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 CrossDisciplinary Research Grant (2020LKSFG04D).

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

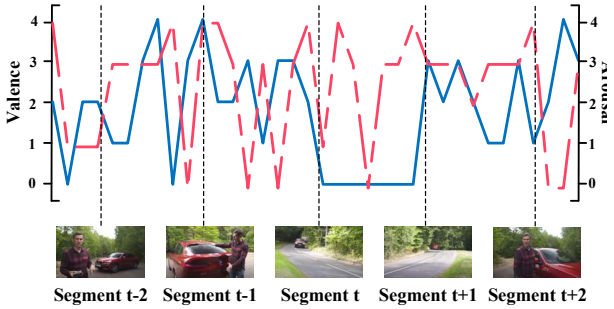

Fig. 1: Task illustration: predicting the five sentiment classes for arousal and valence every 250ms in the video. The solid blue line and red dotted line indicate valence and arousal.

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 dependences 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.

2375-9259/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDMW53433.2021.00025

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 sentient". 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 value 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:

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

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 θ 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

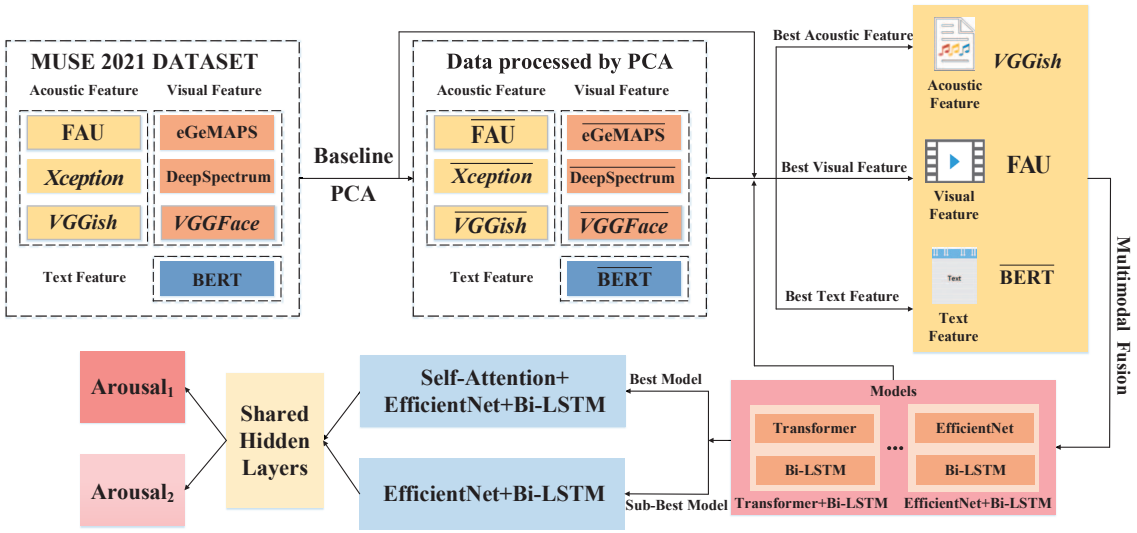

Fig. 2: Overview of the proposed model in the prediction of arousal. The "baseline" is the Self-Attention+Bidirectional Long Short-Term Memory (Bi-LSTM) model and the overline represents the unimodal features after PCA. The best and sub-best models are both selected from six hybrid models according to their performance in the validation set. Additionally, we adopt the early fusion strategy in multimodal fusion and concatenated the output of each subtask in shared hidden layers.

(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:

$$MH(Q,K,V) = concat(head_1, head_2, ..., head_n)W_h \quad (2)$$

$$head_i = Att(QW_i{}^Q, KW_i{}^K, VW_i{}^V) \quad (3)$$

$$Att(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

where $MH(.)$ and $Att(.)$ represent the multi-head attention layer and attention layer, respectively. w_h, W_i^Q, W_i^K and W_i^V denote the learnable parameter matrices. Q, K,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 precoding 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:

$$i_t = \sigma_s(W_i u_t + U_i \vec{h}_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma_s(W_f u_t + U_f \vec{h}_{t-1} + b_f) \quad (6)$$

$$o_t = \sigma_s(W_o u_t + U_o \vec{h}_{t-1} + b_o) \quad (7)$$

$$\vec{h}_t = \sigma_h(o_t \circ (f_t \circ \vec{h}_{t-1} + i_t \circ \sigma_h(W_c u_t + b_c))) \quad (8)$$

where, σ_s and σ_h are the sigmoid and tanh activation functions, respectively. $\circ$ is element-wise multiplication i, f and o are the input, forget and output gate vectors, respectively. W, U, and b are parameters of the model. $\overrightarrow{h_t}$ represents the hidden state of model. Similarly, we can obtain another hidden state $\overleftarrow{h_t}$. The concatenated result of the hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ is the final hidden state $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$.

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 sentient 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.

$$\mathcal{L}_{\mathcal{T}}\left(x, y_{\mathcal{T}}, \widehat{y}_{\mathcal{T}}, W_{\mathcal{T}}\right) = \sum_{\tau \in T} \mathcal{L}_\tau\left(x, y_\tau, \widehat{y}_\tau, W_\tau\right) \cdot \lambda_\tau \quad (9)$$

$$y'_\tau = \operatorname{softmax}\left(W_P h_\tau + b_P\right) \quad (10)$$

$$\widehat{y}_\tau = \underset{k}{\operatorname{argmax}}\left(y'_\tau[k]\right) \quad (11)$$

where, x represents the multimodal features. y_τ and $\widehat{y_\tau}$ are the true and prediction labels, respectively. $\mathcal{T}$ denotes the set of subtasks. $\tau \in \mathcal{T}$ is the current subtask and λ_τ is the corresponding loss weight. W_τ and $W_{\mathcal{T}}$ 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, λ_τ will greatly affect the final result. However, the manual adjustment process of parameter λ_τ 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.

$$\mathcal{L}_{\mathcal{T}}\left(x, y_{\mathcal{T}}, \widehat{y}_{\mathcal{T}}, W_{\mathcal{T}}\right) = \sum_{\tau \in T} \frac{1}{2\delta_\tau^2} \mathcal{L}_\tau\left(x, y_\tau, \widehat{y}_\tau, W_\tau\right) + \log\left(\delta_\tau\right) \quad (12)$$

where, δ_τ is a trainable parameter in each subtask, with an initial value of 1. To speed up the convergence of the models and avoid the over-fitting problem, we redefine the loss function in the following form:

$$\begin{aligned}\mathcal{L}_\tau\left(x, y_{\mathcal{T}}, \widehat{y}_{\mathcal{T}}, W_\tau\right) = \sum_{\tau \in \mathcal{T}} \frac{1}{2\delta_\tau^2} \mathcal{L}_\tau\left(x, y_\tau, \widehat{y}_\tau, W_\tau\right) \\ + \log\left(1 + \delta_\tau\right) + \eta\|\theta\|\end{aligned} \quad (13)$$

where, η is the L2 regularization term and θ 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.

Features	Modality	Dimension	PCA Features	Dimension
DeepSpectrum	A	4096	$\overline{DeepSpectrum}$	38
$VGGish$	A	128	$\overline{VGGish}$	75
eGeMAPS	A	88	$\overline{eGeMAPS}$	10
$Xception$	V	2048	$\overline{Xception}$	304
$VGGFace$	V	512	$\overline{VGGFace}$	34
FAU	V	35	$\overline{FAU}$	14
BERT	T	768	$\overline{BERT}$	412

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 proposed 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 II, we can find that the quality of the $\overline{BERT}$ feature is significantly improved, while the performance of the other two features is just the opposite. In particular, because the $\overline{BERT}$ feature shows a surprising effect, the subsequent multimodal fusion process will also replace the original BERT feature with the $\overline{BERT}$.

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+$\overline{BERT}$ 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+$\overline{BERT}$+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.

TABLE II: F1 score performance of unimodal features and bimodal features on the validation set. The Way1 to Way6 represent six hybrid temporal models respectively, and the best values under different features are highlighted in bold.

Features		Modalities	Arousal						Valence					
			Way1	Way2	Way3	Way4	Way5	Way6	Way1	Way2	Way3	Way4	Way5	Way6
Uni-modal feature	$VGGish$	A	17.42	31.35	26.28	34.67	**37.74**	21.62	18.08	26.83	17.94	31.72	**32.47**	17.29
	FAU	V	14.19	26.27	26.28	27.45	**36.30**	26.40	14.78	24.65	23.26	27.14	30.75	**28.01**
	$BERT$	T	18.29	**30.26**	26.72	26.41	26.40	26.40	16.97	**31.31**	20.78	24.66	23.36	24.87
	$\overline{VGGish}$	A	17.85	30.11	26.41	34.32	**36.14**	26.40	20.06	**31.59**	17.41	30.59	31.38	17.55
	$\overline{FAU}$	V	14.17	26.28	26.42	26.53	**29.72**	26.40	14.89	24.72	23.30	25.40	25.44	**26.55**
	$\overline{BERT}$	T	19.12	34.78	26.40	**35.03**	26.49	31.67	17.03	**32.38**	23.34	25.85	23.49	23.30
Bi-modal feature	$FAU + \overline{BERT}$	V+T	17.19	**34.78**	30.17	26.55	26.59	26.37	19.09	**31.96**	22.11	30.52	25.60	17.86
	$VGGish + FAU$	A+V	17.83	**31.98**	26.47	31.70	26.26	26.36	19.03	**32.23**	18.80	31.06	32.14	17.35
	$VGGish + \overline{BERT}$	A+T	19.59	23.30	26.21	**34.60**	26.53	26.07	17.39	**31.58**	18.26	24.75	27.20	17.60

TABLE III: F1 score performance of multimodal features on the arousal dimension on the validation set.

Features		Modalities	Arousal					
			Way1	Way2	Way3	Way4	Way5	Way6
Multimodal feature	$VGG_{ISH} + FAU + \overline{BERT}$	A+V+T	17.27	35.61	26.53	**36.39**	26.47	20.06
	$VGGish + FAU + \overline{BERT} + \overline{VGGFace}$	A+V+T	17.27	32.26	26.30	**37.21**	26.5	25.94
	$VGGish + FAU + \overline{BERT} + \overline{eGeMAPS}$	A+V+T	17.53	26.62	26.03	26.57	**37.05**	20.48
	$VGGish + FAU + \overline{BERT} + \overline{X_{CEPTION}} + \overline{VGGFace} + \overline{DeepSpectrum} + \overline{eGeMAPS}$	A+V+T	17.93	**36.97**	29.22	26.60	26.43	20.85

TABLE IV: F1 score performance of multimodal features on the valence dimension on the validation set.

Features		Modalities	Valence					
			Way1	Way2	Way3	Way4	Way5	Way6
Multimodal feature	$VGGish$+FAU+$\overline{BERT}$	A+V+T	17.43	**32.17**	19.45	31.24	23.41	18.14
	$VGGish + FAU + \overline{BERT} + \overline{VGGFace}$	A+V+T	19.13	**33.47**	23.70	29.44	25.54	18.38
	$VGGish + FAU + \overline{BERT} + \overline{eGeMAPS}$	A+V+T	18.79	31.57	20.70	**31.58**	25.79	17.58
	$VGGish + FAU + \overline{BERT} + \overline{Xception} + \overline{VGGFace} + \overline{DeepSpectrum} + \overline{eGeMAPS}$	A+V+T	18.93	**32.02**	21.94	27.20	23.69	17.66

TABLE V: F1 score performance of multitask learning model on the validation set.

	Main task		Related task		Best F1 score
	Feature	Model	Feature	Model	
Arousal	$VGGish$	Way5	$VGGish + FAU + \overline{BERT} + \overline{VGGFace}$	Way4	0.3807
Valence	$VGGish + FAU + \overline{BERT} + \overline{VGGFace}$	Way2	$VGGish$	Way5	0.3366

TABLE VI: The best submission results of our method on validation and test set.

Emotion	Partition	Baseline	Proposed
Arousal	Val	0.3772	0.3807
Valence	Val	0.3017	0.3366
Arousal	Test	0.3512	0.3345
Valence	Test	0.3291	0.3379

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 $VGGish + FAU + \overline{BERT} + VGGFace$, 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.

REFERENCES

[1] L. Stappen, L. Schumann, B. Sertolli, A. Baird, B. Weigel, E. Cambria, and B. W. Schuller, "Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," *arXiv preprint arXiv:2107.11757*, 2021.

[2] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[3] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognition Letters*, vol. 125, pp. 264–270, 2019.

[4] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 105–114.

[5] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *ACL*, 2019, pp. 527–536.

[6] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *arXiv preprint arXiv:2101.06053*, 2021.

[7] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3s, pp. 1–32, 2019.

[8] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, vol. 27, p. 46, 1997.

[9] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[10] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[11] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 597–600.

[12] R. Li, J. Zhao, J. Hu, S. Guo, and Q. Jin, "Multi-modal fusion for video sentiment analysis," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, 2020, pp. 19–25.

[13] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, 2020, pp. 27–34.

[14] E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, and J. L. Alba-Castro, "Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 31–40.

[15] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 562–566.

[16] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 65–72.

[17] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1114–1125.

[21] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1009–1021.

[22] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[23] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE transactions on affective computing*, 2019.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 27–33.

[28] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[29] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[34] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *arXiv preprint arXiv:2104.00298*, 2021.

[35] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.