A Multitask Learning Framework for Multimodal Sentiment Analysis

Dazhi Jiang†, Runguo Wei†, Hao Liu†, Jintao Wen†, Geng Tu†, Lin Zheng†, Erik Cambria§
†Department of Computer Science, Shantou University, Shantou, China
§School of Computer Science and Engineering, NTU, Singapore
{dzjiang,20rgwei,20hliu2,20jtwen,19gtu,lzheng}@stu.edu.cn, cambria@ntu.edu.sg

Abstract—Mapping continuous dimensional emotion to discrete classes is an extremely difficult task. In this paper, we predict the intensity classes of emotions based on valence and arousal in segments of audio-visual recordings about car reviews. Consequently, for unimodal features, we first employ baseline methods and principal component analysis to search for the best unimodal features in different modalities, which can simplify the relationship between feature attributes. For multimodal features, we perform multimodal fusion on the best and other unimodal features through an early fusion strategy. For sentiment analysis, we propose six hybrid temporal models for modeling complex time dependencies. To avoid overfitting the validation set and providing complementary information between different modalities, we propose a multitask learning framework, which can adaptively change the weight of loss per subtask.

Index Terms—Multimodal sentiment analysis; Multitask learning

I. INTRODUCTION

With the development of social networks, the ways people convey their emotions are becoming increasingly diverse, multifaceted, and multimodal. However, how to analyze sentiment from multimodal data is an opportunity and challenge in the field of affective computing. Fortunately, many excellent works [1]–[3] and datasets [4]–[6] have been proposed recently on multimodal sentiment analysis, which are constantly promoting this field. In affective computing, there are two mainstream models to describe the emotion, one is the categorical model, the other is the dimensional model [7]. For categorical models, Ekman et al. [8] divided each emotion into independent labels such as joy, sadness, fear, and other emotions, which has natural interpretability, but the differences and relations between labels are not able to compute better. Therefore, Hanjalic et al. [9] divided emotion into two dimensions: arousal and valence, which is more suitable than the categorical model in computing affective. The greater the value of arousal and valence, the more positive the emotion is, and vice versa. Especially, Cambria et al. [10] proposed an hourglass-shaped model that is both discrete and dimensional.

This work was supported by National Natural Science Foundation of China (61902232, 61902231), Natural Science Foundation of Guangdong Province (2019A1515010943), Key Project of Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Natural Science) (2018KZDXM035), The Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence) (2019KZDZX1030) and 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D).

Geng Tu is the corresponding author (e-mail: 19gtu@stu.edu.cn).

MuSe-Sent, sub-challenge of MuSe 2021, requires participants to predict the corresponding five advanced sentiment classes for arousal and valence as shown in Fig. 1 and encourage participants to use multimodal information for a more robust multimedia content emotion analysis. Actually, mapping continuous dimensional emotion to discrete classes is an extremely difficult task. So far, dimensional emotion has been successfully calculated only in a time-continuous fashion [11]. In the previous Muse challenge, Li et al. [12] explored various low-level descriptors (LLDs) and depth features in a different modality and proposed several effective multimodal fusion strategies. Additionally, Sun et al. [13] extracted manual features and depth representation, and utilized a long short-term memory (LSTM) network and self-attention mechanism to capture time dependencies in the video.

In this paper, we construct a multitask learning framework for sentiment analysis. To be specific, we first utilize the baseline methods and principal component analysis (PCA) to obtain the best unimodal features in different modalities. Then, we propose various hybrid temporal models to capture the time dependences of multimodal features. Noticeably, we adopt the early fusion strategy to fuse features in different modalities. Because the combination of different features and models has a great impact on the prediction results, we employ the two combinations with the best F1 score to conduct multitask learning without breaking the original best combination of features and models, which provides complementary information between different modalities and further improves the generalization ability.

Fig. 1: Task illustration: predicting the five sentiment classes for arousal and valence every 250ms in the video. The solid blue line and red dotted line indicate valence and arousal.
Additionally, our framework can adaptively change the weight of loss per subtask instead of regulating parameters manually. Our contribution to this challenge is as follows:

- We utilize baseline methods and PCA to explore various unimodal features so as to simplify the relationship between feature attributes, which paves the way for the subsequent multimodal feature fusion.
- We fuse the best and some other unimodal features based on an early fusion strategy for obtaining multimodal features. Additionally, for performing sentiment analysis, six hybrid temporal models are proposed to model complex time dependencies in the videos.
- We proposed a multitask learning framework to make up for the deficiency of the early fusion at capturing the complementary information between different modalities. Moreover, the generalization ability of the model has been further improved.

II. RELATED WORK

Multimodal Features: For multimodal features, some studies are still adopted handwritten features. For example, In AVEC 2013, Lozano et al. [14] extracted Gabor features and local binary patterns from visual modality. Sun et al. [13], the winner of MuSe 2020, extracted handwritten acoustic features such as the IS13 feature from audio. In recent years, because of the strong representation ability of the deep network models, researchers prefer to utilize deep networks to learn the representation of multimodal data. For example, in EmotiW 2019, Zhou et al. [15] employed three CNN backbones and AlexNet to extract visual and audio features, respectively. In AVEC 2018, the depth audio representation generated by the VGGish model, which is better than the acoustic features based on expert knowledge [16]. Additionally, in multimodal sentiment analysis, text modality also plays a vital role [17]. In AVEC 2017, various word vectors model such as Word2Vec [18] and GloVe [19] are widely used.

Multimodal Fusion: Multimodal fusion strategies are always one of the research focuses in multimodal sentiment analysis. In MuSe 2020, Sun et al. [13] combined early fusion and late fusion strategies. Li et al. [12] proposed a variety of effective multimodal fusion strategies to integrate LLDs and depth features. Additionally, Zadeh et al. [20] presented a tensor fusion network (TFN) to capture the dependency relationships within and between three modality data. Yang et al. [21] introduced a modal temporal attention graph, which can convert misaligned multimodal sequence data into a graph with heterogeneous nodes and edges, which can obtain rich information across modalities and time. And Hazarika et al. [22] proposed a new framework, MISA, which projects each modality data into two different subspaces for multimodal fusion.

Model Architecture: Because recurrent neural networks (RNNs) show extraordinary advantages in sequence modeling, in MuSe 2020, the winners without exception adopted a LSTM network for continuous dimension emotion recognition [12], [13], which is a variant of RNNs. Moreover, some researchers also put forward a lot of novel work. For example, Zadeh et al. [23] proposed long-short term hybrid memory on the basis of LSTM. Chaturvedi et al. [3] introduced a combined model of convolutional neural network and fuzzy logic for predicting the degree of a specific emotion. In addition, because of the dependency between each emotion-related subtask, sentiment analysis will perform better in a suitable joint framework. And people pay increasing attention to emotion analysis based on multitask learning recently. For example, Akhtar et al. [24] proposed a multitask learning framework, which completes four emotion and sentiment analysis tasks together, such as "3-class categorical & 5-class ordinal classification for sentiment". And in the multitask learning framework, all experimental results are better than the single task framework.

III. METHODOLOGY

A. Task Definition

Let \( X_j^{(k)} = [A_j^{(k)}, V_j^{(k)}, D_j^{(k)}] \) represents the multimodal feature extracted from the jth segment of the kth video, where \( A_j^{(k)}, V_j^{(k)}, D_j^{(k)} \) represents the unimodal features from acoustic vision and text modalities. Then \( k \in \{1, 2, \ldots, K\} \) and \( K \) denotes the number of videos. Additionally, let \( Y_j^{(k)} = \{y_j^{(k)}\} \) is the corresponding sentiment label of valence or arousal, where \( j \in \{1, 2, \ldots, N\} \) and \( N \) stands for the number of video segments. According to the overall framework shown in Fig. 2 show, (1) we first apply baseline methods and PCA to explore the best unimodal features \( \bar{A}_j, \bar{V}_j, \bar{D}_j \) (2) Then, we fuse the best and some other unimodal features based on an early fusion strategy and conduct experiments in a variety of hybrid deep temporal models. (3) Finally, we design a multitask learning framework that can adaptively change the loss weights of subtasks to predict sentient classes for each emotion dimension. And the goal of multimodal sentiment analysis in this paper is to maximize the following function:

\[
\Phi = \prod_{i=1}^{T} \prod_{j=1}^{N} \prod_{k=1}^{K} p\left(y_j^{(i,k)} | \bar{A}_j^{(k)}, \bar{V}_j^{(k)}, \bar{D}_j^{(k)}; \theta\right) \tag{1}
\]

where, \( y_j^{(i,k)} \) represents the sentiment label of valence or arousal corresponding to the jth video segment of the kth video in the ith subtask. \( T \) denotes the number of subtasks, and \( \theta \) is the set of model parameters in all subtasks.

B. Multimodal Features

Emotion can be conveyed in various modalities. For example, the changes of voice and intonation in speech, facial expressions, and body movements in vision, and semantic information in the text. In this section, we introduce the features adopted by our model.

1) Acoustic: eGeMAPS Feature: eGeMAPS, an extension of GeMAPS [25], adds some features on the basis of 18 acoustic LLDs, including 5 spectral features and 2 frequency-related features. In addition, a total of 88 statistical features can be obtained on these LLDs. DeepSpectrum Feature: DeepSpectrum feature can be obtained by feeding the spectral map into the pre-trained image convolutional neural network.
(such as VGGNet [26]), and it has been proved that it can capture useful emotional information in speech [27]. **VGGish Feature**: VGGish Feature can be obtained by feeding audio into the pre-trained VGGish model. Among them, VGGish [28] is a variant of VGGNet, which is pre-trained on AudioSet [29] dataset containing more than 2 million human-labeled video soundtracks and more than 600 audio event classes.

2) Vision: **VGGFace Feature**: VGGFace is the facial feature, which can be extracted by feeding the picture obtained by multitask revolutionary neural network (MTCNN) into the pre-trained VGGNet. the training data of the VGGNet consists of 2.6 million faces and more than 2500 identities. Compared with other face recognition models, VGGFace can consume less data and show higher performance. **Xception Feature**: Xception Feature is the environmental features provided by Xception [30] using stacked residual blocks. The network was pre-trained on an ImageNet dataset [31] containing 350 million images and 17000 categories.

3) Text: **BERT Feature**: The extraction process of BERT Feature adopts a transformer-based [32] bidirectional encoder BERT [33], which has been widely applied in various NLP tasks. BERT Feature can be obtained by feeding unlabeled text and their context into pre-trained BERT model. Especially, our feature of 768 dimensions is the sum of the last four BERT layers.

**C. Principal Component Analysis**

PCA is often used to reduce the dimension of high-dimensional data. Importantly, the data after the dimension reduction process can remove the noise and improve the quality. To be specific, PCA uses orthogonal transformation to replace the original n-dimensional features with m-dimensional features with fewer dimensions. These new features are the linear combination of the old features, which is the linear combination that maximizes the sample variance. It is worth noting that in the experiment, we employ the PCA algorithm based on eigenvalue decomposition of the covariance matrix. Let n-dimensional data $X = \{x_1, x_2, \ldots, x_n\}$ is the input and needs to be reduced to k-dimensional. The specific calculation process is as follows:

- Processing the data using the method of the mean-residual normalization.
- Calculating covariance matrix $XX^T$.
- Solving the eigenvalue and eigendirection of the covariance matrix $XX^T$ using the eigenvalue decomposition method.
- Sorting the eigenvalues, and selecting the largest $K$ of them. Then, the corresponding $K$ eigenvectors are used as row vectors to form the eigenvector matrix $P$.
- Mapping the data into a new space constructed by K eigenvectors, namely: $Y = PX$

**D. Multitask sentiment analysis model**

For achieving the more robust sentiment analysis model, we perform multitask learning based on the two best F1 score combinations of feature and hybrid temporal model.

1) **Hybrid temporal models**: To determine the network structure of the multitasking learning framework, we fused different unimodal features based on an early fusion strategy similar to the method in [11], and proposed six hybrid deep temporal models: Transformer (Way1), Self-Attention+Bi-LSTM (Way2), Transformer+Bi-LSTM (Way3), EfficientNet+Bi-LSTM (Way4), Self-Attention-EfficientNet+Bi-LSTM (Way5), Transformer+EfficientNet+Bi-LSTM (Way6).
Transformer: Like most seq2seq models, the structure of the transformer is also composed of an encoder and decoder.

- Encoder: The encoder consists of six layers with the same structure, and each layer contains two sub-layers: a multi-head self-attention and a fully connected feed-forward network. Admittedly, the Transformer can work partly because the multi-head attention mechanism plays an important role. The calculation process of multi-head attention mechanism is as follows:

\[ MH(Q, K, V) = \text{concat}(\text{head}_1, \text{head}_2, ..., \text{head}_n)W_h \]  

\[ \text{head}_i = \text{Att}(QW_i^Q, KW_i^K, VW_i^V) \]  

\[ \text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \]

where \( MH(.) \) and \( \text{Att}(.) \) represent the multi-head attention layer and attention layer, respectively. \( w_h, W_h^Q, W_h^K \) and \( W_h^V \) denote the learnable parameter matrices. \( Q, K, \text{and} V \) are the set of queries, keys, and values, respectively. In addition, the main difference between Self-Attention and traditional attention mechanisms is \( Q = K = V \).

- Decoder: Because the decoder needs to receive the global semantic information of the encoder and the preceding results of the model at the same time, the decoder adds an attention sub-layer with the mask on the basis of the encoder. Other structures of the decoder are the same as those of encoder.

Bi-LSTM: Bi-LSTM is a bidirectional LSTM, which makes up for the lack that traditional LSTM cannot encode information from back to front. Noteworthy, forward and backward information is all particularly important in emotion analysis tasks. Let the input is \( u_t \), the forward calculation process of Bi-LSTM is as follows:

\[ i_t = \sigma_s(W_iu_t + U_i\tilde{h}_{t-1} + b_i) \]  

\[ f_t = \sigma_s(W_fu_t + U_f\tilde{h}_{t-1} + b_f) \]  

\[ o_t = \sigma_s(W_ou_t + U_o\tilde{h}_{t-1} + b_o) \]  

\[ \tilde{h}_t = \sigma_h(o_t \odot (f_t \odot \tilde{h}_{t-1} + i_t \odot \sigma_h(W_cu_t + b_c))) \]

where, \( \sigma_s \) and \( \sigma_h \) are the sigmoid and tanh activation functions, respectively. \( \odot \) is element-wise multiplication \( i, f \) and \( o \) are the input, forget and output gate vectors, respectively. \( W, U, b \) and \( b \) are parameters of the model. \( h_t \) represents the hidden state of model. Similarly, we can obtain another hidden state \( \tilde{h}_t \). The concatenated result of the hidden states \( \vec{h}_t \) and \( \vec{h}_t \) is the final hidden state \( h_t = [\vec{h}_t; \tilde{h}_t] \).

EfficientNet: EfficientNet V2 [34] is a model scaling method. It uses a composite coefficient to enlarge the network from the three dimensions of depth, width, and resolution. Instead of the arbitrarily scaling method, it is based on neural structure search technology to obtain the optimal set of parameters (composite coefficients). Moreover, EfficientNet is not only much faster than other convolutional neural networks but also has higher accuracy. It is worth noting that in the experiment, in order to better model the complex time-dependences in the video, we use a one-dimensional convolution layer to replace the two-dimensional convolution layer in the original model.

2) Multitask learning framework:

In this section, the multitask learning framework is proposed for solving two problems: (1) the early fusion strategy is not able to make full use of the complementary information between multimodal data. (2) the proposed hybrid temporal models may overfit the verification set and thus lack generalization ability. The design idea of the framework is to find a common representation in the middle layer of the hybrid temporal models to fully complement information between multimodal data, and then independently predict the sentiment classes for each emotion dimension in each subtask. Additionally, the hybrid temporal models have better generalization ability in predicting sentiment classes for each emotion dimension, because of the information they share.

Since the loss of each subtask may behave differently, it is essential to balance each loss weight of subtasks. Most multitask learning framework adopts the following form of the combined loss function.

\[ \mathcal{L}_T(x, y_T, \hat{y}_T, W_T) = \sum_{\tau \in T} \mathcal{L}_\tau(x, y_T, \hat{y}_T, W_T) \cdot \lambda_\tau \]  

\[ y'_\tau = \text{softmax}(W_p h_\tau + b_p) \]  

\[ \hat{y}_\tau = \arg\max_k (y'_\tau[k]) \]

where, \( x \) represents the multimodal features. \( y_T \) and \( \hat{y}_T \) are the true and prediction labels, respectively. \( T \) denotes the set of subtasks. \( \tau \in T \) is the current subtask and \( \lambda_\tau \) is the corresponding loss weight. \( W_T \) and \( W_T \) are parameter matrixes of the model. In addition, in our proposed task learning framework, each subtask uses a fully connected network as a classifier. \( W_p \) and \( b_p \) represent the weight and bias of the classifier, respectively.

It is not difficult to find that in the above loss function, \( \lambda_\tau \) will greatly affect the final result. However, the manual adjustment process of parameter \( \lambda_\tau \) will take a lot of expenditure of time and effort. Therefore, we urgently need a strategy that can change the loss weight of each subtask efficiently, adaptively, and dynamically. Fortunately, inspired by [35], we designed the following loss function, which can automatically adjust the loss weight of each subtask based on homoscedastic uncertainty.

\[ \mathcal{L}_T(x, y_T, \hat{y}_T, W_T) = \sum_{\tau \in T} \frac{1}{2\delta^2_\tau} \mathcal{L}_\tau(x, y_T, \hat{y}_T, W_T) + \log (\delta_\tau) \]

\[ \mathcal{L}_T(x, y_T, \hat{y}_T, W_T) = \sum_{\tau \in T} \frac{1}{2\delta^2_\tau} \mathcal{L}_\tau(x, y_T, \hat{y}_T, W_T) + \log (1 + \delta_\tau) + \eta\|\theta\| \]

where, \( \eta \) is the L2 regularization term and \( \theta \) is the set of parameter matrixes of all subtasks.
TABLE I: Comparison between unimodal features and the features after PCA. A, V, T represent audio, video, and text.

<table>
<thead>
<tr>
<th>Features</th>
<th>Modality</th>
<th>Dimension</th>
<th>PCA Features</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpectrum</td>
<td>A</td>
<td>4096</td>
<td>DeepSpectrum</td>
<td>38</td>
</tr>
<tr>
<td>VGGish</td>
<td>A</td>
<td>128</td>
<td>VGGish</td>
<td>75</td>
</tr>
<tr>
<td>eGeMAPS</td>
<td>A</td>
<td>88</td>
<td>eGeMAPS</td>
<td>10</td>
</tr>
<tr>
<td>Xception</td>
<td>V</td>
<td>2048</td>
<td>Xception</td>
<td>304</td>
</tr>
<tr>
<td>VGGFace</td>
<td>V</td>
<td>512</td>
<td>VGGFace</td>
<td>34</td>
</tr>
<tr>
<td>FAU</td>
<td>V</td>
<td>35</td>
<td>FAU</td>
<td>14</td>
</tr>
<tr>
<td>BERT</td>
<td>T</td>
<td>768</td>
<td>BERT</td>
<td>412</td>
</tr>
</tbody>
</table>

IV. EXPERIMENT

A. Dataset

In the MuSe-Sent sub-challenge of MuSe 2021, participants need to predict five sentiment classes for each emotion dimension (arousal or valence) on a segment level, based on audio-visual recordings from the MuSe-CaR dataset [6]. It consists of 291 videos collected from YouTube. And the training set, verification set, and test set contain 166, 62, and 64 videos respectively in the MuSe-Sent sub-challenge.

B. Experimental Setup

In this paper, all the deep learning models used are based on the Pytorch framework. Additionally, the presented six hybrid temporal models, The detailed hyperparameters of the proposed six hybrid temporal models are the following: the initial value of learning rate, epoch, and optimizer is 0.001, 100, and Adam in all six hybrid temporal models. Especially, the activation function of EfficientNet is a sigmoid weighted liner unit (SiLU) but that of other models is ReLU. Moreover, in Transformer, the hidden size in the position-wise feed-forward layer and the number of heads are 128 and 4, respectively. As for Bi-LSTM, the hidden size is 64 which is the same as that of Self-attention.

C. The results of PCA

In this part, we specify the explained variance of PCA as 95% and conduct PCA on different modal features for the reduced dimensionality. The experimental results are shown in Table I. In the multi-dimensional feature vectors, some features contribute little to prediction results. So, our purpose is to eliminate these irrelevant or low correlation information to improve the quality of samples, which can also reduce noise and improve the robustness of the model.

D. The results of hybrid temporal model

Uni-modal Results: Firstly, we use the Self-Attention + Bi-LSTM model to obtain the best unimodal features under different modalities, namely: A: VGGish, V: FAU, T: BERT. Then, the best unimodal features after PCA are fed into six hybrid temporary models. From results shown in Table I, we can find that the quality of the BERT feature is significantly improved, while the performance of the other two features is just the opposite. In particular, because the BERT feature shows a surprising effect, the subsequent multimodal fusion process will also replace the original BERT feature with the BERT.

Bi-modal Results: We fuse the best unimodal features, based on early fusion strategy and conduct experiments in six hybrid temporal models. From results shown in Table II, we can see that when the feature is VGGish+BERT and the model is EfficientNet+Bi-LSTM, the best effect is achieved in the prediction of sentiment classes for arousal. However, when the model or feature changes, the performance of the model will be greatly reduced. In a word, the performance of models will fluctuate greatly with different combinations of features and hybrid temporal models.

Multimodal Results: On the basis of bimodal features, we continue to fuse more related features (such as VGGFace, etc.) that have an average performance on baseline methods, based on early fusion strategy. Then, we obtained four groups of multimodal features and carried out experiments in six hybrid temporal models. From results shown in Table III and IV, we can conclude that when the multimodal feature is VGGish+FAU+BERT+VGGFace, the prediction results of both valence and arousal have achieved the best performance, but the model architecture is different in predicting the sentiment classes for valence or arousal. Additionally, we can see that more features do not bring performance improvement and even play a negative role. A possible explanation is that the fusion of various features based on early fusion strategy is not sufficient.

E. The results of the multitask learning

We select the two combinations of the best features and hybrid temporal models in each prediction task (arousal/valence) as the main task and related task for multitask learning. Our purpose is to make full use of the complementary information between multimodal data. According to the experiment results in Table V, we can infer that to a certain extent, conducting multitask learning does complement the feature of the current task by introducing the shared information from other tasks, which further improves the generalization ability of the model of the main task.

F. Submission Results

The best submission results are shown in Table VI. Although our method is superior to the baseline method with the valence of 0.3379 versus 0.3291, it lacks generalization ability in the prediction task of the sentiment classes for arousal. Fortunately, our method surpasses the baseline method with the combined arousal and valence of 0.3362 versus 0.3282 on the test set.

V. CONCLUSION

In this paper, we explore various features from three modalities (audio, video, and text) and carry out a lot of experiments in six hybrid temporal models. Additionally, for multimodal sentiment analysis, we also present a multitask learning framework that can adaptively change the loss weight per subtask. Firstly, we employ the baseline methods and PCA for obtaining the unimodal features with the best F1 score from each modality.
In addition, we find that the quality of the BERT feature is significantly improved after PCA, while other features after PCA will play a negative role. Secondly, we introduce six hybrid temporal models to capture the time-dependences in segments of videos. When the multimodal feature is VGGish + FAU + BERT + VGGFace, on the arousal and valence dimension, the best prediction results are both achieved. Finally, to make up for the defect that the early fusion strategy can not make full use of the complementarity between various multimodal features, we use the two best combinations of features and models as the main task and related task in our framework, which achieves information sharing between various multimodal features. Moreover, our method surpasses the baseline method with the combined arousal and valence of 0.3362 versus 0.3282 on the test set.


