

# CubeMLP: An MLP-based Model for Multimodal Sentiment Analysis and Depression Estimation

Hao Sun  
sunhaoxx@zju.edu.cn  
College of Computer Science and  
Technology  
Zhejiang University  
Hangzhou, China

Hongyi Wang  
whongyi@zju.edu.cn  
College of Computer Science and  
Technology  
Zhejiang University  
Hangzhou, China

Jiaqing Liu  
liu-j@fc.ritsumei.ac.jp  
College of Information Science and  
Engineering  
Ritsumeikan University  
Shiga, Japan

Yen-Wei Chen  
chen@is.ritsumei.ac.jp  
College of Information Science and  
Engineering  
Ritsumeikan University  
Shiga, Japan

Lanfen Lin  
llf@zju.edu.cn  
College of Computer Science and  
Technology  
Zhejiang University  
Hangzhou, China

## ABSTRACT

Multimodal sentiment analysis and depression estimation are two important research topics that aim to predict human mental states using multimodal data. Previous research has focused on developing effective fusion strategies for exchanging and integrating mind-related information from different modalities. Some MLP-based techniques have recently achieved considerable success in a variety of computer vision tasks. Inspired by this, we explore multimodal approaches with a *feature-mixing* perspective in this study. To this end, we introduce CubeMLP, a multimodal feature processing framework based entirely on MLP. CubeMLP consists of three independent MLP units, each of which has two affine transformations. CubeMLP accepts all relevant modality features as input and *mixes* them across three axes. After extracting the characteristics using CubeMLP, the mixed multimodal features are flattened for task predictions. Our experiments are conducted on sentiment analysis datasets: CMU-MOSI and CMU-MOSEI, and depression estimation dataset: AVEC2019. The results show that CubeMLP can achieve state-of-the-art performance with a much lower computing cost.

## CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; *Sentiment analysis*; • **Computing methodologies** → *Neural networks*.

## KEYWORDS

multimodal processing; multimodal fusion; multimodal interaction; multimedia; MLP; sentiment analysis; depression detection

### ACM Reference Format:

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. CubeMLP: An MLP-based Model for Multimodal Sentiment Analysis and Depression Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3503161.3548025>

## 1 INTRODUCTION

Multimodal data has become an important means of communication for individuals and the public as social media has grown in prevalence. In this scenario, estimating human psyche states from multimodal data, such as sentiment tendencies and depression levels, becomes increasingly important. Multimodal data commonly includes textual ( $t$ ), acoustic ( $a$ ), and visual ( $v$ ) information. The characteristics of features can be extracted from the multimodal data as two-dimensional matrices  $R^{L_m \times D_m}$ , where  $L_m$  and  $D_m$  are the sequential length and feature-channel size of modality  $m$ , respectively.

To effectively process multimodal features, Zadeh et al. [37] first introduced Cartesian product in Tensor Fusion Network (TFN) to blend the features from all involved modalities at  $L_m$  axis. Many researchers then proposed that there are bidirectional relationships and complementary information among modalities, and apply the attention mechanism to calculate the coattention in modality pairs (like textual and acoustic) [9, 39]. Most recently, with the remarkable success of Transformer-based structures [34], some works attempt to employ the self-attention mechanism for modality interactions [3, 4, 11, 35]. The core of these trending methods is mostly the exchange of information between modalities.

The consequences of these information-exchanging methods could be viewed as feature *mixing*. For instance, the TFN [37] and some sequential-wise coattention approaches [19, 39] attempted to mix the features on  $L_m$  axis between modalities, while the channel-wise coattention approaches [21] tried to perform the mixing on  $D_m$

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '22, October 10–14, 2022, Lisboa, Portugal

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3548025>

axis. As for the Transformer-based approaches [3, 4, 11, 35], they employ the self-attention mechanism and perform the sophisticated mixing on  $L_m$  axis between paired modalities, which can also be treated as enhancing one modality by the other.

A number of variations, such as ViT [7] and ViViT [1] have been proposed recently as a result of the increasing use of transformers in computer vision applications. Transformers, on the other hand, have a significant memory requirement for self-attention, which is a major drawback of the Transformer's architecture. Consequently, structures composed entirely of multilayer perceptrons (MLPs) have gained interest. MLP-mixer [31] and ResMLP [32], for example, use MLPs to replace the self-attention mechanism design. By substituting self-attention with MLPs, these techniques significantly reduce computational costs while maintaining high performance.

Inspired by MLP-based techniques, here, we propose CubeMLP, a simple yet effective MLP-based framework for multimodal feature processing. During preprocessing, we integrate the modality features into a multimodal tensor  $X \in R^{L \times M \times D}$  where  $M$  is the number of modalities,  $L$  is sequential length and  $D$  is the size of feature channels. CubeMLP is composed of three MLP units for three respective axis ( $L$ ,  $M$ , and  $D$ ). The first MLP unit is designed to mix up features on the  $L$  axis, a process called *sequential-mixing*. *Modality-mixing* ( $t$ ,  $a$ , and  $v$ ) is performed by the second MLP unit on the  $M$  axis. Finally, the third MLP unit on the  $D$  axis performs *channel-mixing*. Every MLP unit contains two fully-connected layers, each of which contains an affine transformation that can be represented mathematically as a matrix  $W$  with bias  $B$ . In CubeMLP, we mix multimodal features on each possible axis using the proposed three MLP design structures. After that, the mixed features are flattened and fed to the classifier for the predictions. During this procedure, the multimodal features are fused and multimodal information is exchanged on any axis.

Our contributions can be summarized as follows:

- We propose CubeMLP, a multimodal feature processing framework based exclusively on MLP. CubeMLP mixes features on three axes: sequence ( $L$ ), modality ( $M$ ), and channel ( $D$ ). The distinct multimodal information ( $t$ ,  $a$ , and  $v$ ) is effectively transmitted and shared during the mixing process to extract important features for sentiment analysis and depression detection.
- We propose to use MLPs to largely reduce the computational burden while achieving competitive results with some state-of-the-art approaches, which proves that our CubeMLP is an efficient structure for multimodal feature processing.
- We conduct thorough experiments on two mind state estimation tasks to validate the effectiveness of CubeMLP: multimodal sentiment analysis and multimodal depression detection. The results show that our method is favorably competitive with the state-of-the-art methods for sentiment analysis, while achieving great progress for depression detection.

## 2 RELATED WORKS

### 2.1 Multimodal Sentiment Analysis

The task of multimodal sentiment analysis is to predict the sentiment tendencies from human's facial expression( $v$ ), acoustic tone( $a$ ),

and spoken words( $t$ ) in each utterance. Zadeh et al. firstly proposed Tensor Fusion Network (TFN) [37] and Memory Fusion Network (MFN) [38] to fuse the multimodal features on sequential level. Recent works are mostly committed to calculate correlations among involved modalities. For example, Chen et al. [2] developed a Sentimental Words Aware Fusion Network (SWAFN) to compute the coattention between text and other modalities. Deng et al. [5] proposed a deep dense fusion network with multimodal residual (DFMR) to integrate multimodal information in a paired manner. With the significant progress of Transformer [34] in natural language processing and computer vision, some works [4, 11, 35] applied its self-attention mechanism to perform modality interactions. For example, Delbrouck [3] designed a Transformer-based joint-encoding (TBJE) that takes acoustic and textual features as input and jointly encode two modalities. Tsai [33] proposed a Multimodal Transformer (MulT) and attempted to calculate cross-modal attention between paired modalities( $t$  with  $a$ , for instance).

### 2.2 Multimodal Depression Detection

Compared to multimodal sentiment analysis, depression detection requires longer time sequences because it is a persistent long-term characteristic obtained from a human. Joshi et al. [15] used the bag-of-words model to encode acoustic and visual features and then fused them via principal component analysis (PCA) and support vector mechanisms (SVM). Rodrigues et al. [26] used audio-translated texts with their hidden embedding extracted from pre-trained BERT [6] model, and employ CNNs to obtain cross-modality information. Kaya et al. [16] designed a new Automatic Speech Recognizer (ASR) transcription based features while Ray [24] proposed a multi-layer attention network for estimating depressions. Aside from acoustic, visual, and textual features, Kroenke et al. [17] showed that body gestures have a positive contribution to the accuracy of depression estimation. Furthermore, Sun et al. [29] employed Transformer model to extract multimodal features and proposed an adaptive late fusion scheme for the final predictions. Zhao [42] introduced a hybrid feature extraction structure which combines self-attention and 3D convolutions for different kinds of features.

### 2.3 MLP-based Models

MLP-based models are new proposed structures for vision tasks. In ViT [7], Transformer is firstly used for image processing. It splits images into several patches and feeds them to Transformer. The significant performance boost prompts many other variations [1, 20], but the computational burden is still huge. Therefore, some MLP-based models are proposed including MLP-Mixer [31], ResMLP [32], and Hire-MLP [10]. These methods discard the self-attention mechanism and replace with MLPs, which is more efficient. In general, two independent MLPs are included in these models, one MLP processes channels while the other processes the tokens.

Inspired of the inherent structure of MLP-based models, we find it can be naturally transferred to multimodal feature processing. Because multimodal features usually consist of three axis (sequential, modality, and channel), we add an additional MLP to comprehensively mix the features. The detailed structures are illustrated in Section 3.

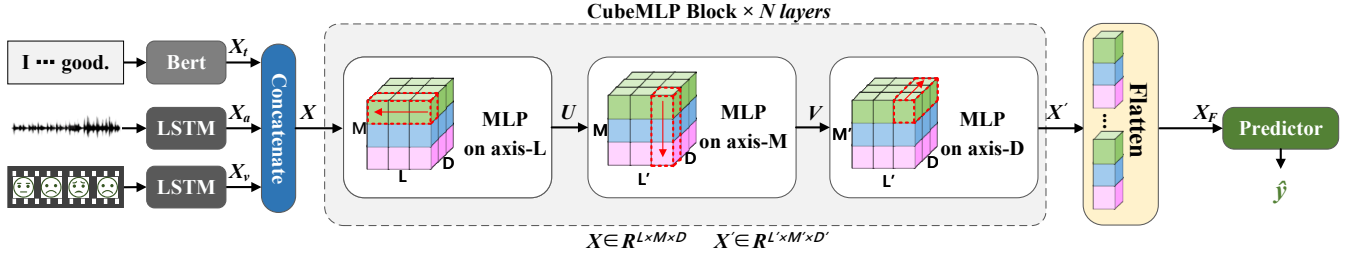


Figure 1: The overview of the MLP-based mixer. The extracted features are fed to  $N$  layer stacked CubeMLP blocks to be mixed. The mixed features are flattened for the prediction.

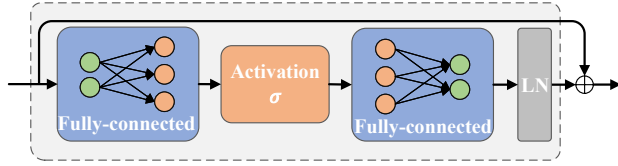


Figure 2: The structure of MLP unit. LN means the layer normalization. A residual short-cut is employed in the unit.

### 3 PROPOSED METHODS

CubeMLP is a simple but effective multimodal feature processing structure. Our task is to predict sentiment tendency or depression level from human utterances in videos, where each utterance is an input sample to the model. In one utterance, three modalities are provided including textual( $t$ ), acoustic( $a$ ) and visual( $v$ ). The overview of our approach is shown in Figure 1. Each modality features are first extracted by specific methods. After features being extracted, we don't exchange cross-modality information on sequential level or channel level, as in previous approaches [4, 41]. Instead, we perform the mixing individually on sequential, channel, and modality levels. Specifically, the CubeMLP is used to *mix up* multimodal features on all axis by respective MLP units. The mixed multimodal features are then passed to the classifier to perform the predictions for sentiment analysis or depression detection.

#### 3.1 Feature Extraction

Before mixing the modalities, LSTM [13] is employed to extract acoustic and visual features while pretrained BERT [6] is used for extracting textual features. The extracted features are represented as  $X_m \in R^{L \times D}$  in which  $L$  denotes the length of the utterance,  $D$  denotes the respective feature channel dimension and  $m \in \{t, a, v\}$ . The extracted modality features share the same sequential lengths and feature channel dimensions. The goal is to mix the modality features  $X_t$ ,  $X_a$ , and  $X_v$ , and predict the continuous affective tendency  $\hat{y} \in R$  for each utterance.

#### 3.2 The Structure of CubeMLP

After the features being extracted, we first extend  $X_m \in R^{L \times D}$  at the second dimension into  $R^{L \times 1 \times D}$  and then concatenate them to comprise the multimodal features  $X \in R^{L \times M \times D}$  along the extended axis, where  $M$  is the number of modalities. The multimodal features are then passed to  $N$  layer stacked CubeMLPs to be mixed-up as

shown in Figure 1. The CubeMLP block consists of three MLP units, and each MLP unit is designed to *mix* the multimodal features on its respective axis. Specifically, the first MLP  $f_L : R^{L \times M \times D} \rightarrow R^{L' \times M \times D}$  aims to perform *sequential-mixing* which acts on  $L$  axis. The second one  $f_M : R^{L' \times M \times D} \rightarrow R^{L' \times M' \times D}$  is *modality-mixing* MLP and acts on  $M$  axis. The third MLP  $f_D : R^{L' \times M' \times D} \rightarrow R^{L' \times M' \times D'}$  executes the *channel-mixing* on  $D$  axis. Each MLP unit is composed of two fully-connected layers and a nonlinear activation as shown in Figure 2. The fully-connected layers can also be treated as two affine transformations.

Let consider the first sequential-mixing MLP unit on  $L$  axis. Tensor  $X$  can be seen as a set of vectors of  $X_{*,m,d} \in R^{L \times 1 \times 1}$ , where  $(m, d) \in \{(1, 1), (1, 2), \dots, (2, 1), (2, 2), \dots, (M, D)\}$ .  $X_{*,m,d}$  is the vector of  $m$ -th modality and  $d$ -th channel. Each affine transformation in the sequential-mixing MLP can be represented as:

$$Aff_L(X_{*,m,d}) = W_L X_{*,m,d} + B_L, \quad (1)$$

where  $W_L \in R^{L' \times L}$  and  $B_L \in R^{L'}$  are two matrix-represented learnable parameters.  $L'$  is the reduced dimension on  $L$ -axis, which is a hyperparameter and will be discussed in more detail in Section 5.3. Equation 1 indicates that all  $X_{*,m,d}$  share parameters  $W_L$  and  $B_L$ . Therefore, the complete MLP unit can be mathematically represented as:

$$U_{*,m,d} = LN(Aff_L(\sigma(Aff_L(X_{*,m,d}))) + X_{*,m,d}) \in R^{L' \times 1 \times 1}, \quad (2)$$

for  $(m, d) \in \{(1, 1), (1, 2), \dots, (2, 1), (2, 2), \dots, (M, D)\}$ ,

where  $\sigma$  is the nonlinear activation,  $LN$  is the layer normalization. The output tensor  $U \in R^{L' \times M \times D}$  of the first MLP unit can be considered as a set of vectors of  $U_{*,m,d} \in R^{L' \times 1 \times 1}$ , where  $(m, d) \in \{(1, 1), (1, 2), \dots, (2, 1), (2, 2), \dots, (M, D)\}$ .

As well as the first MLP unit on the axis of  $L$ , the output  $V \in R^{L' \times M' \times D}$  of the second MLP on  $M$ -axis and the output  $X' \in R^{L' \times M' \times D'}$  of the third MLP on  $D$ -axis can be considered as a set of their elementary vectors  $V_{l,*,d} \in R^{1 \times M' \times 1}$  and  $X'_{l,m,*} \in R^{1 \times 1 \times D'}$ , where  $M'$  and  $D'$  are reduced dimensions on  $M$ -axis and  $D$ -axis, respectively, which will also be discussed in Section 5.3. As well as Equation 2,  $V_{l,*,d}$  and  $X'_{l,m,*}$  can be calculated as Equation 3 and 4, respectively.

$$V_{l,*,d} = LN(Aff_M(\sigma(Aff_M(U_{l,*,d}))) + U_{l,*,d}) \in R^{1 \times M' \times 1}, \quad (3)$$

for  $(l, d) \in \{(1, 1), (1, 2), \dots, (2, 1), (2, 2), \dots, (L', D)\}$ ,

$$X'_{l,m,*} = LN(Aff_D(\sigma(Aff_D(V_{l,m,*}))) + V_{l,m,*}) \in R^{1 \times 1 \times D'}, \quad (4)$$

for  $(l, m) \in \{(1, 1), (1, 2), \dots, (2, 1), (2, 2), \dots, (L', M')\}$ .

The  $X' \in R^{L' \times M' \times D'}$  is the mixed multimodal features.

### 3.3 Prediction Head

Following the previous works [11, 37], the mixed multimodal features  $X' \in R^{L' \times M' \times D'}$  are flattened as  $X' \in R^{L'M'D'}$ . Then the flattened features are fed to the classifier  $f_c : R^{L'M'D'} \rightarrow R$  to predict the sentiment tendency or depression level.

We use the mean absolute error (MAE) as the loss function during training for multimodal sentiment analysis, which is a regression task:

$$\mathcal{L}_{mae} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (5)$$

where  $N$  is the number of samples. The absolute errors between the prediction and the ground truth are calculated using MAE. MAE has a substantial effect on small errors than higher-order errors, allowing models to achieve better accuracy on subtle sentiments. As a result, MAE is often used as the key performance metric for sentiment analysis [2, 11, 12, 37].

For depression detection, we train the model to regress the depression tendency with Concordance Correlation Coefficient (CCC) loss as the loss cost:

$$\mathcal{L}_{CCC} = 1.0 - \frac{2S_{\hat{y}y}}{S_y^2 + S_{\hat{y}}^2 + (\bar{\hat{y}} - \bar{y})^2} \quad (6)$$

CCC is widely used for estimating depressions [15, 26], because it is not only unbiased by changes in scale and location, but also includes measure on both correlation and accuracy [18].

## 4 EXPERIMENTS

### 4.1 Datasets

We conduct the experiments on two multimodal mind state estimation tasks: sentiment analysis and depression detection, as previous works find that there is a strong connection between them [23]. For multimodal sentiment analysis, we use two popular benchmark datasets: CMU-MOSI [40] and CMU-MOSEI [41]. For multimodal depression detection, we use the AVEC2019 dataset [25] to evaluate the effectiveness of CubeMLP.

**4.1.1 CMU-MOSI.** CMU-MOSI [40] is a popular dataset for multimodal sentiment analysis. The samples in CMU-MOSI are utterance-based videos that are collected from the Internet. In each sample the speakers express their subjective opinions on certain topics. 1283 training utterances, 229 validation utterances, and 686 test utterances are provided in the dataset. The dataset is annotated with sentiment tendencies in the interval  $[-3, 3]$ .

**4.1.2 CMU-MOSEI.** CMU-MOSEI [41] is enlarged from the CMU-MOSI. It has the same annotations as the CMU-MOSI. In CMU-MOSEI, there are 16315 utterances for training, 1817 utterances for validation, and 4654 utterances for testing.

**4.1.3 AVEC2019.** The AVEC2019 DDS dataset [25] is obtained from audiovisual recordings of patients' clinical interviews. The interviews are conducted by a virtual agent to preclude human interference. Different from the above two datasets, each modality in AVEC2019 provides several different kinds of features. For example, acoustic modality consists of MFCC, eGeMaps, and deep features extracted by VGG [28] and DenseNet [14]. In previous researches [29], Hao et al. found that MFCC and AU-poses are two most discriminant features in acoustic and visual modalities, respectively. Therefore, for the purpose of simplicity and efficiency we just employ MFCC and AU-poses features for depression detection. The dataset is annotated by PHQ-8 scores in the interval  $[0, 24]$  and bigger PHQ-8 score means the depression tendency is more severe. There are 163 training samples, 56 validation samples, and 56 test samples in this benchmark dataset.

### 4.2 Experimental Setup

For multimodal feature extraction, we set  $L$  to 100 for sentiment analysis and 1000 for depression. Because the lengths of the samples vary, we zero pad smaller sequences and cut off longer sequences to match the length. For each modality features, we set  $D$  to 128. In our study,  $M$  is fixed to 3 because we have three involved modalities ( $t$ ,  $a$ , and  $v$ ). For  $L'$ ,  $M'$ , and  $D'$ , we explore them in Section 5.3. In the experiments, we find CubeMLP is a so effective structure that the performance can reach the state-of-the-art when we just set  $N$  to 3. During the training, the learning rate is initialized to 0.004 and multiplied by 0.1 every 50 epochs. Our models are implemented with PyTorch [22] framework and validated on two GTX 1080Ti GPU cards.

### 4.3 Evaluation Metrics

**4.3.1 CMU-MOSI & CMU-MOSEI.** The CMU-MOSI and CMU-MOSEI provide sentiment regression tasks. Following recent works [2, 5], we provide MAE and Pearson correlation (Corr) as measurements. The consecutive sentiment tendency can also be transferred to binary classification task (positive and negative) and 7-class classification task (rounded tendency, e.g., 1.8 is in class-2). For the classification tasks, we provide accuracy (Acc) and F1-score (F1) as measurements.

**4.3.2 AVEC2019 DDS.** The CCC and MAE metrics are used to evaluate the AVEC2019 DDS dataset, which have been utilized in previous depression detection studies. The formula expression of CCC is represented as:

$$CCC = \frac{2S_{\hat{y}y}}{S_y^2 + S_{\hat{y}}^2 + (\bar{\hat{y}} - \bar{y})^2} \quad (7)$$

The CCC is in the interval  $[-1, 1]$  and -1 represents the total negative correlation while 1 means the perfect positive correlation.

## 5 RESULTS & ANALYSIS

### 5.1 Experimental Results

Our experimental results of multimodal sentiment analysis are shown in Table 1. Table 1 also compares the results of previous

**Table 1: The results on two multimodal sentiment analysis benchmark datasets, CMU-MOSI and CMU-MOSEI.**

Models	CMU-MOSI					CMU-MOSEI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
TFN[37]	0.970	0.633	73.9	73.4	32.1	0.593	0.700	82.5	82.1	50.2
MFN[38]	0.965	0.632	77.4	77.3	34.1	-	-	76.0	76.0	-
ICCN[30]	0.862	0.714	83.0	83.0	39.0	0.565	0.713	84.2	84.2	51.6
SWAFN[2]	0.880	0.697	80.2	80.1	40.1	-	-	-	-	-
MuT[33]	0.871	0.698	83.0	82.8	40.0	0.580	0.703	82.5	82.3	51.8
LMF-MuT[27]	0.957	0.681	78.5	78.5	34.0	0.620	0.668	80.8	81.3	49.3
MAT[3]	-	-	-	80.0	-	-	-	82.0	82.0	-
MNT[3]	-	-	-	80.0	-	-	-	80.5	80.5	-
MISA[12]	0.817	0.748	82.1	82.0	41.4	0.557	0.748	84.9	84.8	51.7
BBFN[11]	0.776	0.755	84.3	84.3	45.0	<b>0.529</b>	<b>0.767</b>	<b>86.2</b>	<b>86.1</b>	54.8
CubeMLP(Ours)	<b>0.770</b>	<b>0.767</b>	<b>85.6</b>	<b>85.5</b>	<b>45.5</b>	<b>0.529</b>	0.760	85.1	84.5	<b>54.9</b>

**Table 2: The results on the test set of depression detection benchmark dataset, AVEC2019 DDS.**

Models	AVEC2019	
	CCC(↑)	MAE(↓)
Baseline [25]	0.111	6.37
Adaptive Fusion Transformer [29]	0.331	6.22
EF [16]	0.344	-
Bert-CNN & Gated-CNN [26]	0.403	6.11
Multi-scale Temporal Dilated CNN [8]	0.430	4.39
Hierarchical BiLSTM [36]	0.442	5.50
CubeMLP(Ours)	<b>0.583</b>	<b>4.37</b>

studies according to their provided papers. As is shown in the table, CubeMLP achieves a MAE of 0.770 on CMU-MOSI and 0.529 on CMU-MOSEI, which is competitive with other state-of-the-art approaches. Among all the methods, TFN [37] and MFN [38] are tensor-based approaches that process the sequential signals from three modalities at the same time. ICCN [30] uses mathematical metrics to determine the relationship between modalities. MuT [33] and LMF-MuT [27] employ stacked Transformers to expand the available temporal frames for alignment. Furthermore, MAT and MNT [3] use Transformer’s self-attention mechanism to simultaneously encode two types of modality features. SWAFN [2] calculates the corresponding co-attention for two modality pairings (text-visual and text-acoustic) in order to share cross-modality information between modalities. MISA [12] attempts to learn modality-invariant and modality-specific representations while BBFN [11] learns complementary information using two symmetric Transformer-based structure.

For multimodal depression detection, our results are shown in Table 2. On the AVEC2019 DDS dataset, we achieve a CCC of 0.583 and an MAE of 4.37. As for other approaches, Baseline [25] uses late fusion and averages the final predictions from involved modalities. To adaptively fuse the final predictions, Sun et al. proposed the adaptive fusion Transformer network. Sun et al. [29] propose the adaptive fusion Transformer network to adaptively fuse the

final predictions. EF [16] introduces simple linguistic and word-duration features to estimate the depression level. Bert-CNN & Gated-CNN [26] employs the gate mechanism to fuse the information attained from involved modalities. Multi-scale Temporal Dilated CNN [8] uses the dilated CNN to extract multimodal features using a larger receptive field. Hierarchical BiLSTM [36] applies the hierarchical biLSTM to capture the sequential information in a pyramid-like structures.

## 5.2 Ablation Study

To study the effectiveness of each MLP units, we perform the ablation studies on the CMU-MOSI datasets shown in Table 3. We can infer from the results that the sequential-mixing (MLP-L) and channel-mixing (MLP-D) play more important roles than modality-mixing (MLP-M). The accuracy tends to get better when combining more MLPs. The performance is best when performing the mixing at all three axis ( $L$ ,  $M$ , and  $D$ ).

## 5.3 Dimension Exploration

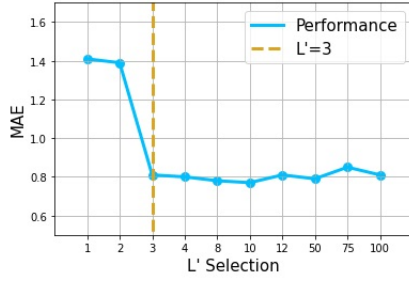
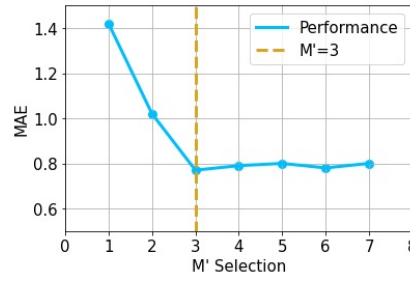
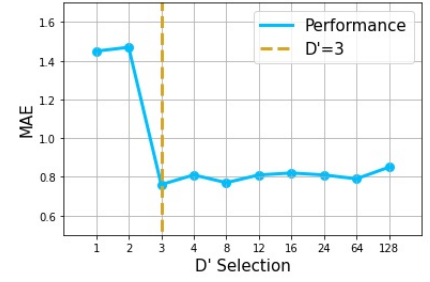
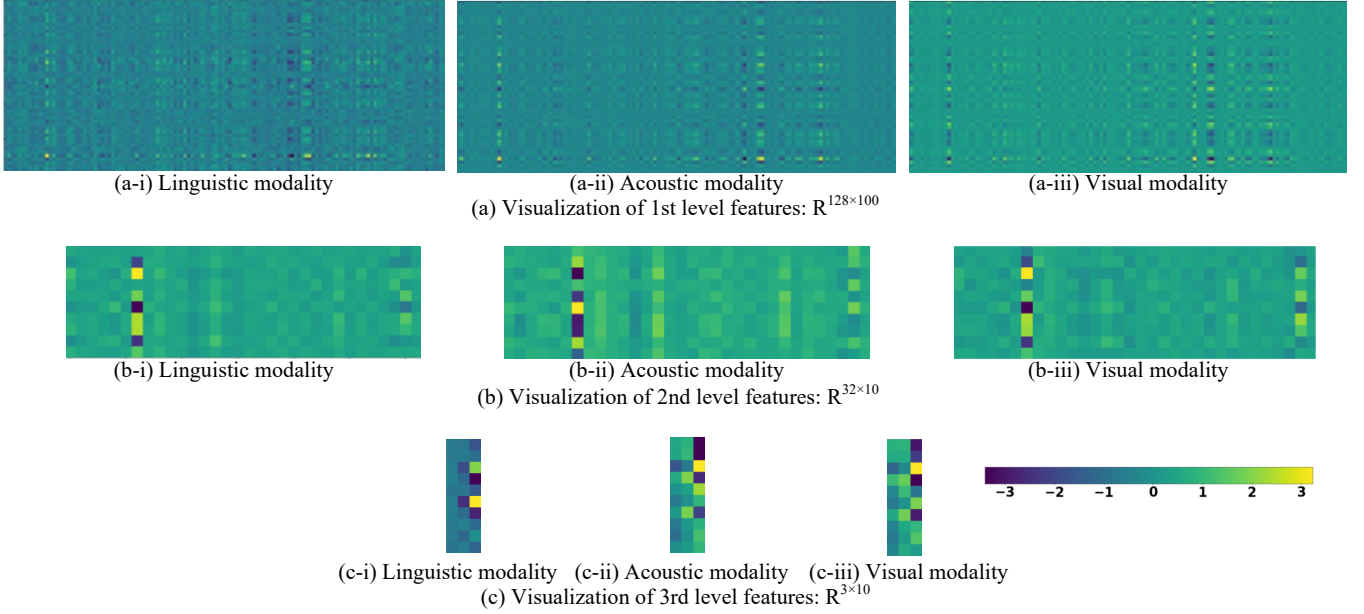
Because of the MLPs’ inherent structures, we can tune the output dimensions of each MLP ( $L'$ ,  $M'$ , and  $D'$ ). From a mathematical point of view,  $L'$ ,  $M'$ , and  $D'$  represent the dimensions after the features’ linear (affine) transformation on respective axis. As a result, we investigate, how the selection affects the final prediction accuracy. The visualization results are shown in Figure 3. It is evident that the performance decreases if the output dimensions are less than three ( $< 3$ ) on any axis. However, the performance fluctuation is substantially smaller if the output dimensions are  $\geq 3$ . We consider that the reason for this phenomenon is the fact that we only have three modalities involved during training ( $M = 3$ ). In other words, when treating the multimodal features as a tensor with shape  $R^{L \times 3 \times D}$  ( $L > 3, D > 3$ ), the rank of this tensor is 3. As a result, there is little information loss while converting this tensor to  $R^{3 \times 3 \times 3}$  using affine transformations on all axis, resulting in constant performance but a significantly lower computing cost.

## 5.4 Feature Visualization

To further prove that the features of different modalities can be mapped to similar hidden spaces by CubeMLP, we visualized the

**Table 3: The ablation study of our methods on CMU-MOSI [40]. MLP-L means the sequential-mixing, MLP-M means the modality-mixing, and MLP-D means the channel mixing.**

	MLP-L	MLP-M	MLP-D	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Model 1	✓			0.860	0.744	80.6	80.7	39.0
Model 2		✓		0.850	0.729	80.3	80.4	39.2
Model 3			✓	0.910	0.717	81.7	81.8	39.0
Model 4	✓	✓		0.806	0.753	81.5	81.6	42.4
Model 5	✓		✓	0.803	0.750	80.6	80.8	39.5
Model 6		✓	✓	0.874	0.718	82.4	82.4	41.6
Model 7(Ours)	✓	✓	✓	<b>0.770</b>	<b>0.767</b>	<b>85.6</b>	<b>85.5</b>	<b>45.5</b>

(a) The exploration of the selection of  $L'$ .(b) The exploration of the selection of  $M'$ .(c) The exploration of the selection of  $D'$ .**Figure 3: The output dimension ( $L'$ ,  $M'$ , and  $D'$ ) exploration of each MLP.****Figure 4: The visualization of each modality's features processed by CubeMLP at different levels. It can be seen that the semantic feature maps are more consistent and denser, as the model level goes deeper.**

features of the model on the CMU-MOSEI dataset. The model in this experiment is formed by three CubeMLP Blocks ( $N = 3$ ). For the three blocks,  $D'$  is set to  $[128, 32, 3]$ , while  $L'$  is set to  $[100, 10, 10]$

respectively. Since CMU-MOSEI has three modalities,  $M'$  is always set to 3 all along. The features of different CubeMLP blocks are visualized in Figure 4. For the convenience of comparison between

**Table 4: The comparison of memory(space) consumption between CubeMLP and other state-of-the-art approaches.**

Models	Space Consumption
TFN [37]	$O(L^M)$
Adaptive Fusion Transformer [29]	$O(L^2)$
ICCN[30]	$O(L \times D^2)$
MNT& MAT[3]	$O(L^2)$
BBFN[11]	$O(L^2)$
CubeMLP(Ours)	$O(\max(L, M, D))$

modalities, each  $R^{L' \times M' \times D'}$  feature is split into  $M'$  two-dimensional  $L' \times D'$  images.

From the visualization result, we can find that in the first CubeMLP Block, features of different modalities still have some noticeable differences between each other((a-i), (a-ii), and (a-iii)). However, as the feature size getting smaller and more semantic information is extracted, the output of the second CubeMLP Block tends to have similar representations for the three different modalities((b-i), (b-ii), and (b-iii)). Finally, the third CubeMLP Block outputs condensed semantic information and the visualization of different modalities' feature become furthermore similar((c-i), (c-ii), and (c-iii)). This indicates that after the processing of CubeMLP, the features of different modalities can be efficiently mapped to a similar hidden space. To put it another way, the multimodal information is better transmitted and semantic features are refined, resulting in a superior modality fusion result. With such fused information, downstream tasks can achieve a higher accuracy.

## 5.5 Computational Complexity Comparison

We compare computational space complexity of CubeMLP with other state-of-the-art techniques. The results are shown in Table 4. One of the advantages of MLP-based structures is its low computational cost. CubeMLP improves performance while utilizing less computational memory, demonstrating that it is a cost-efficient and effective multimodal fusion structure.

## 6 CONCLUSION

In this paper, we treat multimodal fusion as feature mixing and propose the MLP-based CubeMLP for unified multimodal feature processing. In CubeMLP, we perform the mix-up at all axis of multimodal features. CubeMLP can reach the state-of-the-art performance for sentiment analysis and depression detection while keeping the computational burden low. We analyzed CubeMLP's components and compared it to other techniques. Not only that, We conducted extensive ablation studies and visual analysis to illustrate MLP's efficiency and multimodal processing capabilities. We intend to do more MLP research in the future, as well as more experiments in other multimodal fusion domains, such as emotion detection. MLP, we believe, will become a widely used solution for a variety of tasks.

## ACKNOWLEDGMENTS

Thanks to Rahul for the careful English proofing and language grammar check. Thanks to the paper reviewers for the constructive research comments.

## REFERENCES

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.
- [2] Minping Chen and Xia Li. 2020. SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1067–1077.
- [3] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. *arXiv preprint arXiv:2006.15955* (2020).
- [4] Jean-Benoit Delbrouck, Noé Tits, and Stéphane Dupont. 2020. Modulated fusion using transformer for linguistic-acoustic emotion recognition. *arXiv preprint arXiv:2010.02057* (2020).
- [5] Huan Deng, Peipei Kang, Zhenguo Yang, Tianyong Hao, Qing Li, and Wenyin Liu. 2021. Dense Fusion Network with Multimodal Residual for Sentiment Classification. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [8] Weiquan Fan, Zhiwei He, Xiaofen Xing, Bolun Cai, and Weirui Lu. 2019. Multi-modality depression detection via multi-scale temporal dilated cnns. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 73–80.
- [9] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2018. NIH Public Access, 2225.
- [10] Jianyuan Guo, Yehui Tang, Kai Han, Xinghao Chen, Han Wu, Chao Xu, Chang Xu, and Yunhe Wang. 2021. Hire-mlp: Vision mlp via hierarchical rearrangement. *arXiv preprint arXiv:2108.13341* (2021).
- [11] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 6–15.
- [12] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.
- [13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [15] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces* 7, 3 (2013), 217–228.
- [16] Heysem Kaya, Dmitrii Fedotov, Denis Dresvyanskiy, Metehan Doyran, Danila Mamontov, Maxim Markitantov, Alkim Almila Akdag Salah, Evrim Kavcar, Alexey Karpov, and Albert Ali Salah. 2019. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 27–35.
- [17] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. , 509–515 pages.
- [18] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.
- [19] Xia Li and Minping Chen. 2020. Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language. In *China National Conference on Chinese Computational Linguistics*. Springer, 359–373.
- [20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.
- [21] Sijie Mai, Songlong Xing, and Haifeng Hu. 2021. Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1424–1437.
- [22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019.

- Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- [23] Syed Arbaaz Qureshi, Gael Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine* 15, 3 (2020), 47–59.
  - [24] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*. 81–88.
  - [25] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 3–12.
  - [26] Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 55–63.
  - [27] Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman. 2020. Low Rank Fusion based Transformers for Multimodal Sequences. *arXiv preprint arXiv:2007.02038* (2020).
  - [28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
  - [29] Hao Sun, Jiaqing Liu, Shurong Chai, Zhaolin Qiu, Lanfen Lin, Xinyin Huang, and Yenwei Chen. 2021. Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors* 21, 14 (2021), 4764.
  - [30] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.
  - [31] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems* 34 (2021).
  - [32] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. 2021. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404* (2021).
  - [33] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
  - [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
  - [35] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*. 2514–2520.
  - [36] Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang. 2019. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 65–71.
  - [37] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).
  - [38] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
  - [39] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
  - [40] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.
  - [41] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.
  - [42] Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn W Schuller. 2020. Hybrid Network Feature Extraction for Depression Assessment from Speech.. In *INTERSPEECH*. 4956–4960.