MSE-Adapter: A Lightweight Plugin Endowing LLMs with the Capability to Perform Multimodal Sentiment Analysis and Emotion Recognition

Yang Yang, Xunde Dong*, Yupeng Qiang

South China University of Technology, School of Automation Science and Engineering
China, Guangzhou
audxd@scut.edu.cn

Abstract

Current Multimodal Sentiment Analysis (MSA) and Emotion Recognition in Conversations (ERC) methods based on pre-trained language models exhibit two primary limitations: 1) Once trained for MSA and ERC tasks, these pre-trained language models lose their original generalized capabilities. 2) They demand considerable computational resources. As the size of pre-trained language models continues to grow, training larger multimodal sentiment analysis models using previous approaches could result in unnecessary computational cost. In response to this challenge, we propose **M**ultimodal **S**entiment Analysis and **E**motion Recognition **Adapter** (MSE-Adapter), a lightweight and adaptable plugin. This plugin enables a large language model (LLM) to carry out MSA or ERC tasks with minimal computational overhead (only introduces approximately 2.6M to 2.8M trainable parameters upon the 6/7B models), while preserving the intrinsic capabilities of the LLM. In the MSE-Adapter, the Text-Guide-Mixer (TGM) module is introduced to establish explicit connections between non-textual and textual modalities through the Hadamard product. This allows non-textual modalities to better align with textual modalities at the feature level, promoting the generation of higher-quality pseudo tokens. Extensive experiments were conducted on four public English and Chinese datasets using consumer-grade GPUs and open-source LLMs (Qwen-1.8B, ChatGLM3-6B-base, and LLaMA2-7B) as the backbone. The results demonstrate the effectiveness of the proposed plugin.

Code — https://github.com/AZYoung233/MSE-Adapter

Introduction

Making artificial intelligence (AI) comprehend human sentiment is a significant issue in the development of AI. Multimodal Sentiment Analysis (MSA) and Emotion Recognition in Conversations (ERC) have received widespread attention from the Natural Language Processing (NLP) community in recent years (Mai, Zeng, and Hu 2023; Wang et al. 2024). Limited by various factors, early sentiment analysis could only be conducted through text, which is far from sufficient for the diversified information of the real world. The same sentence, accompanied by different facial expressions or tones, can convey entirely different emotions. Therefore, utilizing multimodal information to perceive human emotions can enable AI to make more accurate judgments.

In recent years, the emergence of pre-trained Large Language Model (LLM) has introduced some new paradigms to the research community in NLP. Thanks to the powerful capabilities of LLM, they have demonstrated outstanding performance on many downstream tasks. Consequently, numerous researchers have begun to develop domain-specific LLM (such as for healthcare, weather forecasting, etc.) (Toma et al. 2023; Bi et al. 2023), by fine-tuning LLM with high-quality data to exhibit strong capabilities in relevant fields. This paradigm of LLM-for-specific-field is an excellent approach, but it demands high quality and quantity of data, making its implementation somewhat challenging.

Some researchers in the field of sentiment analysis also leveraged this paradigm to tailor models specifically for MSA and ERC tasks. Hu et al. (2022b) introduced UniMSE, which for the first time unifies MSA and ERC as generative tasks. UniMSE assigns emotional category labels to samples in MSA tasks and sentiment intensity labels to samples in ERC tasks by calculating textual similarity. Subsequently, a model is trained based on the T5 (Raffel et al., 2020) architecture, exhibiting remarkable proficiency in both MSA and ERC tasks. Li et al. (2023b) developed UniSA, a comprehensive framework for sentiment analysis. Built upon the BART (Lewis et al. 2019) model and a series of pre-training tasks, UniSA excels in a wide range of multimodal and text-only sentiment analysis tasks, offering a more comprehensive performance compared to UniMSE. Although the two aforementioned works exhibit remarkable performance, they still exhibit certain limitations. 1) **Significant computational overhead**. Large number of trainable parameters and multi-task training lead to considerable computational expenses, where UniSA's pre-training requires three days on eight NVIDIA RTX V100 32G GPUs. 2) **Losing the generalization capability inherent to the base models**. UniMSE and UniSA, trained respectively on T5 (Raffel et al. 2020) and BART (Lewis et al. 2019), are tailored for the sentiment analysis domain, restricting these models from performing tasks outside of sentiment analysis.

Currently, there is a trend towards developing plugins that freeze the backbone LLM and train them to perform non-textual tasks while retaining their inherent capabilities

---

*Xunde Dong is the corresponding author.

(Tsimpoukelli et al. 2021; Alayrac et al. 2022; Chen et al. 2023; Sun et al. 2023a; Szot et al. 2023) Given the expectation of continued growth in the parameter size of LLM, we propose developing a plugin with reduced training overhead. This will enable LLM to maintain its intrinsic capabilities while effectively executing a range of multimodal sentiment analysis tasks with minimal computational resources. We believe this approach is more promising than training a specialized sentiment analysis LLM. Our contributions are summarized as follows:

- We proposed **M**ultimodal **S**entiment Analysis and **E**motion Recognition **Adapter** (MSE-Adapter), a plug-and-play lightweight plugin. This plugin is trained with minimal overhead without altering the inherent capabilities of the LLM, enabling it to perform MSA or ERC task via autoregressive generated sentiment labels.

- In the MSE-Adapter, we introduce the Text-Guide-Mixer (TGM) and Multi-Scale Fusion (MSF) modules. The TGM establishes explicit connections between non-textual and textual modalities. This facilitates feature-level alignment between them and encourages the LLM to better understand content from non-textual modalities. The MSF module enables early fusion of non-textual features across multiple scales, further augmenting the LLM's efficiency in integrating information from diverse modalities.

- Extensive experiments were conducted on four publicly available Chinese and English datasets using consumer-grade GPUs with open-source LLMs (Qwen-1.8B, ChatGLM3-6B-base, and LLaMA2-7B) as the backbone. The results demonstrate the effectiveness of the MSE-Adapter.

Related Work

Pre-trained language model for MSA and ERC

Currently, in the MSA and ERC communities, a number of outstanding works have emerged, including contrastive learning-based methods (Yang et al. 2023a; Mai, Zeng, and Hu 2023), graph-based methods (Li et al. 2023a; Hu et al. 2021; Lin et al. 2022), transformer-based methods (Sun et al. 2023b; Zhang et al. 2023; Yang et al. 2023b), and methods based on Pre-trained Language Model (PLM). PLM-based methods typically use a designated PLM, such as BERT (Devlin et al. 2018) or T5 (Raffel et al. 2020), as their foundation. These methods convert non-textual modality features into tokens with equivalent dimensions to those of textual modalities, enabling training within the PLM framework.

Rahman et al. (2020) proposed the MAG, where completed word embeddings are fused with non-textual modality features to generate new embeddings. These embeddings are then fine-tuned within PLMs (such as BERT and XL-Net (Yang et al. 2019)) to achieve notable performance improvements. Similarly, Guo et al. (2022) proposed CHFN, which uses its designed Multimodal Interaction layer to integrate non-textual modal information into textual embedding at the word level, and then fine-tunes the integrated multimodal information by feeding it into BERT. Additionally,

Hasan et al. (2023) presented TextMI, a method that converts audio and vision information into corresponding textual descriptions. This approach links these descriptions with the textual content, transforming multimodal information into purely textual information. By inputting this enhanced text into BERT, TextMI achieves competitive performance results. Hu et al. (2022b) introduced UniMSE, an approach that uses the T5 (Raffel et al. 2020) model as its foundation. UniMSE encodes text using the initial layers of T5's encoder and then trains the remaining layers using a combination of non-textual and textual features. This training strategy incorporates contrastive learning to enhance the model's representation learning capabilities. Benefiting from its multitask training paradigm, UniMSE shows capabilities in both MSA and ERC tasks, demonstrating exceptional performance. Li et al. (2023b) developed UniSA, a comprehensive framework for sentiment analysis. UniSA uses PLMs (GPT2-medium (Radford et al. 2019), T5 and BART) as a foundation, standardizing the data formats of various types of sentiment analysis sub-tasks for input into PLMs. It leverages pre-training tasks and contrastive learning to pre-train the PLM, followed by fine-tuning on downstream task datasets. The UniSA$_{BART}$ achieve comprehensive results across multiple sentiment analysis sub-tasks.

Compared to aforementioned works, our proposed approach, MSE-Adapter, is a lightweight plugin that requires fewer training parameters (approximately 2.6M to 2.8M trainable parameters for base models sized 6/7B). Notably, MSE-Adapter preserves the inherent generalization capability of the LLM without sacrificing efficiency. Therefore, when assigned MSA or ERC tasks, the user can invoke the relevant pre-trained MSE-Adapter plugin to carry out the designated task. This design enhances parameter efficiency while preserving the effectiveness and adaptability of LLM, offering a new and efficient solution for using LLM in MSA and ERC tasks.

Adapters enabling LLM to perform non-plain text tasks

Adapters were usually utilized for efficiently fine-tuning large pre-trained models (Houlsby et al. 2019; Pfeiffer et al. 2020; He et al. 2022; Hu et al. 2023). By freezing the main body of the pre-trained model and only training the Adapter, the essence is to use gradient backpropagation to let the Adapter generate pseudo tokens that can be recognized by the pre-trained model. This process is aimed at prompting the pre-trained models to further adapt to certain downstream tasks. Meanwhile, since the pre-trained model is frozen during the training phase, it retains its strong generalization capability, avoiding the issue of catastrophic forgetting (Liu et al. 2021, 2023). Inspired by these relevant works of Adapter, some researchers have argued that it is possible to convert information from non-textual modalities into information understandable by LLMs through an Adapter, enabling them to perform downstream tasks involving non-plain text modalities. Tsimpoukelli et al. (2021) introduced Frozen, which utilizes a vision encoder to convert images into a series of tokens. These tokens are concatenated with a prompt and used to train a LLM for visual question
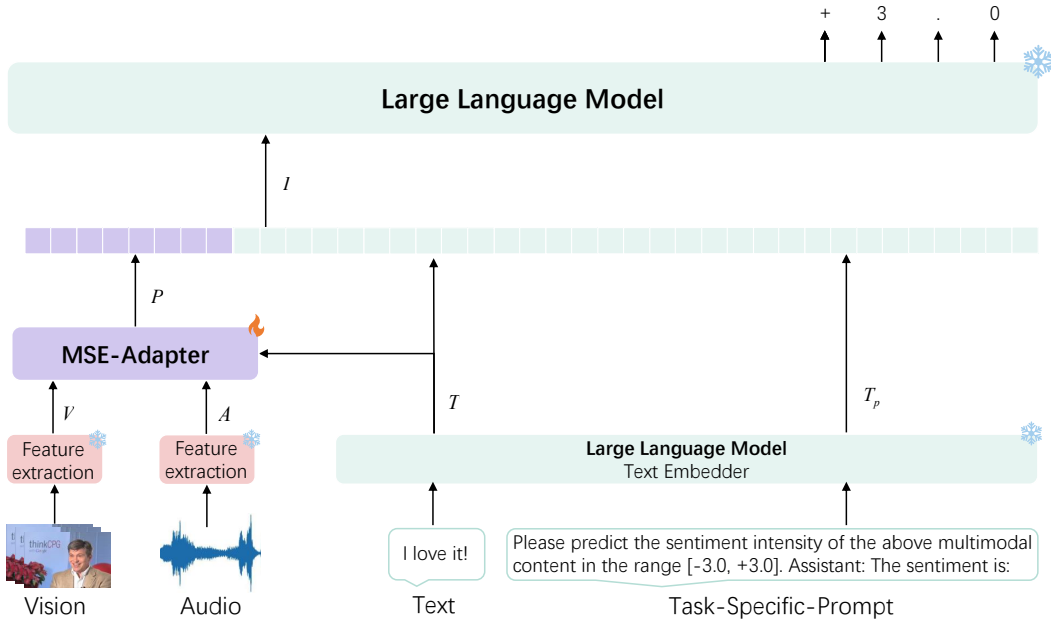
Figure 1: The comprehensive framework integrating MSE-Adapter with LLM.

answering (VQA) and captioning tasks, with gradient back-propagation guiding updates to the parameters of the vision encoder. Similarly, Alayrac et al. (2022) proposed Flamingo, which incorporates trainable cross-attention layers into a frozen LLM to fuse textual and vision modalities after embedding vision modality information using a pre-trained vision encoder. Flamingo exhibits remarkable performance across various video/visual-related tasks following training. Chen et al. (2023) presented X-LLM, a model that leverages X2L interfaces to convert vision, image, and audio modalities into "foreign languages" that can be processed by the LLM. X-LLM demonstrates impressive performance after instruction-tuning on a high-quality multimodal instruction dataset. Sun et al. (2023a) developed TEST, which utilizes contrastive learning to train an encoder for time series (TS) data, applies similarity constraints to align it with text, and fine-tunes the LLM with a soft-prompt approach to effectively process TS-related tasks.

In this paper, we introduce a lightweight plugin named MSE-Adapter, which enable the LLM's to perform MSA or ERC task without affecting its inherent capabilities. Unlike previous works, we introduce a novel module named TGM in the MSE-Adapter. TGM facilitates feature-level alignment between non-textual and textual modalities, which aids the LLM to better understand content from non-textual modalities.

Method

Overall architecture

Figure 1 presents the comprehensive framework integrating MSE-Adapter with LLM. Given a sample x_i, the first step is to convert each modality into a sequence embedding. For the textual modality, we use the Text Embedder within the LLM to tokenize (Kudo and Richardson 2018) the input from the text modality and the Task-Specific-Prompt, and then convert them into sequence embeddings. For audio and vision modalities, we employ pre-trained toolkits (Yu et al. 2021; Liu et al. 2022; Hu et al. 2022b; Sun et al. 2023b) for feature extraction to transform them into feature sequence embeddings (See Appendix A for more details). After the encoding process, the sample's textual modality, Task-Specific-Prompt, vision, and audio modalities are represented as $T \in \mathbb{R}^{l_t \times d_t}$, $T_p \in \mathbb{R}^{l_t \times d_t}$, $V \in \mathbb{R}^{l_v \times d_v}$ and $A \in \mathbb{R}^{l_a \times d_a}$, respectively, where $l_{m \in \{t,v,a\}}$ denotes the sequence length of each modality embedding, and $d_{m \in \{t,v,a\}}$ represents the corresponding feature vector dimension.

Subsequently, T, V and A are fed into the MSE-Adapter for further processing and generating pseudo tokens P. Finally, we concatenate P, T, T_p to obtain I (i.e., $I = [P; T; T_p]$) and input it into the frozen LLM, which returns the logits g_i corresponding to the input sample and the generated text y_i for the entire sentence (including input and output tokens). This can be expressed as:

$$\{y_i, g_i\} = LLM(I, \theta) \tag{1}$$

where θ represents the parameters of the LLM. The LLM predicts the conditional probability $\rho(\gamma_j | I, \theta)$ of each token γ_j of the generated text y_i until the end-of-sequence symbol $<$eos$>$. For the logits $g_i \in \mathbb{R}^{I_l \times V_s}$, where I_l and V_s represent the length of the input I and the size of the vocabulary used by the LLM, respectively.

Following the original training methodology of the LLM, we utilize the next-token prediction loss to measure the output error of the model. Hence, the loss calculation for the model task L_{task} is defined as follows:

$$L_{task} = \sum_{k=1}^{N} -log\rho(Y_k | I, \theta) \tag{2}$$

25644

where Y_k represents the k-th token of the sentiment label corresponding to the sample x_i, N denotes the number of sentiment label tokens. Based on the aforementioned loss, we optimize the parameters of the MSE-Adapter through gradient backpropagation, thereby enhancing the MSE-Adapter's adaptation to the LLM. Upon training completion, the LLM can utilize the MSE-Adapter to receive multimodal inputs and generate autoregressive sentiment labels, similar to text generation.
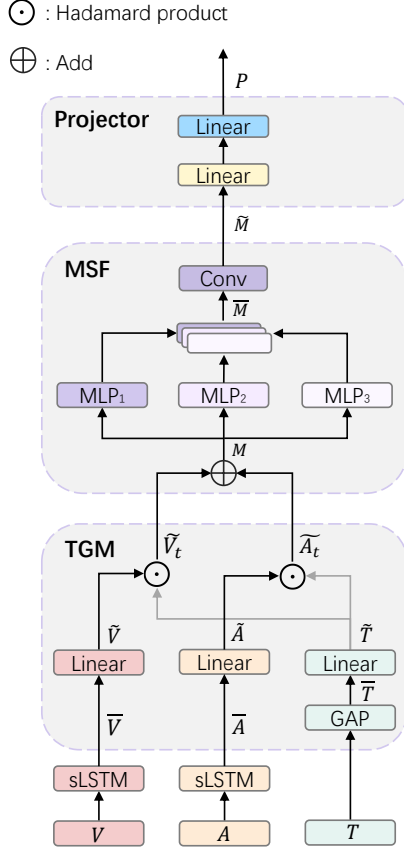

Figure 2: The architecture of MSE-Adapter.

MSE-Adapter

In this section, we introduce the proposed MSE-Adapter, whose structure is depicted in Figure 2. The MSE-Adapter consists of two separate single directional Long Short-Term Memory (sLSTM) modules, the TGM module, the MSF module, and a Projector module. The sLSTM initially performs temporal modeling separately for V and A, and then it captures their end-state hidden representations to derive $\overline{V} \in \mathbb{R}^{h_v \times 1}$ and $\overline{A} \in \mathbb{R}^{h_a \times 1}$. These outputs are then fed into the TGM to obtain $\widetilde{V}_t \in \mathbb{R}^{h \times 1}$ and $\widetilde{A}_t \in \mathbb{R}^{h \times 1}$. Subsequently, these are input into the MSF for fusion, resulting in $\widetilde{M} \in \mathbb{R}^{h \times 1}$. Finally, a Projector composed of two linear layers expands this into pseudo tokens P.

Text-Guide-Mixer Feature alignment across different modalities has always been a significant issue in multimodal

tasks. A key to modal feature alignment is establishing connections between modalities. In the research on recommendation systems and search engines, researchers have applied the Hadamard product to achieve feature crossing, thus creating explicit connections between features at the vector level (Lian et al. 2018; Wang et al. 2021). Inspired by these works, we propose the TGM module. TGM establishes an explicit connection by computing the Hadamard product between the feature vectors of the text modality and those of the non-text modalities. This strategy not only preserves the original individual features of the non-text modalities but also encourages non-text feature vectors to align with text modality feature vectors, narrowing the gap between non-text modality features and text features, enabling the MSE-Adapter to generate pseudo tokens that are more easily understood by LLM.

The implementation is as follows: we first do global average pooling (GAP) on the textual modal inputs T to obtain $\overline{T} \in \mathbb{R}^{1 \times d_t}$, and then we use a linear layer to project $\overline{T}$, $\overline{V}$ and $\overline{A}$ to the same dimension:

$$\begin{aligned} \widetilde{T} &= W_t \overline{T}^T \\ \widetilde{V} &= W_v \overline{V} \\ \widetilde{A} &= W_a \overline{A} \end{aligned} \tag{3}$$

where $W_t \in \mathbb{R}^{h \times d_t}$, $W_v \in \mathbb{R}^{h \times h_v}$, and $W_a \in \mathbb{R}^{h \times h_a}$.

Subsequently, we perform the Hadamard product of $\widetilde{V}$ and $\widetilde{A}$ with $\widetilde{T}$, respectively:

$$\begin{aligned} \widetilde{V}_t &= \widetilde{V} \odot \widetilde{T} \\ \widetilde{A}_t &= \widetilde{A} \odot \widetilde{T} \end{aligned} \tag{4}$$

where $\odot$ represents the Hadamard product, that is, the element-wise multiplication of matrices. Through this process, we obtain the new non-textual modality representations $\widetilde{V}_t$ and $\widetilde{A}_t$.

Multi-Scale-Fusion Entrusting the task of complete multimodal fusion to a frozen LLM presents certain challenges. In response to this challenge, we adopt an early fusion approach for the non-textual modalities prior to their input into the LLM. Specifically, we introduc a module named MSF, dedicated to performing low-level fusion of the non-textual modalities. Subsequently, the high-level fusion between the textual and non-textual modalities is deferred to the LLM. This layered fusion approach enables the LLM to capture more refined and detailed multimodal fusion, thereby boosting model performance.

The implementation of the MSF module is as follows: Firstly, we sum the outputs $\widetilde{V}_t$ and $\widetilde{A}_t$ derived from the TGM to obtain M. Subsequently, we utilize three Multi-Layer Perceptrons (MLPs) with diverse hidden layer dimensions to conduct feature fusion at multiple scales on M. Specifically, for the ith ($i \in \{1, 2, 3\}$) MLP, we obtain m_i:

$$m_i = W_2^i \sigma(W_1^i M) \tag{5}$$

where $W_1^i \in \mathbb{R}^{h/k \times h}$, $W_2^i \in \mathbb{R}^{h \times h/k}$, $k \in \{8, 16, 32\}$, σ represents the GELU activation function. Subsequently,

we stack the fusion results from the three different scales to obtain $\overline{M} = [m_1, m_2, m_3] \in \mathbb{R}^{h \times 3}$. To further integrate information from various scales, we use a 1×1 convolution (Conv) to compress this information, resulting in $\widetilde{M} \in \mathbb{R}^{h \times 1}$.

Projector After extracting the fused information from non-textual modalities, the size of $\widetilde{M}$ is adjusted to meet the input requirements of the LLM using a linear layer. Subsequently, another linear layer is utilized to increase the number of pseudo tokens:

$$P = W_4(W_3\widetilde{M})^T \qquad (6)$$

where $W_3 \in \mathbb{R}^{d_t \times h}, W_4 \in \mathbb{R}^{n \times 1}$ (n is our predetermined number of pseudo tokens). The resulting pseudo tokens, denoted by P, can then be utilized as input into the LLM.

Experiment

Datasets

To more comprehensively evaluate the performance of the MSE-Adapter, we conducted MSA and ERC experiments on two popular English datasets and two popular Chinese datasets, namely MOSEI (Zadeh et al. 2018), MELD (Poria et al. 2018), SIMS-V2 (Liu et al. 2022), and CHERMA (Sun et al. 2023b). The details of the dataset are shown in the appendix B.

Experimental Settings

In this subsection, we briefly introduce the detailed setup of our experiments. Our investigation utilizes the Qwen-1.8B (Bai et al. 2023), ChatGLM3-6B-base (Du et al. 2022), and LLaMA2-7B (Touvron et al. 2023) models as the backbone. For ease of presentation, we add the prefix 'MSE' to the backbone to indicate the integration of the MSE-Adapter (e.g., MSE-Qwen-1.8B). The trainable parameters of MSE-Adapter in various LLMs are presented in Table 1. For a fair comparison, we selected [1111, 2222, 3333, 4444, 5555] as random seeds for the experiments and reported the average results achieved across these five random seeds. All experiments were conducted on a single NVIDIA RTX 4090 GPU. The optimizer used was AdamW with a warmup learning rate strategy. The rest of the settings can be found in Appendix C.

Model	Trainable parameters			
	MOSEI	SIMS-V2	MELD	CHERMA
MSE-Qwen-1.8B	1.35M	1.40M	1.35M	1.56M
MSE-LLaMA2-7B	2.63M	2.68M	2.63M	2.84M
MSE-ChatGLM3-6B	2.63M	2.68M	2.63M	2.84M

Table 1: Trainable parameters of LLMs integrating MSE-Adapter.

To standardize labels for the MSA task, for models where the tokenizer does not automatically generate tokens to distinguish between positive and negative labels, we manually add a '+' sign to labels greater than or equal to 0. Furthermore, to streamline the answer generation process for the ERC task, we translated the emotion labels into distinct numerical values and incorporated these details into the prompts, as illustrated in Appendix C.

Baselines

We compared the performance of "MSE-Qwen-1.8B", "MSE-LLaMA2-7B" and "MSE-ChatGLM3-6B" to that of previous state-of-the-art models: TFN (Zadeh et al. 2017), LMF (Liu et al. 2018), MISA (Hazarika, Zimmermann, and Poria 2020), MAG-BERT (Rahman et al. 2020), Self-MM (Yu et al. 2021), MMIM (Han, Chen, and Poria 2021), CHFN (Guo et al. 2022), UniMSE (Hu et al. 2022b), UniSA$_{BART}$, UniSA$_{T5}$, UniSA$_{GPT2}$ (Li et al. 2023b), AV-MC (Liu et al. 2022), MMGCN (Hu et al. 2021), MM-DFN (Hu et al. 2022a), EmoCaps (Li et al. 2022), GA2MIF (Li et al. 2023a), EFT, LFT, MulT (Tsai et al. 2019), PMR (Lv et al. 2021), and LFMIM (Sun et al. 2023b). The details of these baseline models are given in the Appendix D.

Metrics

In this study, due to the differing labels across datasets, we report our experimental results using a variety of metrics tailored to each dataset. For MOSEI, we report the mean absolute error (MAE), Pearson correlation (Corr), seven-category accuracy (Acc-7), binary accuracy (Acc-2), and F1 score as evaluation metrics (where Acc-2 and F1 are calculated based on a non-negative/negative standard). For SIMS-V2, we report MAE, Corr, Acc2_weak, Acc-2, and F1 (where Acc-2 and F1 are calculated based on a non-positive/positive standard, and Acc2_weak is used to further validate the model's performance on weakly emotional instances within the [-0.4, 0.4] range). For MELD and CHERMA, we report seven-category accuracy (Acc) and weighted F1 (WF1).

Results

Tables 2 and 3 present the results of the performance comparison of multiple methods on the MSA and ERC tasks, respectively. It is noteworthy that, overall, all three LLMs incorporating the MSE-adapter exhibited outstanding performance.

It is evident that MSE-ChatGLM3-6B, achieved the most comprehensive performance, outperforming baseline models on most metrics. Notably, it showed a great performance improvement on CHERMA, the dataset with the largest volume of data, where its WF1 was 2.19% higher than the best baseline LFMIM. Although MSE-LLaMA2-7B has a larger parameter size than MSE-ChatGLM3-6B, its overall performance was not as good, especially on the Chinese datasets SIMS-V2. This might be due to MSE-LLaMA2-7B's inherent limitations in processing Chinese text. Since the LLM is frozen, the performance of the model is highly dependent on the inherent capabilities of the LLM.

Interestingly, MSE-LLaMA2-7B still performed better than the baseline on CHERMA (even though it is not particularly skilled in Chinese), somewhat confirming that LLMs

---

[1]https://github.com/thuiar/MMSA/blob/master/results/result-stat.md

Model	MOSEI					SIMS-V2				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc2_weak	MAE	Corr
TFN*	78.50	78.96	51.60	0.573	0.714	76.51	76.31	66.27	0.323	0.667
LMF*	80.54	80.94	51.59	0.576	0.717	77.05	77.02	69.34	0.343	0.638
MulT*	81.15	81.56	52.84	0.559	0.733	79.50	79.59	69.61	0.317	0.703
MAG-BERT*	82.51	82.77	50.41	0.583	0.741	79.79	79.78	71.87	0.334	0.691
MISA	83.60	83.80	52.20	0.555	0.756	80.53	80.63	70.50	0.314	0.725
Self-MM*	82.81	82.53	53.46	0.530	0.765	79.01	78.89	71.87	0.335	0.640
MMIM	82.24	82.66	54.24	0.526	0.772	80.95	80.97	72.28	0.316	0.707
AV-MC	-	-	-	-	-	82.50	82.55	74.54	0.297	**0.732**
CHFN	83.70	83.90	54.30	0.525	0.778	-	-	-	-	-
UniMSE	85.86	85.79	54.39	0.523	0.773	-	-	-	-	-
UniSA$_{GPT2}$	71.02	-	41.36	0.838	-	-	-	-	-	-
UniSA$_{T5}$	84.22	-	52.50	0.546	-	-	-	-	-	-
UniSA$_{BART}$	84.93	-	50.03	0.587	-	-	-	-	-	-
MSE-Qwen-1.8B	84.12	83.45	52.02	0.558	0.725	80.44	80.24	73.09	0.311	0.678
MSE-LLaMA2-7B	86.74	86.51	**55.57**	**0.501**	**0.787**	75.53	75.44	68.61	0.382	0.553
MSE-ChatGLM3-6B	**86.91**	**86.77**	54.56	0.515	0.783	**83.77**	**83.76**	**75.24**	**0.296**	0.720

Table 2: Experimental results of the MSA task on the MOSEI and SIMS-V2 datasets: 1) Results for models marked with * on MOSEI are sourced from the official repository[1], while results for other models are extracted from relevant published papers; 2) All models' results on SIMS-V2 are cited from literature (Liu et al. 2022).

Model	MELD		CHERMA	
	Acc	WF1	Acc	WF1
TFN*	60.77	57.74	-	68.37
LMF*	61.15	58.30	-	68.23
EFT	-	-	-	68.72
LFT	-	-	-	69.05
MulT	-	-	-	69.24
PMR	-	-	-	69.53
LFMIM	-	-	-	70.54
MMGCN*	60.42	58.31	-	-
MM-DFN*	62.49	59.46	-	-
EmoCaps	-	64.00	-	-
GA2MIF	61.65	58.94	-	-
UniMSE	65.09	**65.51**	-	-
UniSA$_{GPT2}$	48.12	31.26	-	-
UniSA$_{T5}$	64.52	62.17	-	-
UniSA$_{BART}$	62.34	62.22	-	-
MSE-Qwen-1.8B	62.18	59.87	70.38	70.21
MSE-LLaMA2-7B	65.14	63.66	71.58	71.41
MSE-ChatGLM3-6B	**66.23**	65.13	**72.90**	**72.73**

Table 3: Experimental results of the ERC task on the MELD and CHERMA datasets: 1) Results for models marked with * on MELD are cited from the the literature (Hu et al. 2022a), while results for other models are extracted from relevant published papers; 2) All models' results on CHERMA are cited from literature (Sun et al. 2023b).

are essentially pattern machines (Mirchandani et al. 2023), capable of learning mapping relationships from large data volumes, even for tasks they are less proficient in. This was observed similarly on MSE-Qwen-1.8B. Although MSE-Qwen-1.8B's performance was lower than the other two backbones, it requires the least trainable parameters. Since MSE-Qwen-1.8B is deployable on mobile devices, such a small-parameter plugin can offer users a more efficient interaction experience.

Discussion

Ablation Study

To evaluate the contribution of each module within the MSE-Adapter, we conducted ablation experiments on the English dataset MELD and the Chinese dataset SIMS-V2. For MELD, we reported Acc and WF1 score, and for SIMS-V2, we reported Acc-2 and F1 score. Given that MSE-ChatGLM3-6B achieved the most comprehensive performance in our experiments, all of our ablation experiments were based on MSE-ChatGLM3-6B and reported the average results achieved across the five random seeds (same random seeds as experimental). Table 4 shows the results of the ablation experiments.

Effectiveness of TGM and MSF Initially, we evaluated the TGM and MSF module within the MSE-adapter. The results shown in Table 4 indicated that removing TGM decreased model performance, highlighting the importance of establishing cross-modal connections between non-textual and textual features for modality alignment and model performance. Similarly, omitting MSF also resulted in a decline in performance, with a more significant impact than that

	MELD		SIMS-V2	
	Acc	WF1	Acc-2	F1
w/o A	66.24	65.09	83.46	83.42
w/o V	66.13	65.00	78.84	78.68
w/o T	46.44	36.53	72.78	72.21
w/o A, V	57.11	53.92	76.54	75.98
w/o TGM	65.07	63.85	82.98	82.97
w/o MSF	62.72	61.75	81.37	81.32
MSE-Adapter	**66.23**	**65.13**	**83.77**	**83.76**

Table 4: The ablation experiments results on MELD and SIMS-V2, where T, V, and A represent textual, vision, and audio modalities, respectively. And the 'w/o' means remove a modality or module. Furthermore, in the case of "w/o T", the TGM module becomes ineffective. $\overline{V}$ and $\overline{A}$ are mapped to the same dimension through two independent linear layers and then directly summed before being fed into the MSF module. In the case of "w/o A, V", $\overline{V}$ and $\overline{A}$ are randomly initialised into the corresponding dimensions into the TGM and the model is trained anyway.

of TGM. This observation further illustrates that the multimodal fusion capacity of frozen LLM is limited. Therefore, to enhance the fusion of the three modalities, early fusion of non-textual modalities prior to LLM input is necessary.

The Impact of Absent Modalities Additionally, we removed one or several modalities from the multimodal signals to verify their impact on model performance. The ablation results are presented in Table 4. From the results, we discovered that eliminating either the vision or audio modalities, or both, led to a decrease in performance, indicating the necessity of non-textual modalities (i.e., vision and audio) for solving MSA and ERC tasks. Notably, the removal of the textual modality had the most severe impact on performance, underscoring the critical role of language in identifying emotions in the real world.

Further Discussion with Models Adapted for BERT

In previous work, MAG (Rahman et al. 2020) and CHFN (Guo et al. 2022) conducted similar research, primarily utilizing their designed adaptation model to fine-tune BERT with a classification header for MSA tasks. We previously reported their performance metrics. For a fair comparison, we reproduced their adaptation model based on the original paper and adapted it to ChatGLM3-6B within our framework. We conducted experiments on the MOSEI, MELD, SIMS-V2, and CHERMA dataset, maintaining consistent experimental settings with the aforementioned chapters. It should be noted that the hyperparameters are only reported in the original paper for the MOSEI dataset in the MSA task. Consequently, in our reproduction, we maintain consistency with the original paper by keeping the hyperparameters on MOSEI only. The hyperparameters on the rest of the datasets are set with reference to MOSEI. Due to space limitations, we only show partial results for MOSEI and MELD in table 5, and complete results are detailed in Appendix E.

Model	MOSEI		MELD		🔥Paras
	Acc-2	F1	Acc	WF1	
MAG-ChatGLM3-6B	85.10	84.73	60.38	59.81	34.47M
CHFN-ChatGLM3-6B	85.58	85.26	63.03	62.08	50.79M
MSE-ChatGLM3-6B	**86.91**	**86.77**	**66.23**	**65.13**	**2.63M**

Table 5: Experimental results on the MOSEI and MELD datasets. The "🔥Paras" represents the trainable parameter of the model.

The experimental results demonstrate that the MSE-Adapter outperforms both MAG and CHFN when using the same backbone network. Additionally, while replacing BERT with LLM in MAG and CHFN can improve performance, the improvements are limited. Despite integrating more comprehensive multimodal information, MAG and CHFN introduce a degree of information redundancy. When the backbone network is trained, it benefits from this rich information, which enhances its performance. However, when the backbone network is frozen, this redundancy complicates its ability to effectively interpret multimodal content. In contrast, the MSE-Adapter is more efficient in capturing key information and extracting key features used to complete sentiment analysis, which reduces the difficulty of the backbone's comprehension. In conclusion, MSE-Adapter not only effectively reduces the number of trainable parameters, but also enhances the performance of the backbone on MSA or ERC tasks.

Conclusion

This paper presents the MSE-Adapter, a lightweight plug-and-play plugin that empowers an LLM to handle MSA or ERC task without compromising its inherent capabilities. The MSE-Adapter includes a module called TGM, which is designed to facilitate the alignment of non-textual modalities with textual ones. By employing the Hadamard product, TGM establishes explicit connections between non-textual and textual modalities at the feature level, thereby enhancing the LLM's comprehension of content from non-textual sources. MSF is another module of MSE-Adapter that employs MLPs of varying scales for early fusion of non-textual modalities, followed by further integration of this information using convolutional layers. This further enhances the efficiency of LLM in fusing information from diverse modalities. Deployable on consumer-grade GPUs and utilizing open-source LLMs (Qwen-1.8B, ChatGLM3-6B-base, and LLaMA2-7B) as the backbone, we conducted extensive experiments across four public English and Chinese datasets. The competitive results demonstrate the efficacy of MSE-Adapter. It not only optimizes parameter efficiency but also maintains the LLM's efficiency and flexibility, providing a new solution and baseline for the application of LLM in MSA and ERC.

Acknowledgements

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538.

Chen, F.; Han, M.; Zhao, H.; Zhang, Q.; Shi, J.; Xu, S.; and Xu, B. 2023. X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Guo, J.; Tang, J.; Dai, W.; Ding, Y.; and Kong, W. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM international conference on multimedia*, 3394–3402.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2109.00412*.

Hasan, M. K.; Islam, M. S.; Lee, S.; Rahman, W.; Naim, I.; Khan, M. I.; and Hoque, E. 2023. TextMI: Textualize Multimodal Information for Integrating Non-verbal Cues in Pre-trained Language Models. *arXiv preprint arXiv:2303.15430*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

He, S.; Ding, L.; Dong, D.; Zhang, M.; and Tao, D. 2022. SparseAdapter: An Easy Approach for Improving the Parameter-Efficiency of Adapters. *arXiv preprint arXiv:2210.04284*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, D.; Hou, X.; Wei, L.; Jiang, L.; and Mo, Y. 2022a. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7037–7041. IEEE.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022b. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*.

Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.-P.; Lee, R. K.-W.; Bing, L.; and Poria, S. 2023. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2304.01933*.

Kudo, T.; and Richardson, J. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2023a. GA2MIF: Graph and Attention Based Two-Stage Multi-Source Information Fusion for Conversational Emotion Detection. *IEEE Transactions on Affective Computing*.

Li, Z.; Lin, T.-E.; Wu, Y.; Liu, M.; Tang, F.; Zhao, M.; and Li, Y. 2023b. UniSA: Unified Generative Framework for Sentiment Analysis. *arXiv e-prints*, arXiv–2309.

Li, Z.; Tang, F.; Zhao, M.; and Zhu, Y. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*.

Lian, J.; Zhou, X.; Zhang, F.; Chen, Z.; Xie, X.; and Sun, G. 2018. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1754–1763.

Lin, Z.; Liang, B.; Long, Y.; Dang, Y.; Yang, M.; Zhang, M.; and Xu, R. 2022. Modeling Intra- and Inter-Modal Relations: Hierarchical Graph Contrastive Learning for Multimodal Sentiment Analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, 7124–7135. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. *arXiv preprint arXiv:2110.07602*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2023. GPT understands, too. *AI Open*.

Liu, Y.; Yuan, Z.; Mao, H.; Liang, Z.; Yang, W.; Qiu, Y.; Cheng, T.; Li, X.; Xu, H.; and Gao, K. 2022. Make Acoustic and Visual Cues Matter: CH-SIMS v2. 0 Dataset and AV-Mixup Consistent Module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 247–258.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2554–2562.

Mai, S.; Zeng, Y.; and Hu, H. 2023. Learning from the global view: Supervised contrastive learning of multimodal representation. *Information Fusion*, 100: 101920.

Mirchandani, S.; Xia, F.; Florence, P.; Ichter, B.; Driess, D.; Arenas, M. G.; Rao, K.; Sadigh, D.; and Zeng, A. 2023. Large Language Models as General Pattern Machines. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*.

Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; and Gurevych, I. 2020. AdapterHub: A Framework for Adapting Transformers. *arXiv preprint arXiv:2007.07779*.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *arXiv preprint arXiv:1810.02508*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Sun, C.; Li, Y.; Li, H.; and Hong, S. 2023a. TEST: Text prototype aligned embedding to activate LLM's ability for time series. *arXiv preprint arXiv:2308.08241*.

Sun, J.; Han, S.; Ruan, Y.-P.; Zhang, X.; Zheng, S.-K.; Liu, Y.; Huang, Y.; and Li, T. 2023b. Layer-wise Fusion with Modality Independence Modeling for Multi-modal Emotion Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 658–670.

Szot, A.; Schwarzer, M.; Agrawal, H.; Mazoure, B.; Talbott, W.; Metcalf, K.; Mackraz, N.; Hjelm, D.; and Toshev, A. 2023. Large Language Models as Generalizable Policies for Embodied Tasks. arXiv:2310.17722.

Toma, A.; Lawler, P. R.; Ba, J.; Krishnan, R. G.; Rubin, B. B.; and Wang, B. 2023. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv preprint arXiv:2305.12031*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

Wang, R.; Shivanna, R.; Cheng, D.; Jain, S.; Lin, D.; Hong, L.; and Chi, E. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In *Proceedings of the web conference 2021*, 1785–1797.

Wang, W.; Ding, L.; Shen, L.; Luo, Y.; Hu, H.; and Tao, D. 2024. WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2282–2291.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023a. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7617–7630.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNET: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yang, Z.; Fang, Y.; Zhu, C.; Pryzant, R.; Chen, D.; Shi, Y.; Xu, Y.; Qian, Y.; Gao, M.; Chen, Y.-L.; et al. 2023b. i-Code: An Integrative and Composable Multimodal Learning Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10880–10890.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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)*, 2236–2246.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2310.05804*.