following environment was configured for **CMU-MOSI**:

- GPU: NVIDIA V100 GPU with 16 GiB VRAM.
- Driver version of GPU: 440.33.01.
- CUDA version: 10.2.
- OS: Ubuntu 18.04.6 (docker-isolated environment).
- Python for actual code, version 3.8.0.
- Pytorch as the library, version 1.12.0+cu102.
- Transformers as the library as well, version 4.15.0.

The following environment was configured for **CMU-MOSEI**:

- GPU: NVIDIA A100 GPU with 80 GiB VRAM.
- Driver version for GPU: 450.80.02.
- CUDA version: 11.0.
- OS: Ubuntu 20.04.4 (docker-isolated environment).
- Python for actual code, version 3.8.10.
- Pytorch as the library, version 1.12.0+cu113.
- Transformers as the library as well, version 4.15.0.

Appendix B. Parameter size

Our "Full" model (employing BERT-Large for textual modality, having L+A+V modalities with 24 encoder layers and fused using 18 aMLP layers) contained 524,824,647 parameters. This is approximately x1.5 the number of parameters of the BERT-Large model and x4.8 that of the BERT-Base model.

We also discovered that using an encoder with 12 layers for the audio and visual modalities had only a slight impact on accuracy. In this case, the total number of parameters was 476,831,847 (approximately x1.4 that of BERT-Large and x4.3 that of BERT-Base).