2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389694

BOOSTING MODALITY REPRESENTATION WITH PRE-TRAINED MODELS AND MULTI-TASK TRAINING FOR MULTIMODAL SENTIMENT ANALYSIS

Jiarui Hai, Yu-Jeh Liu*, Mounya Elhilali†*

Laboratory for Computational Auditory Perception, Johns Hopkins University, Baltimore, USA

ABSTRACT

Sentiment analysis has traditionally leveraged information from text data. More recently, it has become increasingly clear that multimodal data provides a rich space to drastically boost interpretation of human sentiments by harnessing information across multiple modalities. In this study, we incorporate pre-trained feature extractors and propose a multi-task training strategy to improve modality representations for Multimodal Sentiment Analysis (MSA). The experimental results on the CH-SIMS v2 dataset demonstrate the superior performance of the proposed system compared to existing state-of-the-art methods, validating the effectiveness of our proposed approach. Furthermore, our framework reduces reliance on textual data, achieving competitive outcomes even when utilizing only auditory and visual modalities.

Index Terms— Sentiment Analysis, Multimodal Fusion, Transfer Learning

1. INTRODUCTION

Sentiment analysis, a rapidly evolving field in the realm of artificial intelligence, enables the interpretation of subjective information within textual data. In recent years, this discipline has taken a step forward, leading to the emergence of Multimodal Sentiment Analysis (MSA). MSA extends beyond mere textual analysis to incorporate other modalities such as audio and visual cues, thus providing a comprehensive understanding of the user's sentiment [1]. This is particularly pertinent in the context of opinion videos where different modes of expression converge to form a holistic perspective. With the boom in digital content, especially in the form of videos, understanding user sentiments is no longer confined to text but has transcended to include auditory and visual elements.

As in any emerging field, various exciting aspects of MSA are being established. In some studies, different modalities and corresponding features have been identified as useful for MSA [2, 3]. New datasets consisting diverse categories of modal data from various sources and carefully curated annotations have been published to further the efforts in MSA research [1, 4, 5, 6, 7]. Deep MSA models with state-of-the-art structures have been developed to enhance the efficiency and accuracy [8, 9, 10, 11, 12]. As performance improves, various MSA applications, including human-computer interaction (HCI), emotion recognition, educational feedback, and recommendation systems, have been introduced [13, 14].

Despite this growing interest, MSA studies remain limited in a number of ways. First, there is a predominant over-reliance on the textual modality. Some studies show a significant drop in classification accuracy when the textual modality is absent [15, 16]. This phenomenon, called text-predominance, violates the motivation behind integrating multimodal resources [4]. Nonetheless, some studies even follow the text-centric strategy to design multimodal fusion strategies [17]. While the textual modality provides rich information that directly reflects the underlying emotional state of a subject, text-predominance highly constraints the applicability of MSA systems, effectiveness and reliance on potentially noisy textual inputs. For instance, the textual modality is sometimes imperfect because of meaningless interjection words or automatic speech recognition errors. Furthermore, as stated in [5], unimodal sentiment is not always consistent with the unified multimodal sentiment.

A true multimodal approach to sentiment analysis must leverage the richness of data from multiple modalities and ensure effective integration without favoring one modality over another. The text-predominance issue in MSA involves challenges related to utilizing auditory and visual modalities, which can be summarized as follows: (1) Inadequate optimization of auditory and visual representations in MSA: Previous studies rely on low-level hand-crafted acoustic and facial features for auditory and visual information encoding [6, 15, 16], while high-level textual embeddings from pre-trained language models are widely used as textual features [5, 17]. While these hand-crafted audio and visual features are rich in elements conducive to sentiment analysis, learning sentiment cues is substantially easier from high-level textual embeddings generated by pre-trained language models [5, 15]. This discrepancy creates an imbalance in feature representation, causing MSA models to overly depend on textual features rather than considering auditory and visual cues holistically. Improved feature extraction and sentiment cues learning for auditory and visual modalities are necessary

*Indicates equal contribution.

†Authors supported by ONR N00014-23-1-2050 and N00014-23-1-2086

to address this challenge. (2) Constraints in emotional cues from auditory and visual behaviors: Limited emotional cues in auditory and visual modalities, such as neutral expressions and monotone speech, may not sufficiently contribute to overall sentiment understanding. Discordance between unified multimodal annotations and single modality sentiments further complicates the learning of visual and auditory features. Introducing unimodal annotations and incorporating subtasks for unimodal sentiment recognition can enhance the learning of emotion-laden auditory and visual representations. To address this, researchers have proposed introducing unimodal annotations and adding subtasks for unimodal sentiment recognition, enhancing the learning of emotion-laden auditory and visual information [5, 6, 18].

In this work, we propose a new perspective for multimodal sentiment analysis[1], and our contributions are summarised as follows: (1) To improve the representation of auditory and visual modalities, we use pre-trained speech and video models to operate on par with text channel for a true multimodal integration. (2) To address the issues of convergence imbalance and overfitting when jointly fine-tuning multiple large models, we propose a two-step multi-task training strategy: each feature extractor is first fine-tuned on unimodal sentiment analysis, and these feature extractors are then frozen and incorporated into the multimodal sentiment analysis system. (3) We conduct a series of ablation studies to verify the efficacy of the proposed MSA method and to investigate the factors influencing MSA performance.

2. RELATED WORK

2.1. Feature Extractors and Pre-trained Models

In a typical MSA application, the first step involves feature extraction to get essential information from raw multimodal data. In recent years, large pre-trained models have demonstrated their efficacy in providing crucial features for downstream tasks such as textual sentiment analysis, speech emotion recognition, and human action recognition. When it comes to MSA, for textual data, embeddings from pre-trained models like BERT are commonly favored over predefined features [19]. However, when dealing with auditory and visual data, a static set of hand-crafted features is usually employed for each modality. In the case of audio, Librosa and openSMILE are often utilized to extract features such as Chroma, MFCC, and PLP cepstral coefficients [20, 21]. For visual data, OpenFace has gained popularity for its ability to record facial landmarks, head poses, and gazes within videos [22, 23]. Some recent Speech Emotion Recognition studies have applied pre-trained speech models [24, 25] , but the potential of pre-trained speech and video models remains largely untapped in MSA applications.

[1]Source code: https://github.com/JHU-LCAP/BoostingMSA

2.2. Multimodal Fusion Methods

With the embeddings extracted for all modalities, the downstream task is handled by a fusion method that learns the interaction among the modalities. Depending on the task at hand, one of two possible fusion methods is generally adopted, either global-scaled, or sequential fusion methods [26]. Global-scaled fusion aims to understand multimodal interactions by observing only the global embeddings from the various modalities. One example of global-scaled fusion method applies tensor computations [26]. This approach might incorporate low-rank tensor approximations and regularization, creating expressive feature embeddings without extra training parameters. Compared with global-scaled fusion methods, sequential fusion methods based on recurrent neural networks and cross-attention transformers are utilized to simultaneously capture both global and local information to improve the efficiency of multimodal fusion [27, 28]. While these methods have exhibited excellent performance, they require additional trainable parameters and could cause performance degradation when using pre-trained models [25].

3. METHODOLOGY

3.1. Model Pipeline

As illustrated in Fig 1, the proposed MSA framework includes three main components: unimodal feature extractors, multimodal feature fusion blocks, and decision-level fully-connected blocks for sentiment predictions. In the initial phase, individual signals are processed by their corresponding feature extractors, leveraging pre-trained models. Next, the fusion module combines high-level embeddings from different modalities into a comprehensive multimodal embedding, allowing the model to capture interactions and dependencies across modalities. In the final stage, unimodal features are fed into their respective prediction blocks for unimodal sentiment analysis, while the multimodal embeddings are used by the multimodal prediction block for the multimodal sentiment prediction.

3.2. Feature Extractors

To address the text-predominant problem, we expand upon the use of pre-trained language models by incorporating pre-trained speech and video models in our MSA system. The configuration of our feature extractors is as follows:

Text: In line with [17, 5], we utilize the pre-trained BERT model for processing transcriptions. Each transcription, including a [CLS] token designed to learn the global context of the sentence, is first tokenized by the BERT tokenizer and then fed to the BERT model to obtain sequential embeddings. We employ the last hidden layer of BERT for both unimodal and multimodal sentiment analysis.

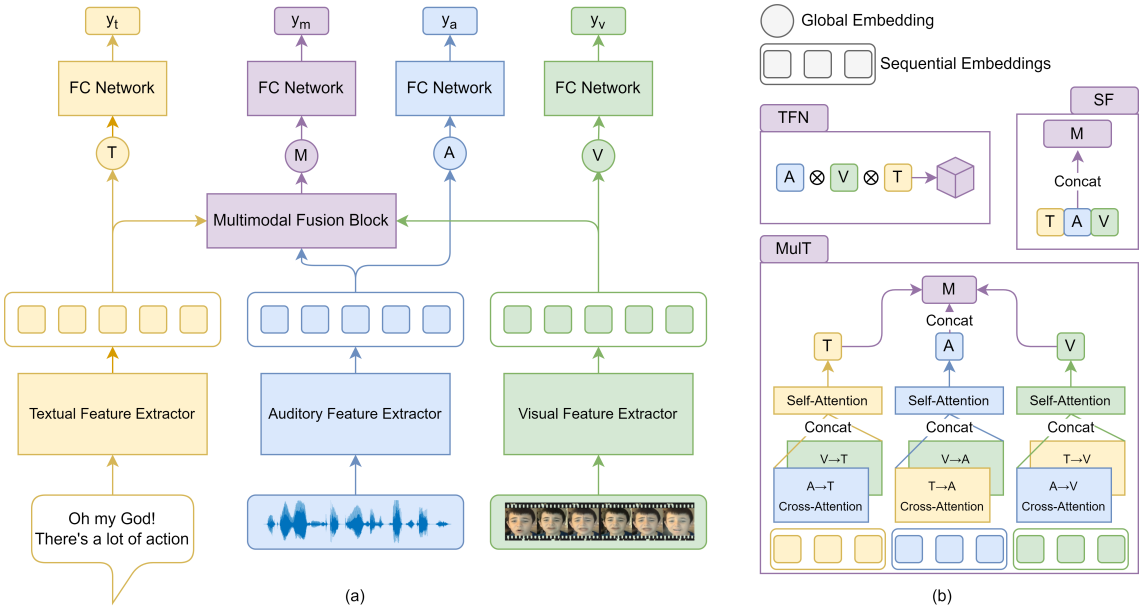

Fig. 1. (a) The framework of multimodal multi-task MSA. FC network denotes the fully connected neural network. The variables y_k, where $k \in m, t, a, v$, represent the sentiment scores for each unimodal or multimodal task. (b) Three fusion methods used in this study.

Audio: We deploy HuBERT, a large-scale pre-trained speech model [29] with demonstrated efficacy in speech emotion recognition [30], as our auditory feature extractor. Audio waveforms are resampled to 16kHz and padded and then processed by the pre-trained speech model. Similar to the text model, we use the embeddings from the last layer. To create global embeddings for the entire audio clip, we apply average time pooling to the sequential embeddings.

Visual: Since facial expressions are usually considered to have rich emotional information and there could be multiple speakers in the video scenes [5], we first employ TalkNet [31] as an Active Speaker Detection (ASD) tool to extract the speaker's facial video within each video clip. We then use the Video Swin Transformer [32], pre-trained on the Kinetics dataset [33], as our visual feature extractor. This choice is driven by the Kinetics dataset's diversity in human-centric actions, including facial expressions like laughing and crying. We opt for a 128×128 resolution for the facial video and a frame rate of 10 fps, balancing performance and computational demands. To extract time sequence embeddings and global embeddings for the entire video clip, we apply average spatial and temporal pooling, respectively.

3.3. Multimodal Fusion Network

Fusion mechanisms have been recognized as a crucial component in previous MSA frameworks. In the proposed system, we explore the performance of three representative fusion strategies when using the proposed feature extractors.

SF: Concatenation Fusion, referred to as Shallow Fusion due to its simplicity, is a basic strategy that concatenates global features from different modalities. It represents one of the most straightforward fusion methods, extensively used in various multimodal fusion studies. In our MSA system utilizing shallow fusion, we concatenate the global embeddings from the three feature extractors to form the multimodal feature, which is then passed to the subsequent prediction block.

TFN: The Tensor Fusion Network [8] explicitly models both view-specific and cross-view dynamics by creating a multidimensional tensor based on the outer product, which can capture interactions across one, two, or all three modalities. Like Shallow Fusion, TFN does not introduce any additional trainable parameters.

MulT: The Multimodal Transformer [28] employs directional pairwise cross-modal attention to facilitate frame-level interaction among the three modalities. This approach encourages interaction between two multimodal sequences across different time steps, utilizing pairwise cross-modal attention to subtly adapt one modality's streams to another. In the case of fusing three modalities, the Multimodal Transformer requires six bimodal cross-attention blocks and three unimodal self-attention blocks, which results in a large number of additional trainable parameters compared to both SF and TFN.

3.4. Multi-task Learning

In our approach, besides the primary unified multimodal sentiment prediction, we utilize three unimodal sentiment pre-

diction subtasks connected to each feature extractor for joint optimization of feature extractors. These subtasks aim to enhance the MSA system's ability to recognize fine-grained sentiments, while also assisting large pre-trained models to better adapt to the scenarios in which unimodal and multimodal sentiments are not consistent and prevent overfitting. We use the L_1 loss for supervision for both the unimodal tasks and the main multimodal task as follows:

$$L_k = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_k^i - y_k^i| \tag{1}$$

where $k \in m, t, a, v$, N is the number of supervised instances, y is the ground truth sentiment, and $\hat{y}$ is the model prediction.

When it comes to the joint multi-task training, the three subtasks and the main task are trained simultaneously. The final sentiment regression loss is formulated as the weighted average of the unimodal and multimodal tasks:

$$L_= \sum_k \alpha_k L_k \tag{2}$$

where α is the hyper-parameter controlling the contribution of unimodal and multimodal tasks.

However, directly fine-tuning large pre-trained models through joint multi-task learning can lead to convergence imbalance and overfitting which can limit the performance of fusion methods with additional trainable parameters. To overcome these challenges, we propose a two-step fine-tuning strategy. In the first stage, the pre-trained models are individually fine-tuned for their respective unimodal sentiment analysis tasks. In the second stage, the pre-trained models are frozen and integrated into the fusion model for multimodal sentiment analysis, with only the parameters of the fusion block and the fully connected block for the main task being updated.

4. EXPERIMENTS

4.1. Datasets

We leverage two datasets to evaluate the proposed MSA system and compare it with existing state-of-the-art methods.

4.1.1. MOSI

The MOSI dataset introduced by Zadeh et. al. [6] contains videos sourced from the YouTube platform. The videos have varying lengths, ranging from 2 to 5 minutes, and the dataset contains a total of 93 videos. These videos are mainly videos uploaded by YouTube users to express their opinions on different topics. Despite having diverse ethnic backgrounds, all speakers communicated in English. Along with the raw clips, each of the clip underwent manual transcription to extract spoken words. Finally, a single sentiment label linearly ranging from -3 to +3 is given to each clip by averaging human annotators' responses.

4.1.2. CH-SIMS v2

Liu et. al. [4] introduced the second version of the CH-SIMS dataset. It contains 4402 supervised and 10161 unsupervised video segments. The average length of the segments is around 4.4 seconds. The sentiment annotations are ranging from -1 to +1. Unlike the MOSI dataset, which only assigns one multimodal sentiment label per video clip, CH-SIMS v2 provides additional sentiment labels corresponding to the three modalities, enabling multi-task learning. These videos, drawn from various Chinese TV shows, offer diverse content, characters, and background scenes, making the dataset ideal for validating the strength of the proposed MSA system and conducting ablation studies.

4.2. Experimental Details

All experiments are conducted on a single NVIDIA RTX 3090 GPU. Models are trained using the Adam optimizer with learning rates set to 10^{-5} for pre-trained feature extractors and 10^{-4} for the other components. The checkpoints for BERT and HuBERT are from HuggingFace [34], and the checkpoint for the Video Swin Transformer is from [32].

To evaluate MSA frameworks, results are analyzed in both classification and regression tasks. For classification, binary classification accuracy (Acc2) and F1 score are used to gauge the accuracy of basic sentiment polarity prediction, i.e., positive or negative classification. For the CH-SIMS v2 dataset, we additionally utilize Acc2_weak to evaluate model performance with weak emotion instances labeled within the range [-0.4, 0.4]. For regression tasks, we report the mean absolute error (MAE) and Pearson correlation (Corr).

4.3. Performance Comparisons

We evaluate the performance impact of different feature extractors by grouping state-of-the-art MSA frameworks into two categories: those using baseline feature extractors and those employing our proposed feature extractors. In the baseline systems, BERT handles textual modality, while handcrafted features as outlined by [5, 35] are used for auditory and visual modalities. In our proposed system, we employ BERT, HuBERT, and Video Swin Transformer for textual, auditory, and visual modalities respectively. The results on the CH-SIMS and MOSI datasets are demonstrated in Table 1 and Table 2.

Here, we consider two specific frameworks: MAG-BERT [36] and AV-MC [4]. MAG-BERT integrates a Multimodal Adaptation Gate (MAG) with BERT to enable multimodal information fusion. While it exhibits top-level performance, it also has a text-predominance issue as it overlooks the roles

Table 1. Performance comparison of multimodal sentiment analysis on CH-SIMS v2.0 dataset.

Fusion Method	Subtask	Acc2↑	Acc2_weak↑	F1↑	MAE↓	Corr↑
Baseline Feature Extractors: BERT[T] Hand-crafted[A] Hand-crafted[V]						
SF	Single-task	73.95	69.13	73.84	0.381	52.19
TFN	Single-task	76.51	66.27	76.31	0.323	66.65
MulT	Single-task	79.50	69.61	79.59	0.317	70.32
MAG-BERT	Single-task	79.79	71.87	79.59	0.334	69.09
SF	Joint Multi-task	78.04	71.59	78.44	0.326	65.80
TFN	Joint Multi-task	80.26	71.07	80.33	0.318	70.54
MulT	Joint Multi-task	82.76	73.41	82.51	0.293	70.32
AV-MC	Mix-up + Joint Multi-task	82.50	74.54	82.55	0.297	73.17
AV-MC (Semi)	Mix-up + Joint Multi-task	83.46	74.54	83.52	0.286	76.04
Proposed Feature Extractors: BERT[T] HuBERT[A] Swin[V]						
SF	Single-task	82.12	73.00	82.09	0.287	72.63
TFN	Single-task	82.12	73.26	82.14	0.288	71.63
MulT	Single-task	82.02	72.23	81.98	0.279	76.03
SF	Joint Multi-task	**86.06**	**78.14**	**85.99**	0.252	78.35
TFN	Joint Multi-task	85.95	77.37	85.94	**0.240**	**80.69**
MulT	Joint Multi-task	83.72	75.06	83.68	0.272	73.14
SF	**Two-step Multi-task**	86.17	79.69	86.17	0.252	78.28
TFN	**Two-step Multi-task**	86.80	78.66	86.77	0.249	79.98
MulT	**Two-step Multi-task**	**87.02**	**80.20**	**87.01**	**0.246**	**80.05**

Table 2. Performance comparison of multimodal sentiment analysis on MOSI dataset.

Fusion Method	Acc2↑	F1↑	MAE↓	Corr↑
Baseline Features & Single Task Training				
SF	79.39	79.45	0.945	67.5
TFN	78.02	78.09	0.971	65.2
MulT	80.21	80.22	0.912	69.5
MAG-BERT	83.41	83.47	**0.761**	**77.6**
Proposed Features & Single Task Training				
SF	**83.53**	**83.55**	0.785	75.8
TFN	82.01	81.99	0.866	69.9
MulT	80.64	80.72	0.905	66.7

of auditory and visual modalities during the fusion stage [17]. AV-MC, on the other hand, employs data augmentation via mix-up [37] for hand-crafted audio and visual features, and its variant AV-MC (Semi) further incorporates semi-supervised data for pre-training. Despite its use of a straightforward shallow fusion strategy, it represents a state-of-the-art MSA framework for the CH-SIMS v2 dataset.

Table 1 outlines the results of the experiments conducted on the CH-SIMS v2 dataset. The key observations include: (1) In both multi-task and single-task scenarios, models with our proposed feature extractors significantly outperform systems using baseline feature extractors. Particularly, MAG-BERT exhibits inferior performance due to the inability to utilize multimodal annotations. Although AV-MC achieves some improvements through data enhancement, it still lags behind the proposed method. (2) Under the single-task or joint multi-task training, global-scaled fusion methods SF and TFN perform better with pre-trained models while sequential fusion method MulT excels with handcrafted features. (3) The two-step multi-task training, proposed for the pre-trained feature extractor, further improves the MSA performance. In this scenario, MulT outperforms SF and TFN. (4) Systems engaged in multi-task settings, whether joint or two-step, perform superior to those in single-task settings. (5) Multi-task training also improves Acc2_weak scores and regression performance, signifying the importance of unimodal annotations for fine-grained sentiment analysis.

Table 2 presents the experimental results on the MOSI dataset. As the MOSI dataset does not supply unimodal sentiment annotations, the systems are trained using multimodal sentiment annotations. The system integrating our proposed feature extractors with Shallow Fusion demonstrates marginally superior classification performance compared to benchmarks but falls short in terms of regression performance. Given the experiments conducted on the CH-SIMS v2 dataset, we believe the performance can be enhanced further if unimodal annotations were accessible.

5. ABLATION STUDIES AND DISCUSSIONS

5.1. Comparison of Feature Extractors

Tables 1 and 2 highlight that by employing our proposed feature extractors—pre-trained models for all modalities—we achieve superior MSA performance compared to conven-

Table 3. Results for the ablation study on feature extractors for unimodal sentiment analysis on CH-SIMS v2.0 dataset.

Feature Extractor		Acc2↑	F1↑	MAE↓	Corr↑
Textual Input → Textual Sentiment					
❄	BERT	88.48	88.50	0.240	78.26
	BERT	90.06	90.07	0.228	79.91
Auditory Input → Auditory Sentiment					
	openSMILE	60.09	59.67	0.425	23.17
❄	HuBERT	72.04	72.05	0.329	51.71
	HuBERT	79.45	79.44	0.277	65.85
Visual Input → Visual Sentiment					
	OpenFace	78.02	77.90	0.312	58.73
❄	Swin3D	73.52	73.43	0.353	46.09
	Swin3D	86.76	86.73	0.247	74.72

tional baseline feature extractors. This reflects a general enhancement in the performance of all the frameworks. To verify the contribution of each modality, we conduct experiments on each individual modality subnet.

Table 3 exhibits the performance increase for each modality when employing our proposed feature extractors. Here, the snowflake symbol ❄ indicates a pre-trained model of which the parameters remain frozen. A pre-trained model without the snowflake symbol signifies that it undergoes fine-tuning. For all modalities, the fine-tuned pre-trained models yield the best results. Specifically, for the audio and visual modalities, there are notable improvements in both classification and regression scores when a fine-tuned pre-trained model is used instead of hand-crafted features.

5.2. Comparison of Unimodal, Bi-modal and Tri-modal Sentiment Analysis

In this study, all MSA frameworks tested in Tables 1 and 2 are tri-modal, utilizing data from three modalities. However, in scenarios where comprehensive multimodal data is not available, the performance of existing MSA frameworks could be significantly impacted, particularly when textual data is absent. The feature extraction strategy we propose aim to address this issue by improving the representation of auditory and visual modalities to approximate the textual representation. We evaluate the MSA framework using various combinations of unimodal, bi-modal, and tri-modal features. The feature extractor undergoes direct fine-tuning for unimodal cases, while two-step multi-task training and the MulT fusion are applied for bi-modal and tri-modal cases.

The results in Table 4 demonstrate that tri-modal features achieve the best performance, indicating the advantages of utilizing multiple modalities in MSA. Even though the textual modality plays an important role in MSA, the absence of textual modality does not lead to a huge decrease in performance. This suggests that the text-predominance issue can be alleviated by incorporating the proposed feature extractors.

Table 4. Unified multimodal sentiment analysis using unimodal, bi-modal and tri-modal features on CH-SIMS v2.0 dataset.

Modality	Acc2↑	F1↑	MAE↓	Corr↑
T → M	80.00	79.88	0.339	64.66
A → M	78.72	78.50	0.317	64.76
V → M	78.29	78.02	0.340	62.64
T+A → M	83.51	83.50	0.283	70.95
T+V → M	85.31	85.33	0.261	77.96
A+V → M	81.80	81.79	0.302	69.80
T+A+V → M	87.02	87.01	0.246	80.05

5.3. Effectiveness of Multi-task Training

Table 1 highlights the benefits of multi-task training for MSA systems. The benefits of multi-task training over single-task training are evident in both systems using hand-crafted features and those using pre-trained models. In addition to being able to significantly improve binary accuracy, multi-task training brings significant improvement to the regression metrics, which implies that multi-task has an important role for fine-grained sentiment recognition.

5.4. Comparison of Multimodal Fusion Strategies with Pre-trained Feature Extractors

As discussed in Section 4.3, under single-task and joint multi-task training with the proposed feature extractors, frameworks employing global-scaled fusion methods exhibit superior performance compared to systems utilizing sequential fusion methods. However, in the case of two-stage multi-task training where the pre-trained models are frozen, the sequential fusion method MulT provide better performance. This finding suggests that MulT's additional trainable parameters render the model susceptible to overfitting when combined with trainable pre-trained feature extractors. Nevertheless, employing the two-step multi-task training approach, which freezes the fine-tuned feature extractors during fusion model training, helps mitigate overfitting and unlocks the full capabilities of MulT.

6. CONCLUSIONS

The proposed framework effectively improves auditory and visual feature representations and achieves state-of-the-art performance on the CH-SIMS v2 dataset. Key findings from our experiments include the performance gains obtained by using the proposed feature extractors, the superiority of multi-task training over single-task training, and the improvement led by the two-stage multi-task training strategy when employing pre-trained feature extractors. In the future, we would like to design efficiency fine-tuning method for pre-trained feature extractors on datasets without unimodal annotations.

7. REFERENCES

[1] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[2] Angeliki Metallinou, Athanasios Katsamanis, Martin Wöllmer, Florian Eyben, Björn Schuller, and Shrikanth S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, pp. 184–198, 2012.

[3] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE intelligent systems*, vol. 28, no. 3, pp. 38–45, 2013.

[4] Yih-Ling Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao, "Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module," *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022.

[5] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Annual Meeting of the Association for Computational Linguistics*, 2020.

[6] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *ArXiv*, vol. abs/1606.06259, 2016.

[7] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *ArXiv*, vol. abs/1810.02508, 2018.

[8] Amir Zadeh, Minghai Chen, Soujanya Poria, E. Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," in *Conference on Empirical Methods in Natural Language Processing*, 2017.

[9] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, pp. 153–163, 2013.

[10] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, E. Cambria, and Soujanya Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl. Based Syst.*, vol. 161, pp. 124–133, 2018.

[11] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] Wei Han, Hui Chen, Alexander F. Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

[13] Zhongmei Han, Jiyi Wu, Changqin Huang, Qionghao Huang, and Meihua Zhao, "A review on sentiment discovery and analysis of educational big-data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, 2019.

[14] Agnieszka Rozanska and Michal Podpora, "Multimodal sentiment analysis applied to interaction between patients and a humanoid robot pepper," *IFAC-PapersOnLine*, 2019.

[15] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[16] Xia Li and Minping Chen, "Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language," in *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*. Springer, 2020, pp. 359–373.

[17] Xianbing Zhao, Yixin Chen, Wanting Li, Lei Gao, and Buzhou Tang, "Mag+: An extended multimodal adaptation gate for multimodal sentiment analysis," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4753–4757, 2022.

[18] Lijun He, Ziqing Wang, Liejun Wang, and Fan Li, "Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python.," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[21] Florian Eyben, Martin Wöllmer, and Björn Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*. ACM, 2010, pp. 1459–1462.

[22] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "Openface: An open source facial behavior analysis toolkit," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

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

[24] Wei Yang, Satoru Fukayama, Panikos Heracleous, and Jun Ogata, "Exploiting fine-tuning of self-supervised learning models for improving bi-modal sentiment analysis and emotion recognition," in *Proc. Interspeech*, 2022, vol. 2022, pp. 1998–2002.

[25] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," 2020.

[26] Lijun He, Ziqing Wang, Liejun Wang, and Fan Li, "Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

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

[28] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[29] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[30] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[31] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935.

[32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[33] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[35] Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao, "M-sena: An integrated platform for multimodal sentiment analysis," *arXiv preprint arXiv:2203.12441*, 2022.

[36] Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque, "Integrating multimodal information in large pretrained transformers," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020, pp. 2359–2369, 2020.

[37] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.