Information Processing and Management 60 (2023) 103229

Contents lists available at ScienceDirect

Information Processing and Management

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

PS-Mixer: A Polar-Vector and Strength-Vector Mixer Model for Multimodal Sentiment Analysis

Han Lin [a,1], Pinglu Zhang [a,1], Jiading Ling [a], Zhenguo Yang [a,*], Lap Kei Lee [b], Wenyin Liu [a]

[a] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[b] School of Science and Technology, Hong Kong Metropolitan University, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Multimodal fusion
Deep learning
Sentiment classification
MLP

ABSTRACT

Multimodal sentiment analysis aims to judge the sentiment of multimodal data uploaded by the Internet users on various social media platforms. On one hand, existing studies focus on the fusion mechanism of multimodal data such as text, audio and visual, but ignore the similarity of text and audio, text and visual, and the heterogeneity of audio and visual, resulting in deviation of sentiment analysis. On the other hand, multimodal data brings noise irrelevant to sentiment analysis, which affects the effectness of fusion. In this paper, we propose a Polar-Vector and Strength-Vector mixer model called PS-Mixer, which is based on MLP-Mixer, to achieve better communication between different modal data for multimodal sentiment analysis. Specifically, we design a Polar-Vector (PV) and a Strength-Vector (SV) for judging the polar and strength of sentiment separately. PV is obtained from the communication of text and visual features to decide the sentiment that is positive, negative, or neutral sentiment. SV is gained from the communication between the text and audio features to analyze the sentiment strength in the range of 0 to 3. Furthermore, we devise an MLP-Communication module (MLP-C) composed of several fully connected layers and activation functions to make the different modal features fully interact in both the horizontal and the vertical directions, which is a novel attempt to use MLP for multimodal information communication. Finally, we mix PV and SV to obtain a fusion vector to judge the sentiment state. The proposed PS-Mixer is tested on two publicly available datasets, CMU-MOSEI and CMU-MOSI, which achieves the state-of-the-art (SOTA) performance on CMU-MOSEI compared with baseline methods. The codes are available at: https://github.com/metaphysicser/PS-Mixer.

1. Introduction

With the popularity of social platforms (e.g. YouTube, Facebook), the task of sentiment analysis is now not limited to unimodal data but has been extended to multimodal data consisting of multiple sources of information, including visual, audio, and text. The ability to extract users' sentiment in multimodal data can help decision-makers to understand the past, predict the future, and make the right decisions. These sentiments can be broadly classified as positive, negative, or neutral. Multimodal sentiment analysis (MSA) (Balahur, Montoyo, Martínez-Barco, & Boldrini, 2012; Poria, Hazarika, Majumder, & Mihalcea, 2020) is useful in many aspects

* Corresponding author.
E-mail addresses: 2112105225@mail2.gdut.edu.cn (H. Lin), 3119004985@mail2.gdut.edu.cn (P. Zhang), 2112005117@mail2.gdut.edu.cn (J. Ling), yzg@gdut.edu.cn (Z. Yang), lklee@hkmu.edu.hk (L.K. Lee), liuwy@gdut.edu.cn (W. Liu).
[1] Lin Han and Zhang Pinglu contribute equally to this paper.

https://doi.org/10.1016/j.ipm.2022.103229
Received 14 June 2022; Received in revised form 11 October 2022; Accepted 6 December 2022
Available online 21 December 2022

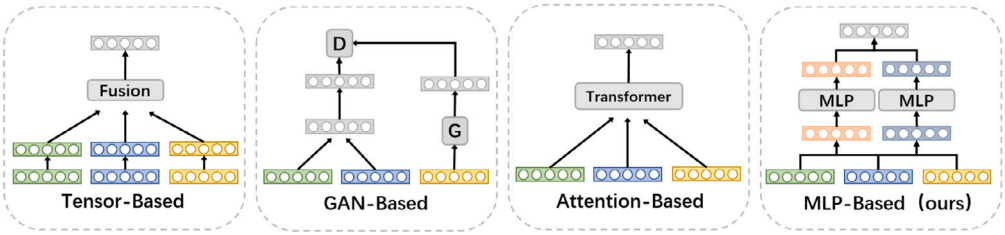

Fig. 1. Several approaches to multimodal fusion (Tensor-based, GAN-based, Attention-based and ours).

of life, including politics (Abbasi, Chen, & Salem, 2008), stock market prediction (Bollen, Mao, & Zeng, 2011), movie box office revenue prediction (Romero, Galuba, Asur, & Huberman, 2011), customer feedback (Gamon, Aue, Corston-Oliver, & K.Ringger, 2005) and so on. By analyzing visual, audio, and text from people, it is possible to understand human sentiment communication and open the way for more humanization in artificial intelligence (Martinez-Miranda & Aldea, 2005; Rubin & Kenneth, 1998).

Since the multimodal data is heterogeneous from one modality to another, it proposes a challenge that how to make full use of the complementary information in each modality to work together for sentiment analysis. Although multimodal data can help analyze users' real sentiment from different perspectives such as visual perception, auditory rhythm, and so on, avoid the information singularity of unimodal data, and improve modeling effect, the use of multimodal data also introduces heterogeneity between different modal data, which increases the difficulty of semantic understanding and destroys the semantic integrity. Secondly, the multimodal data brings a lot of noise unrelated to specific scenes and result in spending more effort on distinguishing the input data, which increases the calculation amount of model training and reduces the efficiency of the model.

Existing methods can be divided into early fusion, late fusion, and hybrid fusion (Lecun, Bengio, & Hinton, 2015) according to different fusion stages. The approaches of early fusion fuse the features immediately after the feature extraction. In Zadeh, Chen, Poria, Cambria, and Morency (2017), the features of visual, audio, and text modality were first extracted and then the fusion vectors for decision making were directly obtained by calculating the 3-fold Cartesian product. However, early fusion of features is not only unable to fully fuse the useful information between different modalities but also brings a lot of redundant information, which affects the learning efficiency of the model. Therefore, the Principal Component Analysis (PCA) (Sun, Wang, Xu, Zhang, & Balezentis, 2022) (Liu, Gui, Xiong, & Zhan, 2021) method is used to reduce the dimensionality of the data and remove the redundant information from the data. In late fusion that is not affected by the distribution of the original data, the data of each modality are first computed by their own unique pre-trained models and then the results of each model are fused into a final vector. For example, Yu, Xu, Yuan, and Wu (2021) proposed a label generation module based on self-supervised learning to obtain labels for each individual modality named self-mm, and a weight-adjustment strategy was designed to guide the subtask to focus on the more different sampling between modalities. Late fusion is a better solution of fusing semantic information than early fusion, but the model structure becomes more complex. Hybrid fusion combines early fusion and late fusion, allowing features to be fused at multiple stages of the model. For example, in Sahu and Vechtomova (2021), the features of any two modalities (e.g., video and audio) were first fused and then the features of the remaining modalities (e.g., text) were trained to obtain the final fusion vector. Kumar and Vepa (2020) proposed a learnable gating mechanism to selectively learn cross attended features, which can control the transfer and discard of information and use a self-attention mechanism to capture long-term context. Moreover, the vectors obtained by gating mechanisms were fitted in a deep multimodal fusion module to obtain the final vectors for the analysis of sentiment states. It can significantly improve the performance of the model by planning the stage of the fusion in the model.

As shown in Fig. 1, there are many fusion mechanisms that can be used to solve the heterogeneity challenges of multimodal data, such as based on tensor fusion, based on Generative Adversarial Network (GAN), based on attention mechanism and so on. For example, Zadeh et al. (2017) devised an early approach based on tensor fusion named TFN. It performed Cartesian product for fusing the features of three modalities after dimensional expansion, and combined the features obtained from multimodal fusion with the respective unimodal features for decision making which could retain not only each modality information but also all modalities information. Although this method was relatively simple, it faced a huge amount of computation and memory consumption and the number of model parameters also increases, which is prone to the risk of overfitting. Later, Liu et al. (2018) proposed a low-rank multimodal fusion method called LMF by decomposing the tensors and weights in parallel to solve the problems of TFN, which could reduce the number of parameters while improving the computational speed. Even though LMF has solved some problems of TFN, LMF still led to the problem of parameter dramatic growth once the input feature dimension became large. Later methods based on the GAN mechanism Mai, Hu, and Xing (2020) Tsai, Liang, Zadeh, Morency, and Salakhutdinov (2019) abandoned the direct computation of features and adopted the idea of the two-player game, using the mutual confrontation of generator and discriminator to gradually bring different modal features closer to each other. Mai, Hu, and Xing (2020) proposed a novel generative adversarial network (GAN) based on a layered graphical neural network to achieve unification by learning the embedding space invariant with modality and converting the distribution of the source modality to the distribution of the target modality. Although GAN exploits reconstruction loss and classification loss to impose constraints on the embedding space, the possible collapse in training will lead to unstable training. Later in the field of NLP, due to Transformer (Vaswani et al., 2017) used the multi-head attention mechanism to improve the stable of model training, Zadeh, Liang, Mazumder, et al. (2018), Wang et al. (2019) Tsai, Bai, et al. (2019) and Su, Hu, Li, and Cao (2020) tried to apply the attention mechanism to multimodal fusion. Zadeh, Liang, Mazumder, et al. (2018) proposed

the Memory Fusion Network (MFN) approach to process multimodal sequence data for view-specific interactions and cross-view interactions by interacting the information in the Delta-memory Attention Network (DMAN) module and Multi-view Gated Memory module. Despite the progress made in the above approaches, these fusion methods were still subject to complex challenges due to the heterogeneity between different modalities.

Recently, there are a number of studies focusing on multi-layer perceptron (MLP), such as MLP-Mixer (Tolstikhin et al., 2021), S^2-MLP (Yu, Li, Cai, Sun, & Li, 2022), S^2-MLPv2 (Yu, Li, Cai, Sun, & Li, 2021), hire-MLP (Guo et al., 2022), ResMLP (Touvron et al., 2021), CycleMLP (Chen, Xie, Ge, Liang, & Luo, 2022), gMLP (Liu, Dai, So, & Le, 2021), VIP (Hou et al., 2021) and so on, which confirmed that MLP can be comparable to the Transformer. Tolstikhin et al. (2021) firstly designed an architecture based on multi-layer perceptrons (MLPs) named MLP-Mixer which consists of two parts. The first part applied MLPs to image patches individually in order to mix features at each position, the other part used MLPs to cross patches, which could mix spatial information. In Chen, Xie et al. (2022), a CycleMLP Block to realize local features by aligning features at different spatial locations to the same channel was designed, which achieved an accuracy of 83.2% and outperformed the Transformer-based models. Nie et al. (2021) removed multi-head attention from the Transformer and replace it with MLP, then fed the bimodal data (visual and text) into a pure MLP framework to achieve a similar result as the Transformer after pre-training, which successfully confirmed the feasibility of MLP on fusing multimodal data. The positive effect of MLP in the multimodal field encourages us to apply MLP to multimodal sentiment analysis.

In this paper, we propose a Polar-Vector and Strength-Vector mixer model (PS-Mixer) based on MLP for multimodal sentiment analysis to achieve better communication between different modal data. Specifically, we design a Polar-Vector (PV) that determines the polarity of the sentiment including positive, negative and neutral, and a Strength-Vector (SV) that decides the strength value of sentiment between 0 and 3. In addition, we devise the MLP-Communication module that is able to communicate the input features in both vertical and horizontal directions to reduce the interference of noise and facilitate multimodal interactions. Finally, we propose a polar loss to determine sentiment direction and a strength loss to judge sentiment strength. Our experiments show that PS-Mixer has reached the SOTA result on the CMU-MOSEI dataset, demonstrating the competitiveness of PS-Mixer on the task of multimodal sentiment analysis.

The main contributions of this paper are as follows.

1. We propose a Polar-Vector and Strength-Vector mixer model called PS-Mixer, which is based on MLP-Mixer, to communicate between different modalities features. With our designed MLP-Communicator (MLP-C), features can be communicated and interacted in both vertical and horizontal directions. This is the first time that MLP has been used for a trimodal (visual, audio and text) sentiment analysis task.
2. We design two judgments Polar-Vector (PV) and Strength-Vector (SV) to work together for the decision of sentiment prediction. The PV indicates positive, negative or neutral sentiment. The SV expresses the sentiment strength in the range of 0 to 3.
3. We propose three loss functions: polar loss, strength loss and task loss to make the predicted sentiment more accurate according to the multimodal data by measuring the gap between the predicted value of sentiment polarity, strength and the true value.
4. We conduct extensive experiments and achieve SOTA performance for multimodal sentiment analysis compared with baseline methods, especially on the seven classification task of CMU-MOSEI dataset with the accuracy of 86.1%. The experimental result shows that the number of parameters of our proposed MLP-C module is 1.1M less than the Transformer module.

The remainder of this paper is organized as follows. Section 2 introduces sentiment analysis, multimodal feature extraction and multi-layer perceptron. Section 3 explains the details of the proposed method. Section 4 presents the details of the experiments and the evaluation results. Section 5 concludes the work.

2. Related work

The dramatic growth of multimodal data has led to the development of multimodal sentiment analysis. In this section, we review the literature about sentiment analysis from unimodal to multimodal and analyze the methods they use. Then we introduce the related work on multimodal feature extraction. Finally, we analyze the relevant background of the MLP field.

2.1. Sentiment analysis

Sentiment analysis is an important area of research in deep learning. From the beginning of the research to the present, sentiment analysis task has expanded from unimodal to multimodal.

Initially, the sentiment analysis task focused on unimodal sentiment analysis such as textual sentiment analysis. Meng et al. (2012) proposed a generative cross-lingual mixture model (CLMM) to maximize the likelihood of bilingual parallel data by parameter fitting and determine the polarity of sentences in the parallel corpus by using words in the source and targeted languages. Li, Pan, Jin, Yang, and Zhu (2012) designed a dual-domain adaptation framework to extract the exact words in the target domain without annotation by generating some high-confidence sentiment in the target domain and proposed a Relational Adaptive bootstrapping (RAP) algorithm to extend the seeds of the target domain by using labeled source domain data and the relationship between topic and sentiment words. Jiang, Yu, Zhou, Liu, and Zhao (2011) developed an improved method based on target-dependence and context-awareness to improve the performance of sentiment classification for tweet sentiment classification, especially for very short and

ambiguous tweets. Although unimodal data could express the emotional state, it could not accurately predict the emotional state due to the lack of multiple perspectives.

The visual and audio information in the video can also provide a rich sentiment state. Zhao et al. (2020) proposed a deep Visual Audio Attention Network (VAANet) to guide attention generation by integrating spatial, channel-wise and temporal attentions into 3D CNNs and 2D CNNs. Zhu, Chen, and Wu (2019) designed a noise cancellation framework based on a quality embedding network to derive the corresponding stochastic gradient descent (SGD) optimization objective with variational inference and conditional independence assumption, which can be generalized to other multimodal problems with labeled noise. Wang, Wu, and Hoashi (2019) devised a multiple attention fusion network (MAFN) consisting of two attention mechanisms by modeling human sentiment recognition mechanisms to dynamically extract representative sentiment features and automatically highlight different modal features according to their importance. Ghaleb, Popa, and Asteriadis (2020) proposed a Multimodal Emotion Recognition Metric Learning (MERML) network to capture the complex relationships between two modalities and learn the potential space by learning modality-specific metrics together for audio–video emotion recognition tasks is also scalable framework. Han, Zhang, Ren, and Schuller (2021) presented a new cross-modal emotion embedding framework (EmoBed) to improve the performance of existing sentiment recognition systems for exploring the knowledge of other auxiliary modalities by exploring the underlying semantic sentiment information under a shared recognition network and a shared sentiment embedding space. In contrast, bimodal data are more conducive to increased data richness than unimodal data.

Some videos contain textual information that can also be involved in the analysis of sentiment. Arjmand, Dousti, and Moradi (2021) proposed a Transformer-Based Speech-Prefixed Language Model (TEASEL) for multimodal sentiment analysis which implemented a Lightweight Attentive Aggregation module to generate an efficient spatial encoding. TEASEL could achieve the same level of performance as spending a long time retraining the Transformer without training the full Transformer. Shenoy and Sardana (2020) designed a recurrent neural network architecture for sentiment analysis and sentiment detection in conversation, using a state GRU (sGRU) to model the interlocutor's state, a context GRU (cGRU) to track the context of the conversation, an emotion GRU (eGRU) to track the participant's sentiment state and a pairwise attention mechanism to combine related states for sentiment prediction. The human perception model emphasizes the importance of top-down integration, i.e. cognition affects perception. Paraskevopoulos, Georgiou, and Potamianos (2022) proposed a feedback module named MMLatch that allowed modeling top-down cross-modal interactions between higher and lower level architectures. The architecture used a feedback mechanism in forward propagation to capture top-down cross-modal interactions during network training and extracted high-level representations of each modality, to mask sensory inputs. Hazarika, Zimmermann, and Poria (2020) designed a framework named MISA for learning mode-invariant and mode-specific representations by projecting each modality into two different subspaces to reduce the gap between modalities and capture modality. This method of learning modal invariant and modal characteristics was beneficial to the full exploitation of data from different modalities. Delbrouck, Tits, Brousmiche, and Dupont (2020) devised a Transformer-based model for sentiment analysis named TBJE which relied exclusively on attentional mechanisms and feedforward neural networks (FFN) to map global dependencies between inputs and outputs. Sahu and Vechtomova (2021) proposed an adaptive fusion technique and two networks named automatic fusion network and GAN fusion network that aimed to efficiently model context from different modalities. Two proposed networks could learn to compress information from different modalities while preserving the context and regularized the learning potential space of a given context from complementary modalities. Compared with existing methods, the lightweight adaptive networks could better model the context in different modalities. Lian, Liu, and Tao (2021) devised a multimodal learning framework for conversational sentiment recognition called conversational Transformer network (CTNet), which used Transformer-based structure to model the interactions between multimodal features within and across modalities and a bidirectional GRU component to model the bidirectional dependencies of context-sensitive and speaker-sensitive.

2.2. Feature representation

As the raw data is often high dimensional and contains a lot of redundant information, the raw information can be very sparse. If the raw data is fed into the model, it will result in a huge amount of calculations and inefficient model training. Therefore, the raw data needs to be extracted to feature before being fed into the model. Feature extraction is the process of reducing the dimensionality of the original input data and recombining the features from the original data to facilitate subsequent tasks. Feature extraction can be used to solve the following problems as mentioned by Guyon, Nikravesh, Gunn, and Zadeh (2006). Firstly, the original data is highly dimensional. Secondly, the original data contains too much redundant information. Thirdly, the original data is too sparse. In the following, different feature extraction methods are applicated for different types of data.

1) Text

Since text is a natural language used by humans and lacks understanding by computers, features need to be extracted from text for representing natural language. The traditional method of text feature extraction is one-hot encoding (Lucas, 2014), which was a method for converting a word vector into a binary number by encoding the word vector with 0 and 1 so that each word vector had its own unique encoding. Although one-hot encoding solved the problem that machines could not handle discrete data, it ignored the order of words, so it was not conducive to the semantic accuracy. One of the text feature extraction methods using neural networks is word2vec (Mikolov, Chen, Corrado, & Dean, 2013). The vector representation of each word could be obtained by the word2vec, and the semantical similar words would be close to each other, so the relationship between words could be represented. Two sub models named CBOW (Continuous Bag-of-Word) and Skip Gram respectively were proposed in word2vec to predict the current word with the following words and predicted the contextual words with the current word. Although word2vec considered contextual information, it could not solve the problem of multiple meanings of words because of the one-to-one relationship between

words and word vectors. With the rise of Transformer in NLP, Devlin, Chang, Lee, and Toutanova (2019) proposed Bidirectional Encoder Representations from Transformers (BERT). With the huge amount of pre-trained data and the unique training techniques, which allowed BERT to learn features well. BERT designed a "masked language model" (MLM) for pre-training which randomly replaced the token in each training sequence and then predicted the original word of the mask token. Through the BERT, a deep bidirectional language representation was generated which incorporates contextual information.

2) Vision

Visual data includes images, videos and so on. For image data, the Convolutional Neural Networks (CNN) is commonly used for feature extraction (Krizhevsky, Sutskever, & Hinton, 2012; Lecun et al., 1989). CNN is a deep neural network that consists of fully connected layers, convolution layers and pooling layers and the latter two together form a feature extractor. The convolution layers extract specific image features through convolution kernels, which can be enhanced for specific features while reducing noise. The pooling layer can down sample the feature map while retaining useful information. Because of its special shared convolutional kernel, CNN can easily handle high-dimensional data and avoid too many parameters. However, the pooling layers may loss several useful information, which is conducive to the reconstruction of image.

For video data, C3D Network (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015) is a general purpose network that used 3D convolution. The 3D feature map was obtained after the 3D convolution operation. Since the 3D convolution had one more dimension (time dimension) than the 2D convolution, thus the obtained 3D feature map contained the timing information in the video. Although C3D Network was able to model the time information well, it had not the memory function for the input video as recurrent neural network (RNN) do (Elman, 1990). The traditional RNN approach extracted all of the features without any processing of the input. But this would result that network memorized too much useless data, thus giving rise to the gated recurrent unit (GRU) (Cho et al., 2014). GRU was able to solve the long-term dependency problem of sequences because of the using of gate mechanism for controlling the transmission and loss of features. GRU combined the input gate and the forget gate proposed in LSTM (Hochreiter & Schmidhuber, 1997) into one gate called update gate which controlled how much information would be transmitted to the back. Another gate was called reset gate which controlled how much information would be forgotten. Since GRU is able to selectively memorize sequences, it is widely used for feature extraction of video data.

3) Audio

General audio feature extraction methods sample the original waveform, identifies the useful parts of the audio signal so that facilitate the recognition of semantic information and discards the noise. The methods of audio feature extraction can be classified by the different feature extraction processes, and there are the following methods. For example, the methods based on Zero Crossing Rate (ZCR) extract features directly from the original signal. The ZCR is the number of the speech signal passes through the zero point in each frame. ZRC has been widely used in the fields of speech recognition and it becomes the key feature for the classification of tapped sounds. Although ZCR can clearly determine the starting and ending points of unvoiced sound, the statuses of unvoiced sound and environmental noise are similar, thus it is not possible to distinguish them by ZCR. The other methods are based on Spectral Centroid which is a certain frequency range by energy-weighted averaging and it is with important information about the distribution of frequency and energy of the sound signal. In the field of subjective perception, Spectral Centroid describes the brightness of a sound. In addition, the Mel-scale Frequency Cepstral Coefficients (MFCCs) (Godino-Llorente, Gomez-Vilda, & Blanco-Velasco, 2006) have been at the advanced level. MFCCs used the advantages that humans have different perceptual abilities for different frequencies of speech, and relates the pure tone frequencies to the actual measured frequencies, this allows the features to be closer to what humans hear.

2.3. Multi-layer perceptron

MLP was initially considered to have powerful representational capabilities in the field of computer vision (CV), but early MLP training was limited by the computational power of the devices. Later, as the computational power of devices gradually increased, more and more large-scale models appeared. The disadvantages of MLP requiring larger computational power was also solved, so MLP regained its popularity. Some recent work started to use a pure MLP framework for image classification tasks, which had also broadened the use scenarios of MLP. For example, Tolstikhin et al. (2021) proposed MLP-Mixer, which used MLP to replace traditional convolution operations and attention mechanisms and applied them to image classification tasks. MLP-Mixer divided the images into non-overlapping patches and sent them to the MLP for fusion. MLP-Mixer proposed token-mixer and channel-mixer to achieve information fusion in the spatial and channel domains, respectively. Although the MLP-Mixer framework is very simple, it achieves about the same results as the Transformer-based model in image classification tasks. Due to the MLP framework is simple in design and can replace the Transformer module, the feasibility of the MLP architecture in the computer vision field is verified. A series of MLP-based models such as S2MLP (Yu et al., 2022), s2mlpv2 (Yu, Li et al., 2021), ResMLP (Touvron et al., 2021), Hire-MLP (Guo et al., 2022), CycleMLP (Chen, Xie et al., 2022)and GFNet (Zhou, Chen, Liu, & Yu, 2020) had been generated after MLP-Mixer, all of them are continuously improving the performance and even reaching the result of SOTA. Yu et al. (2022) proposed a spatial-shift framework called S2-MLP based on MLP. Unlike MLP-Mixer, S2-MLP contained only channel-mixer. S2MLP designs spatial-shift to implement the communication between different patches. Specifically, the spatial-shift operation first divided the feature into 4 groups by channel, then the first group was shifted horizontally right by one unit, the second group was shifted horizontally left by one unit, the third group was shifted forward by one unit, and the fourth group was shifted backward by one unit. Spatial-shift was a fixed operation without parameters, which could be achieved by a simple assignment of values. This moved features which are at different locations to the same channel. Then a 1×1 convolution is performed to fuse the information within receptive field to achieve the communication between different patches.

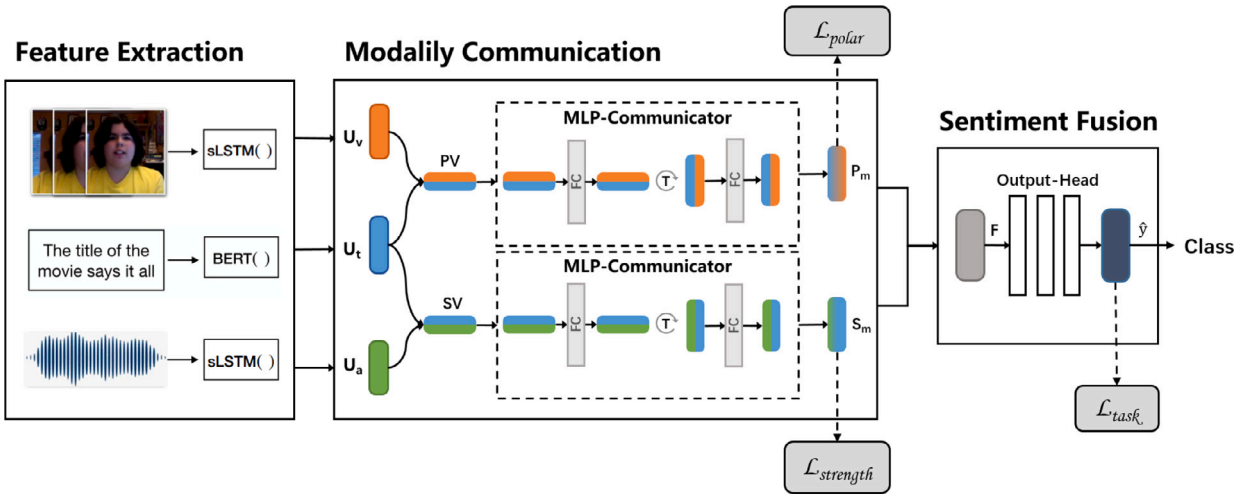

Fig. 2. PS-Mixer consists of three components: feature extraction, modality communication, and sentiment fusion. The feature extraction component generates three kinds of low-dimensional features (visual, audio and text). The modality communication component is used to generate two sentiment scales. The sentiment fusion component fuses two sentiment scales for the classification task.

As substantial results were achieved on the unimodal data, the attention of scientists was turned to the multimodal data. The MLP framework for vision-and-language (VL) fusion was first investigated in Nie et al. (2021). The paper replaced the multi-head attention in Transformer with MLP in order to compare the effects of multi-head attention and MLP. The results were not good enough because they were not trained by using large-scale data. However, when the model was pre-trained with large-scale data, the accuracy improved by 5.73%, competitive to the performce of Transformer. This professes that MLP can replace the Transformer based on a large amount of pre-training, and the Transformer module is not necessary. It also proves that MLP is effective in multimodal fusion. Inspired by this paper, we decide to propose an MLP-based sentiment analysis model applied to the multimodal field.

3. Methodology

3.1. Overview of the framework

In this paper, the proposed method uses the MLP framework to fuse multi-modal data, two scales are generated to determine the direction and strength of the sentiment respectively, then combines two sentiment scales in the sentiment fusion module and output decision result. As shown in Fig. 2, our model contains three main modules: feature extraction module, modality communication module and sentiment fusion module. The feature extraction module is applied to extract multi-modal raw data into three specific low-dimensional vectors representation. The modality communication module is used to interact information between modalities and obtain two different sentiment scales (polarity and strength). Finally, the multi-modal fusion module is set to combine two previous obtained sentiment scales and takes them for the sentiment classification task. The detailed description is provided in the following sections.

3.2. Feature extraction

The first module of PS-Mixer is feature extraction. In this module, we take pre-trained models (BERT, OpenFace and COVAREP) to perform feature extraction for visual, audio and text data (Baltrusaitis, Robinson, & Morency, 2016; Degottex, Kane, Drugman, Raitio, & Scherer, 2014; Devlin et al., 2019). The features generated are used to perform the subsequent multimodal communication. The function of the feature extraction module can be segmented into two main stages: feature extracting and feature embedding. In the stage of feature extracting, pre-trained models are applied to extract features from the raw video files to generate features with different dimensions. During the feature embedding stage, we project three modalities features into the same dimension to get the higher-quality information and the lower computation complexity. With our proposed feature extraction module, we are able to convert multimodal information into low-dimensional features for using in subsequent modules.

3.2.1. Visual and audio feature

According to the description of the official dataset, Openface and COVAREP are used to extract visual and audio features, respectively.

As for Openface, the first step is to find all the faces using the Histogram of Oriented Gradient (HOG) algorithm (Dalal & Triggs, 2005) (Chen, Zhao, Chan, & Kong, 2022; Petsiuk & Pearce, 2022). Openface segments the image into numerous small squares. Then,

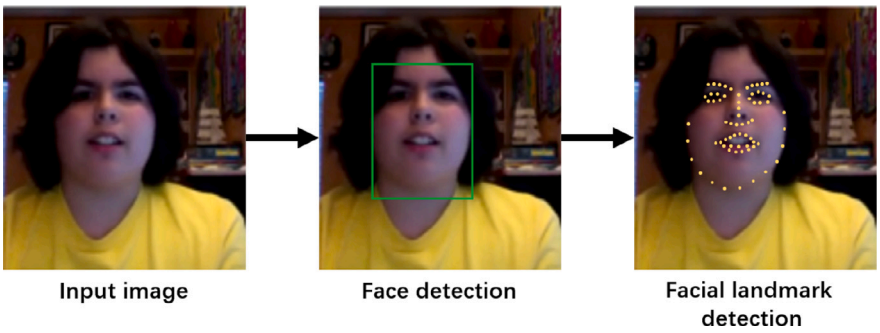

Fig. 3. Flowchart on how openface generates facial landmark. First, the input image is obtained, then the face is detected and circled. Finally the facial landmark is drawn based on the circled face.

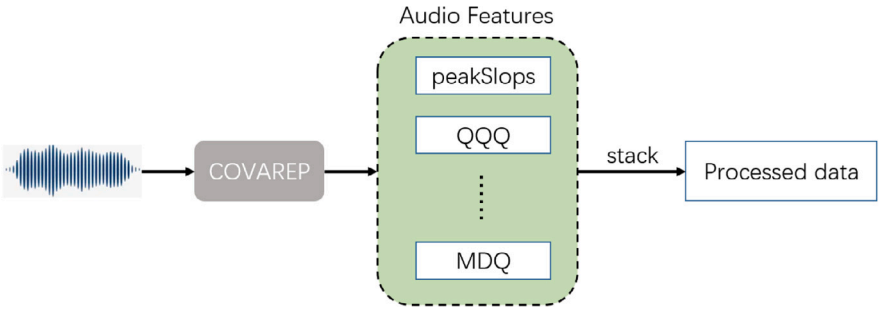

Fig. 4. The general flow of processing audio data. Firstly, the audio features need to be extracted. COVAREP calculates dozens of audio features such as peakSlope, QQQ and MDQ. Then, these features are stacked together to get the processed data.

it calculates how many points there are in each main direction (point up, point right up, point right, etc.) and replaces the original one with the strongest points. The result is that the original image is converted into a very simple representation to capture the basic structure of the face. The second step is to perform facial landmark estimation using an ensemble of regression trees (ERT) (Valle, Buenaposada, Valdés, & Baumela, 2019) on the feature points of the face. The specific flow of facial landmark is shown in Fig. 3. The third step is to encode the face. In this step, a deep convolutional neural network is trained to generate features for the faces. Three faces will be observed during training, the first one is a known face, the second one is a face of the same people, and the third one is a face of different people. The features generated by training the network, so that these features enable the faces of the same person to be as close as possible and the pictures of different people to be as far away as possible.

For audio, the feature extraction module uses a speech processing tool COVAREP, which can extract not only some basic speech features, such as frame energy, fundamental frequency, short-time jitter parameters, but also important speech sentiment feature parameters, such as Mel-scale Frequency Cepstral Coefficients (MFCCs). All features are normalized using zero-mean and variance normalization, and the segment without audio information is set to zero. The general flow of COVAREP processing audio data is as follows. After inputting audio data, COVAREP calculates dozens of audio features such as peakSlope, QQQ and MDQ. These features are stacked together to get the processed data, as shown in Fig. 4. After the above operation, high-quality audio features S_a can be obtained.

Bidirectional LSTM (BiLSTM) considers both forward and backward information to better capture two-way semantic dependencies. BiLSTM is a combination of forward LSTM and backward LSTM and it can concatenate the hidden states of the two LSTMs as the representation of each position. The forward and backward LSTMs are formulated as the following Eqs. (1) and (2):

$$\vec{c}_t, \vec{h}_t = g^{LSTM}\left(\vec{c}_{t-1}, \vec{h}_{t-1}, W_t\right) \tag{1}$$

$$\overleftarrow{c}_t, \overleftarrow{h}_t = g^{LSTM}\left(\overleftarrow{c}_{t+1}, \overleftarrow{h}_{t+1}, W_t\right) \tag{2}$$

where the parameters in the two LSTMs are shared. The g^{LSTM} denotes the one-way LSTM, the arrow denotes the direction of the one-way LSTM, $\rightarrow$ denotes the running process of the forward LSTM, and $\leftarrow$denotes the running process of the backward LSTM. The hidden states $\vec{h}_t$ and memory cell $\vec{c}_t$ of the current state are generated from the states $\vec{h}_{t-1}$ and $\vec{c}_{t-1}$ of the previous time. Each current state ($\vec{h}_t$) will only consider the forward context and not the backward context. To solve the problem that LSTM can only capture one-way information, BiLSTM combines two directions of LSTM, which can consider forward and backward contextual information and better capture bidirectional semantic dependencies. At each current position, the representation of hidden states is $h_t = \vec{h}_t \bigoplus \overleftarrow{h}_t$,

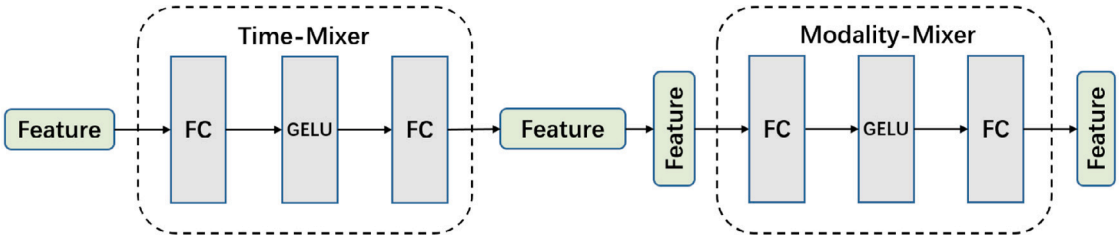

Fig. 5. The framework for the implementation of MLP-Communicator consisting of several fully connected layers as well as activation functions (GELU).

which is a concatenation of the hidden states of the forward LSTM and backward LSTM. In this way, the forward and backward contexts can be considered simultaneously.

$$U_a = sLSTM\left(S_a;\ \theta_a\right) \tag{3}$$

$$U_v = sLSTM\left(S_v;\ \theta_v\right) \tag{4}$$

The audio feature $S_a \in \mathbb{R}^{T_a \times d_a}$ and visual feature $S_v \in \mathbb{R}^{T_v \times d_v}$ are the original feature we used, where $T \times d$ means feature dimensions, fixed-sized vector U is what S produces. θ is separate parameter for each modality. For audio and visual modalities, we apply the stacked bi-directional Long Short-Term Memory (sLSTM) to exact visual and audio feature. The final output $U_a \in \mathbb{R}^{d_h}$ and $U_v \in \mathbb{R}^{d_h}$, the entire sentence representation is processed by the sLSTM network and an embedding layer sequentially.

3.2.2. Text feature

Bidirectional Encoder Representations from Transformers (Bert) is a pre-trained model for the NLP field. BERT uses the Transformer Encoder model as the language model, completely abandoning the RNN and CNN structure and using the attention mechanism to establish long-term dependence. BERT uses multi-head attention to perform self-attention on the input, and the subsequent feed forward operation performs a nonlinear transformation on the vector after the self-attention. Recently, Bert has been used in sentiment analysis as the text feature extractor and has an excellent performance. This model comprises 12 stacked Transformer layers that gives the final output $U_t \in \mathbb{R}^{d_h}$ for the raw sentence $S_t \in \mathbb{R}^{T_m \times d_m}$.

$$U_t = BERT\left(S_t;\ \theta_t\right) \tag{5}$$

3.3. Modality communication

In the modal communication module, the proposed method uses text feature as the reference vector, lets visual feature, audio feature communicate with text feature, respectively. And our proposed MLP-Communicator (MLP-C) generates two vectors called Polar-Vector (PV) and Strength-Vector (SV) to help determine the polarity and strength of sentiment. Specifically, the text and visual features generate PV through MLP-C to determine the direction of the sentiment, and the text and audio features generate SV through MLP-C to determine the strength of the sentiment.

Specifically, Given the audio, visual and text feature for modality $m \in \{a, v, t\}$, we learn the polarity and strength representations employing the encoding functions.

$$h_t^p,\ h_v^p = E_p\left(U_t;\ U_v;\ \theta^p\right) \tag{6}$$

$$h_a^s,\ h_t^s = E_s\left(U_a;\ U_t;\ \theta^s\right) \tag{7}$$

The $h_t^p, h_v^p, h_a^s and h_t^s$ represent the vectors that contain each modal features information. They are generated through simple feed-forward neural layers E which is composed of linear layers and sigmoid activation functions. The superscript p indicates that it is used to determine the polarity, similarly, s indicates that it is used to determine the intensity. Both θ are parameters shared between two modalities.

3.3.1. MLP-communicator

Attention-based networks such as ViT (Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly, Uszkoreit, & Houlsby, 2021) and BERT achieved unparalleled success in all nearly NLP and visual tasks. Recently, Mixer and ResMLP show the potential of MLP architecture that could replace convolutional and attention blocks in numerous fields.

Based on MLP-Mixer architecture (Tolstikhin et al., 2021), the MLP-Communicator we proposed consists of several identical layers, which contains two MLP blocks: Time-mixing MLP and Modality-mixing MLP, as shown in Fig. 5. Modality-mixing MLP affects the modality dimension of the input feature, allowing different modalities to communicate with each other, time-mixing MLP has the same effect on the time dimension.

As a result, information could flow across different modalities and time sequences. Each block is composed of two MLP layers, and one GELU activation function (described as Φ). Additionally, skip connection is also applied in each block. Suppose that $X \in \mathbb{R}^{t \times d}$ is an input feature, where t is the length of time sequences and d is the number of modalities. In each layer, the MLP-Communicator module could be represented as the following:

$$\mathbf{Z}_{*,i} = \mathbf{X}_{*,i} + \mathbf{W}_2 \Phi \left(\mathbf{W}_1 \text{Norm}(\mathbf{X}_{*,i}) \right) \tag{8}$$

$$\mathbf{Y}_{j,*} = \mathbf{Z}_{j,*} + \mathbf{W}_4 \Phi \left(\mathbf{W}_3 \text{Norm}(\mathbf{Z}_{j,*}) \right) \tag{9}$$

where i ranges from 1 to d indicates the number of rows, and j ranges from 1 to t indicates the number of columns. Norm() denotes LayerNorm and W represents the weights of the linear layer in each block. The input feature X first passes through Modality-Mixing MLP and generates Z though skip connection, this step allows the communication of features in the horizontal pairs. Then Z follows Time-Mixing MLP to generate Y, the features are fused in the longitudinal direction. The final obtained feature Y fuses features from two directions. This structure allows each element in the input features could interact with other features along the two dimensions.

3.3.2. Sentiment scale

Generally, text and audio modalities have a stronger correlation with sentiment strength and sentiment polarity. We project three feature vectors into two representations. One is the sentiment direction component that captures the polarity of sentiment whether is positive, negative or neutral. The other is the sentiment strength component that learns the power of sentiment.

We stack modality representations h_t^p and h_v^p into a matrix $PV = [h_t^p, h_v^p] \in \mathbb{R}^{2 \times d_h}$, and apply the same method on h_t^s and h_a^s to make a similar matrix $SV = [h_t^s, h_a^s] \in \mathbb{R}^{2 \times d_h}$.For modality communication, we take text vector as reference vector and apply the MLP-Communicator for these representations. Finally, we acquire P_m that representing the polarity of sentiment and S_m represents the strength of sentiment:

$$P_m = BN\left(MLP - C\left(PV,\ \theta^p\right)\right) \tag{10}$$

$$S_m = BN\left(MLP - C\left(SV,\ \theta^s\right)\right) \tag{11}$$

where, BN is the Batch Normalization function. The algorithm of $MLP - C$ can be found in Section 3.3.1. The parameters θ^p and θ^s are the weights of linear layers in $MLP - C$.

3.4. Modality fusion and prediction

3.4.1. Fusion

After projecting the modalities into their respective sentiment scales, we fuse them into a comprehensive vector for final prediction by extracting the direction of the polar vector and the scale of the strength vector. We construct a joint vector using simple multiplication. The direction of the polar vector and the length of the strength vector are multiplied as follow:

$$F = \|S\|_1 \times \frac{P}{\|P\|_2} \tag{12}$$

where vector $P \in \mathbb{R}^{2d_h}$ consists of one-dimensional vector by splicing the first and second rows of matrix $P_m \in \mathbb{R}^{2 \times d_h}$ and vector $S \in \mathbb{R}^{2d_h}$ is the same situation. The $\|P\|_2$ denotes the 2-Norm of P, it contains information about the length of P, so $\frac{P}{\|P\|_2}$ can be interpreted as a unit vector of P, containing only directional information, with unit length. Similarly, $\|S\|_1$ denotes the 1-Norm of S which can be interpreted as the value of S. The product of the direction of sentiment and the value of sentiment, F, represents the final sentiment score.

3.4.2. Prediction

As for the seven classification task, the predicted values we obtain are grouped between -3 and 3 to represent the strength of the seven sentiments $(-3, -2, -1, 0, 1, 2, 3)$. As for the binary classification task, the predicted values are classified as positive number and negative number, which are used to represent two sentiment direction(positive and negative). The final predicted out is generated by the output-head $\hat{y} = G(F, \theta^o)$ that consists of a linear layer with a ReLU activation which is expressed in the equation as G().

3.5. Objective functions

The overall learning of the model is performed by minimizing loss function:

$$\mathcal{L} = \alpha \mathcal{L}_{task} + \beta \mathcal{L}_{polar} + \gamma \mathcal{L}_{strength} \tag{13}$$

where α, β, γ are the weights which decide the contribution of each regularization component to the total loss $\mathcal{L}$. We discuss them next.

Table 1
Description of TP,TN,FP,FN. The letters T and F in the table stand for true and false. The letters P and N stand for positive and negative predicted results. TP means that both predicted and true values are positive, TN means that both predicted and true values are positive, FP means that the true value is negative but the predicted value is positive, and FN means that the true value is positive but the predicted value is negative.

		Predicted class	
		Class (Yes/+)	Class (No/-)
Actual class	Class (Yes/+)	TP	PN
	Class (No/-)	FP	TN

3.5.1. $\mathcal{L}_{polar}$ - Polar loss

The polar loss helps the sentiment polarity to be more accurate. Minimizing it could improve the accuracy of the binary classification directly. In the process of training, we calculate the Cosine Similarity between predicted mean polar vector and true mean polar vector as $\mathcal{L}_{polar}$. We calculate the distance between the polar vector that is predicted to be positive and that the label is positive. After doing the same for the negative vectors, the weighted sum of them is used as the Polar Loss. so the $\mathcal{L}_{polar}$ is computed as follows:

$$\mathcal{L}_{polar} = 1 - w_1 \cdot Similarity\left(\boldsymbol{p}^+, \boldsymbol{t}^+\right) - w_2 \cdot Similarity(\boldsymbol{p}^-, \boldsymbol{t}^-) \tag{14}$$

where w_1 and w_2 is the weight of positive and negative cosine similarity, and $\boldsymbol{p}$ is the vector of predicted result and $\boldsymbol{t}$ is the vector of true label. The superscript + and - indicates positive numbers and negative numbers. The $Similarity()$ is expressed as the formula for calculating the cosine similarity. Maximizing $w_1 \cdot Similarity\left(\boldsymbol{p}^+, \boldsymbol{t}^+\right)$ and $w_2 \cdot Similarity(\boldsymbol{p}^-, \boldsymbol{t}^-)$ can minimize the $\mathcal{L}_{polar}$.

3.5.2. $\mathcal{L}_{strength}$ - Strength loss

Minimizing the strength loss could help the sentiment strength close to the true value. We compute the correlation coefficient distance as $\mathcal{L}_{strength}$ between $\|S\|_1$, which is the value of strength and the absolute value of the true labels.

The correlation coefficient distance used to calculate the distance between X and Y is defined as follows:

$$D_{corr}(X, Y) = 1 - \frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}} \tag{15}$$

So the $\mathcal{L}_{strength}$ is computed as:

$$\mathcal{L}_{strength} = D_{corr}\left(\left|y^{true}\right|,\ \|S\|_1\right) \tag{16}$$

where $\left|y^{true}\right|$ denotes the absolute value of true labels. $\|S\|_1$ denotes the value of the prediction. Minimizing the correlation coefficient distance between them, i.e. $L_{strength}$, can make the predicted values closer to the true values.

3.5.3. $\mathcal{L}_{task}$ - Task loss

The task-specific loss estimates the quality of prediction during training. We use the mean squared error (MSE) loss. For N sequences in the training data, $\mathcal{L}_{task}$ is calculated as follows:

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{i=0}^{N} \|y_i - \hat{y}_i\|_2^2 \tag{17}$$

where N is the number of training samples, y_i and $\hat{y}_i$ are the true label and predicted label. Minimizing L_{task} reduces the gap between the predicted and true values.

4. Experiments

4.1. Datasets

We have conducted the experiments on two public datasets, CMU-MOSI (Zadeh, Zellers, Pincus, & Morency, 2016) and CMU-MOSEI (Zadeh, Liang, Poria, Cambria, & Morency, 2018), which provide word-aligned multi-modal features for raw data.

4.1.1. CMU-MOSI

The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset (Zadeh et al., 2016) is a popular benchmark that commonly used in the multi-modal sentiment analysis. MOSI dataset is a collection of 2199 opinion video clips. Each opinion video is annotated with sentiment in the range [−3,3], which represents strongly negative/positive sentiment. The dataset is rigorously annotated with labels for subjectivity, sentiment strength, per-frame and per-opinion annotated visual features, and per-milliseconds annotated audio features.

4.1.2. CMU-MOSEI

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset (Zadeh, Liang, Poria, Cambria et al., 2018) is the largest dataset of multimodal sentiment analysis and emotion recognition. The dataset contains more than 23,500 sentence utterance videos from more than 1000 online YouTube speakers. The dataset is gender balanced. All the sentences utterances are randomly chosen from various topics and monolog videos. The videos were transcribed and properly punctuated. The MOSEI dataset is an improvement over MOSI with a higher number of utterances, greater variety in samples, speakers, and topics.

4.2. Baselines

There has been a lot of works conducted in the field of sentiment analysis, especially in the area of multimodal sentiment analysis. As described in Section 2, these approaches can be classified according to different fusion mechanisms, such as tensor-based fusion approaches, GAN-based and attention-based mechanisms, among others. We have conducted a comprehensive comparative study of PS-Mixer, and the baselines of our study are listed below.

1. **Models based on tensor fusion.** Tensor Fusion Network for Multimodal Sentiment Analysis (TFN) (Zadeh et al., 2017), Efficient Low-rank Multimodal Fusion With Modality-Specific Factors (LMF) (Liu et al., 2018), Locally Confined Modality Fusion Network With a Global Perspective for Multimodal Human Affective Computing(LMFN) (Mai, Xing, & Hu, 2020), Divide, Conquer and Combine: Hierarchical Feature Fusion Network with and Global Perspectives for Multimodal Affective Computing (HFFN) (Mai, Hu, & Xing, 2019).
2. **Models based on Generative Adversarial Network.** Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion (ARGF) (Mai, Hu, & Xing, 2020), Dynamic Fusion for Multimodal Data (Sahu & Vechtomova, 2021), Speaker-invariant Affective Representation Learning via Adversarial Training (Li, Tu, Huang, Narayanan, & Georgiou, 2020), Adversarial Multimodal Representation Learning for Click-Through Rate Prediction (Li, Wang, et al., 2020), Learning Factorized Multimodal Representations Tsai, Liang, et al. (2019).
3. **Models based on attention mechanisms.** Memory Fusion Network for Multi-View Sequential Learning (MFN) (Zadeh, Liang, Mazumder, et al., 2018), Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors (RAVEN) (Wang, Shen et al., 2019), Multimodal Transformer for Unaligned Multimodal Language Sequences (MulT) (Tsai, Bai, et al., 2019), Multimodal Split Attention Fusion (MSAF) (Su et al., 2020), Hierarchical Delta-Attention Method for Multimodal Fusion (Panchal, 2020), Attention Is Not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion (Liang, Lin, Feng, Zhang, & Lv, 2021).
4. **Models based on Graph-based fusion.** Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph (Graph-MFN) (Zadeh, Liang, Poria, Cambria et al., 2018), Graph Completion Network for Incomplete Multimodal Learning in Conversation (GCNet) (Lian, Chen, Sun, Liu, & Tao, 2022), Analyzing Unaligned Multimodal Sequence via Graph Convolution and Graph Pooling Fusion (Mai, Xing, He, Zeng, & Hu, 2020), Graph Capsule Aggregation for Unaligned Multimodal Sequences (Wu, Mai, & Hu, 2021), Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion (ARGF) (Mai, Hu, & Xing, 2020).
5. **Models based on time series.** Memory Fusion Network for Multi-view Sequential Learning (MFN) (Zadeh, Liang, Mazumder, et al., 2018), Multi-attention Recurrent Network for Human Communication Comprehension (MARN) (Zadeh, Liang, Poria, Vij, et al., 2018), Extending Long Short-Term Memory for Multi-View Structured Learning (MV-LSTM) (Rajagopalan, Morency, Baltrusaitis, & Goecke, 2016), Multimodal Language Analysis with Recurrent Multistage Fusion (RMFN) (Liang, Liu, Zadeh, & Morency, 2018).
6. **Models based on gating mechanism.** Gated Mechanism For Attention Based Multimodal Sentiment Analysis (GATE) (Kumar & Vepa, 2020), Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition (Beard et al., 2018), Multimodal sentiment analysis based on feature fusion of attention mechanism-bidirectional gated recurrent unit (Xuemei, Hong, Hongyu, & Shanshan, 2021).
7. **Models based on Multi-tasking.** Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis (Chauhan, Akhtar, Ekbal, & Bhattacharyya, 2019), Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis (Akhtar et al., 2019), Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis (Chen, Luo, Xu, & Ke, 2020), Weakly-supervised Multi-task Learning for Multimodal Affect Recognition (Dai, Cahyawijaya, Bang, & Fung, 2021).

4.3. Evaluation metrics

In order to make a comprehensive evaluation on PS-Mixer model, various standard measures like mean absolute error (MAE) and Pearson correlation (Corr) are used. Additionally, the benchmark also has a classification index with seven-class accuracy (Acc-7), binary accuracy (Acc-2) and F-score.

$$\text{Prec} = \frac{TP}{TP + FP} \tag{18}$$

$$\text{Rec} = \frac{TP}{TP + FN} \tag{19}$$

Table 2
Performances of multimodal models in MOSEI. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive". The upward arrow indicates that the higher this indicator is, the better it is, and the downward arrow is the opposite.

Models	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-score (↑)	Acc-7 (↑)
LMF	0.623	0.677	82.0	82.2	48.0
LMFN	–	–	80.85	80.92	–
ARGF	–	–	–	–	–
MFM	0.568	0.717	84.4	84.3	51.3
RAVEN	0.614	0.662	79.1	79.5	50.0
MuIT	0.630	0.664	80.1	80.9	49.0
MSAF	0.559	0.738	85.5	85.5	52.4
MICA	–	–	83.7	83.3	52.4
Graph-MFN	0.710	0.540	76.9	77.0	45.0
Multimodal Graph	0.608	0.675	81.4	81.7	49.7
GraphCAGE	0.609	0.670	81.7	81.8	48.9
MFN	0.612	0.687	80.6	80.0	49.1
MV-LSTM	–	–	76.4	76.4	–
GATE	–	–	81.14/85.27	78.53/84.08	
AMF-BiGRU	–	–	78.48	78.16	–
CIA	0.680	0.590	80.4	78.2	50.1
CIM-MTL	–	–	80.5	78.8	–
DFF-ATMF	–	–	77.1	78.3	–
PS-Mixer	**0.537**	**0.765**	**83.1/86.1**	**83.1/86.1**	**53.0**
Δ SOTA	↓ 0.022	↑ 0.027	↑ 0.61	↑ 0.61	↑ 0.6

The precision and recall are computed using Eqs. (18) and (19). Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the observations in the actual class (Yes / +). As shown in Table 1, these measures can be interpreted as follows. TP means "True Positive" - these are the correctly predicted positive values, which means that the value of the actual class is (Yes / +) and the value of the predicted class is also (Yes / +). FP means "False Positive" - when the actual class is (No / -) and the predicted class is (Yes / +). TN, "True Negatives" - these are the correctly predicted negative values, which means that the value of actual class is (No / -) and value of the predicted class is also (No / -). FN, "False Negatives" - when the actual class is (Yes / +) but the predicted class is (No / -).

There is an anti-correlation between precision and recall. It means that the recall drops when the precision rises and vice versa. In other words, a system that attempts for recall gets lower precision, and a system that attempts for precision gets a lower recall. To consider the two metrics together, a single measure, called F-measure, is used. F-measure is a statistical measure that merges both precision and recall. This is calculated as follows:

$$F_1 = \frac{1}{\delta \cdot \frac{1}{P} + (1-\delta)\frac{1}{R}} = \frac{\left(\gamma^2+1\right)\text{Prec}\cdot\text{Rec}}{\left(\gamma^2\right)\text{Prec}+\text{Rec}} \tag{20}$$

where $\gamma^2 = \frac{1-\delta}{\delta}, \alpha \in [0,1]$, and $\gamma^2 \in [0,\infty]$. If a large value ($\gamma > 1$) assigns to the γ, it indicates that precision has more priority. If a small value ($\gamma < 1$) assigns to the γ, it indicates that recall has more priority. If $\gamma = 1$ the precision and recall are assumed to have equally priority in computing F-measure. F-measure for $\gamma = 1$ is computed as follows:

$$F_1 = \frac{2\cdot\text{Prec}\cdot\text{Rec}}{\text{Prec}\ +\text{Rec}} \tag{21}$$

where Prec is precision and Rec is recall.

4.4. Quantitative evaluations of the approaches

The results of our experiments on two publicly available datasets, CMU-MOSI and CMU-MOSEI, are presented in Table 2 (MOSEI) and Table 3 (MOSI). After we investigated a large number of models for the same task, our model could outperform most baselines on all metrics (MAE, Corr, Acc-2, F1-score, Acc-7). Following the previous works, we report Weighted F1 score (F1-Score) and binary classification accuracy (Acc-2). Specifically, for MOSI and MOSEI datasets, we calculate Acc-2 and F1-Score in two ways: negative/non-negative (non-exclude zero) and negative/positive (exclude zero) (Hazarika et al., 2020). Compared to TFN based on tensor fusion, our model on the CMU-MOSEI dataset has higher binary classification accuracy than TFN, while the seven classification accuracy was higher. We consider that this is because tensor fusion-based model in TFN cannot further explore the connections between multimodal data, while our proposed multimodal communication module just enables good interaction between multimodal data. These results are a good demonstration of the superiority of PS-Mixer.

Table 3
Performances of multimodal models in MOSI.

Models	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-score (↑)	Acc-7 (↑)
TFN	0.970	0.633	73.9	73.4	32.1
LMF	0.912	0.668	76.4	75.7	32.8
LMFN	–	–	80.9	80.9	–
HFFN	–	–	80.2	80.3	–
ARGF	–	–	81.4	81.5	–
MFM	0.877	0.706	81.7	81.6	35.4
RAVEN	0.915	0.691	78.0	76.6	33.2
MuIT	0.871	0.698	83.0	**82.8**	40.0
MICA	–	–	82.6	82.7	40.8
Multimodal Graph	0.933	0.684	80.6	80.5	32.1
GraphCAGE	0.933	0.684	82.1	82.1	35.4
MFN	0.965	0.632	77.4	77.3	34.1
MARN	0.968	0.625	77.1	77.0	34.7
MV-LSTM	1.019	0.601	73.9	74	33.2
RMFN	0.922	0.681	78.4	78.0	38.3
GATE	–	–	**83.91**	81.17	–
AMF-BiGRU	–	–	82.05	82.02	–
CIA	0.914	0.689	79.8	79.5	38.9
DFF-ATMF	–	–	80.9	81.2	–
PS-Mixer	**0.794**	**0.748**	80.3/82.1	80.3/82.1	**44.31**

Table 4
Ablation study of PS-Mixer model on different components and three losses. For component part, we remove MLP-C and test on CMU-MOSI. For loss part, we replace the loss function with an Euclidean distance form.

Method	MAE	Corr	Acc-2	F-score	Acc-7
w/o MLP	0.871	0.714	0.782/0.795	0.782/0.794	38.7
task loss with Euclidean metric	0.864	0.737	0.793/0.814	0.793/0.813	36.5
polar loss with Euclidean metric	0.867	0.714	0.790/0.803	0.789/0.802	37.4
strength loss with Euclidean metric	0.819	0.754	0.794/0.809	0.794/0.809	38.1
PS-Mixer	**0.794**	**0.748**	**80.3/82.1**	**80.3/82.1**	**44.31**

Table 5
Ablation study of PS-Mixer model on different components and three losses. For component part, we remove MLP-C and test on CMU-MOSEI. For loss part, we replace the loss function with an Euclidean distance form.

Method	MAE	Corr	Acc-2	F-score	Acc-7
w/o MLP	0.543	0.760	0.825/0.854	0.822/0.855	52.7
task loss with Euclidean metric	**0.443**	0.756	0.815/0.850	0.815/0.850	52.2
polar loss with Euclidean metric	0.606	0.759	0.817/0.850	0.812/0.850	49.2
strength loss with Euclidean metric	0.543	0.756	0.843/0.855	0.842/0.857	52.7
PS-Mixer	0.537	**0.765**	**83.1/86.1**	**83.1/86.1**	**53.0**

4.5. Ablation study

We conduct an ablation study to evaluate the MLP communication structure and the contribution of the three loss functions. As shown in Tables 4, To explore the role of the different components, we first remove the MLP communication structure and test it on the MOSI dataset, and we can see that the accuracy of the binary classification and seven classifications decreased by 3.9% and 5.5% respectively. Next, as shown in Table 5 we replace the loss function with an Euclidean distance form and the model performance also decreases. Overall, using our proposed MLP communication structure and using the cosine similarity in the polar loss function, the correlation coefficient distance in strength loss can significantly improve the model results.

4.5.1. MLP vs. Transformer

The well-established transformer layer is composed of one or more multi-head attention blocks, which are used to capture position-wise token interactions by aggregating information across tokens. It could also have the function of modality communicating though with a huge number of parameters.

To compare the performance of the MLP and transformer on the Modality Communication module, we replace the MLP-Communicator with one-head and two-head transformer encoder layer. Based on the same super parameter in the MOSI dataset,

Table 6
Comparison with Transformer: We replace the MLP-Communicator with two kinds of transformer and compute the parameter scale of the model.

Model	Param	MAE	Corr
Transformer(One-head)	1.4M	0.83	0.72
Transformer(Two-head)	1.4M	0.81	0.73
MLP	0.3M	0.82	0.73

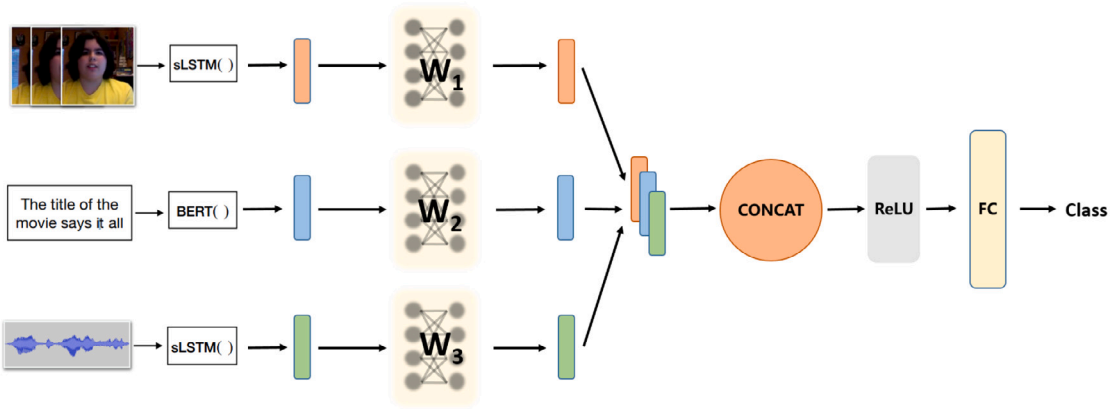

Fig. 6. Modality select model: we design a network with parallel input and same initial weights to select modalities on different data.

Table 7
Contribution of different modalities.

Modality	Sentiment polarity	Sentiment strength
text	48.3%	51.4%
audio	21.1%	29.4%
visual	30.5%	19.1%

we compute the parameter scale of the model without the part for feature representation, and the result is shown in Table 6. We could see that MLP-Communicator achieves nearly the same effect with less parameters.

4.5.2. Polar and strength vector validation

To verify the role of polar and strength vector in PS-Mixer model, We have designed special experiments to demonstrate their performance in judging the direction and strength of the sentiment. For polar vector, the label values are discretized as positive or negative and the MSE loss function is replaced with the Cross-Entropy loss function. In the model, we only use the polar vector for prediction and obtain the experimental results. For strength vector, we take the absolute values of the labels and only use the strength vector in the model for prediction to get the results.

For the polar vector, the Acc-2 reaches 81.5% and 85.9% for MOSI and MOSEI, respectively. Similarly, in the experiments with strength vector, MAE reaches 0.63 in MOSI and 0.61 in MOSEI. This result indicates that the design of the polar vector and strength vector has the expected effect.

4.5.3. Modality selection

Our model learns from sentiment polarity and sentiment strength. Since different modality may have different contributions on them, a feature selection is conducted to determine the model input and modality combination. We propose to input the modality set in parallel channels, and design the model to learn which modality is more relevant for lower loss value and give higher weight to those modalities as shown in Fig. 6.

We intend to weight three modalities with a single scalar value that can reflect the scale of weight parameters. Let w_m denote the weight matrix for the $m \in \{a, v, l\}$ modality vector. Before training, the weights corresponding to each modality vector are initialized with equal values and the input data is normalized to the same interval. Then, the weighted data are concatenated and the output-head is applied.As the weight vector w_m affects the magnitude of the modalities, they also affect the gradients propagated back to the linear layer, which transforms the input features. Therefore, it is important to have a unique weight matrix for each input feature matrix.We treat the $L2$ parametrization of weight vector as a way to evaluate different modalities in polarity and strength. Experiment is conducted on the MOSI dataset and the results are shown on Table 7.

From the result of the contribution in Table 7, we could find that text and visual modalities have close relevance to sentiment polarity, as well as that text and audio modalities on sentiment strength.

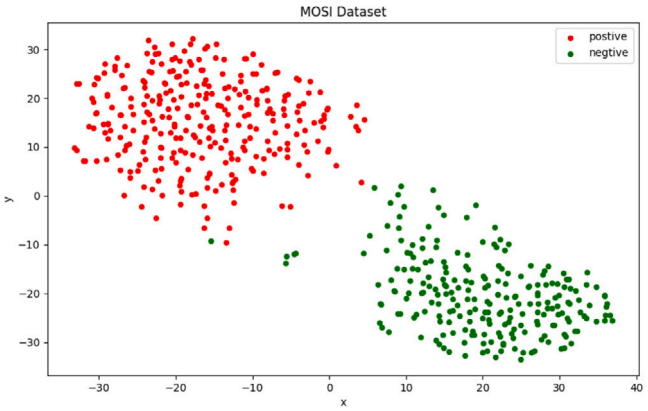

Fig. 7. Polar Vector Visualization: The red points are two-dimension vectors that are positive label, as the same as the green points are negative. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.6. Visualizing polar vector

We visualize the hidden polar vector for the samples in the testing sets. Fig. 7 presents the illustrations. We process the polar vectors into two dimensions using t-distributed stochastic neighbor embedding (T-SNE). It is obvious that the two-dimensions polar vectors are gathered into two clusters using our proposed method. This indicates that our method can well distinguish the sentiment polarity of multimodal data.

5. Conclusion

In this paper, we propose a Polar-Vector and Strength-Vector mixer model based on MLP-Mixer (PS-Mixer) that can effectively fuse multimodal information and improve the accuracy of sentiment analysis. We use MLP for communicating modal information, discarding the Transformer's multi-headed attention mechanism. The polar and strength scales of the sentiment states are together for sentiment analysis. Our experiments show that this MLP-based model achieves SOTA results in multimodal sentiment analysis tasks compared with baseline methods. Our model not only outperforms other models in terms of accuracy due to its innovative use of the MLP mechanism, but also is smaller in parameter than other similar model (e.g. Transformer). Our model demonstrates the feasibility of the MLP mechanism in multimodal tasks, showing that attention is not necessary and can be replaced by MLP. Anyway, the fusion method in PS-Mixer still needs to be improved to be applicable to emotion analysis. For future work, we will explore the way to communicate and fuse multimodal data more efficiently.

CRediT authorship contribution statement

Han Lin: Conceptualization, Methodology, Writing – original draft. **Pinglu Zhang:** Methodology, Writing – original draft, Experiments. **Jiading Ling:** Writing – review & editing. **Zhenguo Yang:** Writing – review & editing, Supervision. **Lap Kei Lee:** Writing – review, Supervision. **Wenyin Liu:** Writing – review & Supervision.

Data availability

Data will be made available on request.

Acknowledgments

This work is partly supported by the Science and Technology Program of Guangzhou (No. 202102020524), and the RGC of HKSAR, China, under Grant UGC/FDS16/E17/21.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, *26*(3).

Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. *1*, In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics* (pp. 370–379).

Arjmand, M., Dousti, M. J., & Moradi, H. (2021). TEASEL: A transformer-based speech-prefixed language model. *CoRR*, *abs/2109.05522*.

Balahur, A., Montoyo, A., Martínez-Barco, P., & Boldrini, E. (Eds.), (2012). *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*. The Association for Computer Linguistics.

Baltrusaitis, T., Robinson, P., & Morency, L. (2016). OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision* (pp. 1–10).
Beard, R., Das, R., Ng, R. W. M., Gopalakrishnan, P. G. K., Eerens, L., Swietojanski, P., et al. (2018). Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 251–259).
Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.
Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5646–5656).
Chen, F., Luo, Z., Xu, Y., & Ke, D. (2020). Complementary fusion of multi-features and multi-modalities in sentiment analysis. In *Proceedings of the 3rd workshop on affective content analysis (AffCon 2020) co-located with thirty-fourth AAAI conference on artificial intelligence* (pp. 82–99).
Chen, S., Xie, E., Ge, C., Liang, D., & Luo, P. (2022). CycleMLP: A MLP-like architecture for dense prediction. In *The Tenth International Conference on Learning Representations*.
Chen, L., Zhao, Y., Chan, J. C.-W., & Kong, S. G. (2022). Histograms of oriented mosaic gradients for snapshot spectral image description. *ISPRS Journal of Photogrammetry and Remote Sensing*, *183*, 79–93.
Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734).
Dai, W., Cahyawijaya, S., Bang, Y., & Fung, P. (2021). Weakly-supervised multi-task learning for multimodal affect recognition. *CoRR*, *abs/2104.11560*.
Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition* (pp. 886–893).
Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP-A collaborative voice analysis repository for speech technologies. In *IEEE international conference on acoustics, speech and signal processing* (pp. 960–964).
Delbrouck, J., Tits, N., Brousmiche, M., & Dupont, S. (2020). A transformer-based joint-encoding for emotion recognition and sentiment analysis. *CoRR*, *abs/2006.15955*.
Devlin, J., Chang, M., Lee, K., & Toutanova, K. (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* (pp. 4171–4186).
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations*.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.
Gamon, M., Aue, A., Corston-Oliver, S., & K.Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Advances in intelligent data analysis VI, 6th international symposium on intelligent data analysis* (pp. 121–132).
Ghaleb, E., Popa, M., & Asteriadis, S. (2020). Metric learning-based multimodal audio-visual emotion recognition. *Ieee Multimedia*, *27*(1), 37–48.
Godino-Llorente, J. I., Gomez-Vilda, P., & Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *IEEE Transactions on Biomedical Engineering*, *53*(10), 1943–1953.
Guo, J., Tang, Y., Han, K., Chen, X., Wu, H., Xu, C., et al. (2022). Hire-MLP: Vision MLP via hierarchical rearrangement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 816–826).
Guyon, I., Nikravesh, M., Gunn, S. R., & Zadeh, L. A. (Eds.), (2006). Feature extraction : Foundations and applications. In *Studies in fuzziness and soft computing*: *207*, Springer.
Han, J., Zhang, Z., Ren, Z., & Schuller, B. W. (2021). EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *IEEE Trans. Affect. Comput.*, *12*(3), 553–564.
Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM international conference on multimedia* (pp. 1122–1131).
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Hou, Q., Jiang, Z., Yuan, L., Cheng, M., Yan, S., & Feng, J. (2021). Vision permutator: A permutable MLP-like architecture for visual recognition. CoRR abs/2106.12368.
Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification. In *The 49th annual meeting of the association for computational linguistics* (pp. 151–160).
Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1106–1114.
Kumar, A., & Vepa, J. (2020). Gated mechanism for attention based multi modal sentiment analysis. In *2020 IEEE international conference on acoustics, speech and signal processing* (pp. 4477–4481).
Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Lecun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.
Li, F., Pan, S. J., Jin, O., Yang, Q., & Zhu, X. (2012). Cross-domain co-extraction of sentiment and topic lexicons. In *The 50th annual meeting of the association for computational linguistics* (pp. 410–419).
Li, H., Tu, M., Huang, J., Narayanan, S., & Georgiou, P. G. (2020). Speaker invariant affective representation learning via adversarial training. In *2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7144–7148).
Li, X., Wang, C., Tan, J., Zeng, X., Ou, D., & Zheng, B. (2020). Adversarial multimodal representation learning for click-through rate prediction. In *WWW '20: The web conference 2020* (pp. 827–836).
Lian, Z., Chen, L., Sun, L., Liu, B., & Tao, J. (2022). GCNet: Graph completion network for incomplete multimodal learning in conversation. *CoRR*, *abs/2203.02177*.
Lian, Z., Liu, B., & Tao, J. (2021). Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 985–1000.
Liang, T., Lin, G., Feng, L., Zhang, Y., & Lv, F. (2021). Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *2021 IEEE/CVF international conference on computer vision* (pp. 8128–8136).
Liang, P. P., Liu, Z., Zadeh, A., & Morency, L. (2018). Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 150–161).
Liu, H., Dai, Z., So, D. R., & Le, Q. V. (2021). Pay attention to MLPs. In *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021* (pp. 9204–9215).
Liu, Q., Gui, Z., Xiong, S., & Zhan, M. (2021). A principal component analysis dominance mechanism based many-objective scheduling optimization. *Applied Soft Computing*, *113*, 107931.
Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 2247–2256).
Lucas, A. (2014). Ising formulations of many NP problems. *Frontiers in Physics*, *2*, 5.
Mai, S., Hu, H., & Xing, S. (2019). Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 481–492).

Mai, S., Hu, H., & Xing, S. (2020). Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *The thirty-fourth AAAI conference on artificial intelligence* (pp. 164–172).
Mai, S., Xing, S., He, J., Zeng, Y., & Hu, H. (2020). Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. *CoRR, abs/2011.13572*.
Mai, S., Xing, S., & Hu, H. (2020). Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia*, *22*(1), 122–137.
Martinez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial intelligence. *Computers in Human Behavior*, *21*(2), 323–341.
Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., & Wang, H. (2012). Cross-lingual mixture model for sentiment classification. In *The 50th annual meeting of the association for computational linguistics* (pp. 572–581).
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International conference on learning representations*.
Nie, Y., Li, L., Gan, Z., Wang, S., Zhu, C., Zeng, M., et al. (2021). MLP architectures for vision-and-language modeling: An empirical study. *abs/2112.04453*,
Panchal, K. (2020). Hierachical Delta-attention method for multimodal fusion. *CoRR, abs/2011.10916*.
Paraskevopoulos, G., Georgiou, E., & Potamianos, A. (2022). Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis. In *IEEE International conference on acoustics, speech and signal processing* (pp. 4573–4577).
Petsiuk, A. L., & Pearce, J. M. (2022). Towards smart monitored AM: Open source in-situ layer-wise 3D printing image anomaly detection using histograms of oriented gradients and a physics-based rendering engine. *Additive Manufacturing*, *52*, 102690.
Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research.
Rajagopalan, S. S., Morency, L., Baltrusaitis, T., & Goecke, R. (2016). Extending long short-term memory for multi-view structured learning. In *Computer vision - ECCV 2016 - 14th European conference* (pp. 338–353).
Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. In *Machine learning and knowledge discovery in databases* (pp. 18–33).
Rubin, & Kenneth, H. (1998). Social and emotional development from a cultural perspective. *Developmental Psychology*, *34*(4), 611–615.
Sahu, G., & Vechtomova, O. (2021). Adaptive fusion techniques for multimodal data. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main Volume* (pp. 3156–3166).
Shenoy, A., & Sardana, A. (2020). Multilogue-net: A context aware RNN for multi-modal emotion detection and sentiment analysis in conversation. *CoRR, abs/2002.08267*.
Su, L., Hu, C., Li, G., & Cao, D. (2020). MSAF: Multimodal split attention fusion. *abs/2012.07175*,
Sun, L., Wang, K., Xu, L., Zhang, C., & Balezentis, T. (2022). A time-varying distance based interval-valued functional principal component analysis method - A case study of consumer price index. *Information Sciences*, *589*, 94–116.
Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., et al. (2021). MLP-mixer: An all-MLP architecture for vision. In *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021* (pp. 24261–24272).
Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., et al. (2021). ResMLP: Feedforward networks for image classification with data-efficient training. CoRR abs/2105.03404.
Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *2015 IEEE international conference on computer vision* (pp. 4489–4497).
Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 6558–6569).
Tsai, Y. H., Liang, P. P., Zadeh, A., Morency, L., & Salakhutdinov, R. (2019). Learning factorized multimodal representations. In *7th International conference on learning representations*.
Valle, R., Buenaposada, J. M., Valdés, A., & Baumela, L. (2019). Face alignment using a 3D deeply-initialized ensemble of regression trees. *Computer Vision and Image Understanding*, *189*, 102846.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017* (pp. 5998–6008).
Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L. (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 7216–7223).
Wang, Y., Wu, J., & Hoashi, K. (2019). Multi-attention fusion network for video-based emotion recognition. In *International conference on multimodal interaction* (pp. 595–601).
Wu, J., Mai, S., & Hu, H. (2021). Graph capsule aggregation for unaligned multimodal sequences. In *ICMI '21: International conference on multimodal interaction* (pp. 521–529).
Xuemei, L., Hong, T., Hongyu, C., & Shanshan, L. (2021). Multimodal sentiment analysis based on feature fusion of attention mechanism-bidirectional gated recurrent unit. *Journal of Computer Applications*, *41*(5), 1268.
Yu, T., Li, X., Cai, Y., Sun, M., & Li, P. (2021). S^2-MLPv2: Improved spatial-shift MLP architecture for vision. CoRR abs/2108.01072.
Yu, T., Li, X., Cai, Y., Sun, M., & Li, P. (2022). S^2-MLP: Spatial-shift MLP architecture for vision. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3615–3624).
Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-fifth AAAI conference on artificial intelligence* (pp. 10790–10797).
Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language Processing* (pp. 1103–1114).
Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the thirty-second AAAI conference on artificial intelligence* (pp. 5634–5641).
Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & 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* (pp. 2236–2246).
Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. (2018). Multi-attention recurrent network for human communication comprehension. In *Proceedings of the thirty-second AAAI conference on artificial intelligence* (pp. 5642–5649).
Zadeh, A., Zellers, R., Pincus, E., & Morency, L. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, *31*(6), 82–88.
Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., et al. (2020). An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *The Thirty-fourth AAAI conference on artificial intelligence* (pp. 303–311).
Zhou, W., Chen, Y., Liu, C., & Yu, L. (2020). GFNet: Gate fusion network with Res2Net for detecting salient objects in RGB-D images. *IEEE Signal Processing Letters*, *27*, 800–804.
Zhu, Y., Chen, Z., & Wu, F. (2019). Multimodal deep denoise framework for affective video content analysis. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 130–138).