Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism

Licai Sun*
School of Artificial Intelligence, University of Chinese Academy of Sciences
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
Beijing, China
sunlicai2019@ia.ac.cn

Zheng Lian*
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
lianzheng2016@ia.ac.cn

Jianhua Tao
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
CAS Center for Excellence in Brain Science and Intelligence Technology
Beijing, China
jhtao@nlpr.ia.ac.cn

Bin Liu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
liubin@nlpr.ia.ac.cn

Mingyue Niu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
mingyue.niu@nlpr.ia.ac.cn

ABSTRACT

Automatic perception and understanding of human emotion or sentiment has a wide range of applications and has attracted increasing attention nowadays. The Multimodal Sentiment Analysis in Real-life Media (MuSe) 2020 provides a testing bed for recognizing human emotion or sentiment from multiple modalities (audio, video, and text) in the wild scenario. In this paper, we present our solutions to the MuSe-Wild sub-challenge of MuSe 2020. The goal of this sub-challenge is to perform continuous emotion (arousal and valence) predictions on a car review database, Muse-CaR. To this end, we first extract both handcrafted features and deep representations from multiple modalities. Then, we utilize the Long Short-Term Memory (LSTM) recurrent neural network as well as the self-attention mechanism to model the complex temporal dependencies in the sequence. The Concordance Correlation Coefficient (CCC) loss is employed to guide the model to learn local variations and the global trend of emotion simultaneously. Finally, two fusion strategies, early fusion and late fusion, are adopted to further boost the model's performance by exploiting complementary information from different modalities. Our proposed method achieves CCC of 0.4726 and 0.5996 for arousal and valence respectively on the test set, which outperforms the baseline system with corresponding CCC of 0.2834 and 0.2431.

KEYWORDS

Dimensional Emotion Recognition; Long Short-Term Memory; Self Attention; Multi-modal Fusion

ACM Reference Format:
Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe'20), October 16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3423327.3423672

1 INTRODUCTION

Affective computing, as an emerging interdisciplinary field, aims to endow machines with the ability to recognize and understand the emotion or sentiment of human automatically [36]. With the advent of the era of big data, a huge amount of multimedia data is uploaded to online websites every day by people around the world. Automatic emotion recognition and analysis is of significant importance to manage and retrieve the large-scale data. Moreover, affective computing is a key step towards harmonious human-machine interaction, which has a wide range of applications, such as healthcare [33] and education [45].

*Equal contribution.

MuSe'20, October 16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8157-4/20/10…$15.00
https://doi.org/10.1145/3423327.3423672

There are two mainstream emotion representation models in affective computing, which are the categorical model and dimensional model [49], respectively. As for the categorical model, a person's emotional state is described by a few affective attributes, such as *happiness* and *sadness*. In contrast, in the dimensional model, the emotional state is mapped to a point in a Euclidean space. Thus, compared to the categorical model, the dimensional model can express subtler and more complicated emotional states. The Multimodal Sentiment Analysis in Real-life Media (MuSe) 2020 [40] adopts the dimensional model for emotion representation. In the MuSe-Wild sub-challenge of MuSe 2020, the task is to predict continuous dimensional values of arousal and valence from three common modalities (i.e., audio, video, and text). Before MuSe, the most influential challenge for continuous dimensional emotion recognition is the Audio-Visual Emotion Challenge (AVEC), which has been held successfully from 2011 to 2019. As the name suggested, AVEC pays more attention to audio and visual modalities, while MuSe aims to extensively explore the fusion of the three modalities.

Constructing a good multi-modal emotion recognition system lies in three aspects: discriminative feature extraction, powerful regression model, and effective multimodal fusion. Extracting discriminative features is an important step for robust emotion recognition. Traditionally, researchers rely on handcrafted features [7, 8, 37]. However, these features require careful engineering and considerable domain-specific expertise [28]. Recently, deep learning-based methods (i.e., convolutional neural network and recurrent neural network) have revolutionized the representation learning. Luo et al. [31] demonstrate that handcrafted features and deep representations describe the emotion information from different aspects. Therefore, in addition to handcrafted features, we also explore several efficient deep representations in this paper. Specifically, we extract deep acoustic representation from the VGGish model [20], which is trained on a large-scale audio dataset. For the visual modality, a ResNet-50 model [1] trained on a facial expression dataset is employed to obtain deep visual representation. For the textual modality, apart from word vectors like Word2Vec [32] or GloVe [35], we also extract contextual word embeddings using a pre-trained BERT model [13]. To the best of our knowledge, as a new and powerful language representation model, BERT has not been used in corresponding sub-challenges of previous AVECs.

After feature extraction, we need to choose a proper regression model for emotion prediction. Support Vector Regression (SVR), as a non-contextual model, is one of the most popular regression models. However, researchers suggest that contextual information is essential to emotion recognition and they propose various context-sensitive models to emphasize the temporal dynamic information. One of the state-of-the-art context-sensitive models is Long Short-Term Memory (LSTM) recurrent neural network [21], which has been successfully applied in previous AVECs [8–10, 12, 22, 23, 47, 48]. Unfortunately, LSTM suffers from the problem of gradient vanishing and its performance usually degrades when encountering very long sequence. Recently, Vaswani et al. [43] propose a new sequence transduction model with no recurrence, called Transformer, which has achieved impressive results in natural language processing. The key module of the Transformer is the self-attention mechanism, which allows for modeling temporal dependencies of different positions in the sequence without regard to their distance. Therefore, in this paper, we propose to augment LSTM with the self-attention mechanism to capture the complex temporal dynamics of emotions. We hypothesize that the integration of LSTM and self-attention mechanism can enhance the ability of long-term contextual modeling and is more suitable for continuous dimensional emotion recognition.

Multi-modal fusion aims to make more accurate predictions than its unimodal part by integrating shared and complementary information from different modalities. Typically, the fusion strategies can be split into three classes: early fusion (i.e., feature-level fusion), late fusion (i.e., decision-level fusion), and hybrid fusion [3]. Early fusion integrates features immediately after they are extracted from multiple modalities. The simplest way of early fusion is to concatenate multiple features and then feed them into the classification or regression model. One advantage of early fusion is that it can capture the interaction between different modalities. However, it needs to align multiple features due to the different sampling rates and may suffer from the problems caused by high dimensionality [12]. For late fusion, it allows for adopting a specific model for each modality and fuses the unimodal predictions by voting, weighting, or additional learned model. Thus, late fusion provides more flexibility for each unimodal model. However, it ignores the low-level interaction between different modalities. Hybrid fusion is the trade-off of early fusion and late fusion, which aims to exploit the advantages of both methods in a unified framework. In this paper, we adopt both early fusion and late fusion to integrate multi-modal features. We also compare the performance of these two fusion strategies.

In summary, our contributions are:

- We evaluate the effectiveness of various features including both handcrafted features and deep representations. Particularly, we extract contextual word representations from textual modality, which has not been utilized in corresponding sub-challenges of previous AVECs.
- We propose to augment LSTM with the self-attention mechanism for continuous dimensional emotion recognition. We show that the combination of these two modules can model long-term temporal dependencies in the sequence. Moreover, we fuse multi-modal features using both early fusion and late fusion and achieve promising results [1].

The remainder of this paper is organized as follows. We present the related works in Section 2. The multi-modal features and the recognition model are introduced in Section 3 and Section 4, respectively. The experimental results and analysis are described in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORKS

Multi-modal Features: As the predecessor of MuSe, various multi-modal features have been utilized in the past series of AVECs. In the early stage, participants usually use handcrafted features for different modalities. For instance, Sánchez-Lozano et al. [37], the winner of AVEC 2013, extract Local Binary Patterns (LBP) and Gabor feature from the visual modality and low-level descriptors such as Mel-frequency Cepstral Coefficients (MFCCs) from the audio modality. As deep learning shows its superior power over

[1] Code is available at https://github.com/youcaiSUN/MuSe-Wild_2020.

carefully-engineered features, participants tend to prefer the deep representations learned by various deep neural networks. In AVEC 2017, the experimental results of Chen et al. [12] and Huang et al. [23] demonstrate that deep visual representations extracted by deep convolutional neural networks achieve comparable or even better performance than handcrafted features (such as local Gabor binary patterns from three orthogonal planes and histogram of oriented gradients). Further, Zhao et al. [47] demonstrate that deep audio representation extracted from the VGGish model outperforms expert-knowledge based acoustic features in AVEC 2018. In the recent AVEC 2019, Chen et al. [9] verify the unparalleled performance of deep audio-visual representations once again by means of an efficient 2D+1D convolutional architecture. In addition to audio-visual modalities, textual modality also plays an important role in multimodal emotion recognition and sentiment analysis [36]. Text modality is first introduced in AVEC 2017. At first, competitors use the classic bag-of-words features [12, 23]. Afterward, word vectors like Word2Vec [32] and GloVe [35] trained on large-scale text corpora are widely adopted for their efficiency and effectiveness.

Model Architecture: SVR is often chosen as the baseline in previous AVECs. However, as a non-temporal model, SVR cannot utilize contextual information to facilitate emotion prediction. With the advent of deep learning, the Recurrent Neural Network (RNN) has shown its extraordinary superiority in sequence modeling. Unexceptionally, the winners of recent AVECs all adopt LSTM, a variant of RNN, for continuous dimensional emotion recognition. Wöllmer et al. [44] and Chen et al. [12] both utilize SVR and LSTM to perform regression analysis. Their studies reveal that LSTM, which captures the temporal information, outperform SVR significantly. In comparison with RNN based contextual regressors, Du et al. [14] propose a fully convolutional network, referred to as Temporal Hourglass Convolutional Neural Network (TH-CNN). TH-CNN can perform emotion prediction in a coarse-to-fine manner by integrating multi-scale features at different levels. Huang et al. [24] utilize three kinds of temporal models, including LSTM, Time-Delay Neural Network (TDNN) and multi-head attention model, to learn long-term context dependencies in the sequence. They show that the combination of these models obtains the best result.

Multi-modal Fusion: Multi-modal fusion is indispensable to achieve better performance for emotion recognition systems. In AVEC 2017, Huang et al. [23] adopt late fusion to combine the predictions of different features, while Chen et al. [12] utilize early fusion to better model the cross-modal dynamics. Further comparison of these two methods is made by Huang et al. [22] in AVEC 2018. The results show that late fusion is good at predicting arousal and valence, while early fusion is more suitable for liking prediction. In the recent AVEC 2019, Chen et al. [9] propose to combine early fusion and late fusion. The authors first train Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks (DBLSTM) for each unimodal feature. Then, several bi-modal features are early fused using other DBLSTMs. Finally, a second level DBLSTM is adopted to late fuse the predictions of unimodal models and early fusion models. Recently, Huang et al. [25] propose to fuse audio-visual modalities via inter-modal and intra-modal attention. The results show that the proposed method achieved better performance than early fusion and late fusion.

3 MULTIMODAL FEATURES

3.1 Acoustic Features

eGeMAPS Feature: The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) contains 23 acoustic low-level descriptors (LLDs), covering spectral, cepstral, and prosodic features [15]. Several statistical functions can be applied over these LLDs to extract segment-level feature, which results in an 88-dimensional vector. We use the eGeMAPS feature provided by the organizers of MuSe 2020, which utilize the freely available openSMILE toolkit [16] to extract it. It's noted that the window size is set to 5 s and a hop size of 0.25 s is applied in order to match with the ground-truth emotion labels.

PyAudio Feature: In addition to the segment-level feature, we also extract short-time frame-level acoustic features. PyAudio feature is a 34-dimensional vector, which includes MFCCs, zero-crossing rate, spectral spread, and chroma vector etc. We use an open-source python library [18] to extract PyAudio feature. The window size and hop size are set to be 0.025 s and 0.01 s, respectively. To align with the ground-truth emotion labels, the frame-level features near each label timestamp are averaged.

IS13 Feature: To reflect a border coverage of paralinguistic information, we extract another frame-level acoustic feature with IS13 configuration using the openSMILE toolkit [16]. The IS13 feature set is a comprehensive acoustic feature set, which is first introduced in the INTERSPEECH 2013 Computational Paralinguistics Challenge [38]. Since then, it has been widely used by the emotion recognition community. It consists of 65 LLDs and the corresponding 1st derivatives. To align with the ground-truth emotion labels, the frame-level features near each label timestamp are averaged.

VGGish Feature: VGGish [20] is a variant of VGGNet [39], which is designed for audio classification. It is trained on a large-scale audio dataset, AudioSet [17], which contains over 2 million human-labeled video soundtracks with more than 600 audio event classes. Zhao et al. [47] show that the feature extracted from VGGish outperforms handcrafted acoustic features. Therefore, in this paper, we utilize a pre-trained VGGish model to extract deep acoustic representation. To match with the ground-truth emotion labels, the recordings are first divided into multiple 0.975 s frames with a hop size of 0.25 s. Then, log spectrograms are extracted from these frames and are fed into the VGGish model. Finally, we extract the high-level 128-dimensional embeddings from the output of $fc2$ layer as the VGGish feature.

3.2 Visual Features

FAUs Feature: OpenFace toolkit [4] provides a wide range of facial features, such as facial landmarks (2D or 3D), head pose features, eye gaze positions and the intensity and presence of 17 Facial Action Units (FAUs). We only use the FAUs feature provided by the organizers.

OpenPose Feature: To exploit the pose information of speaker in the video, the pose feature is extracted using a pre-trained OpenPose model [6]. OpenPose feature includes 2D coordinates and confidence score of a keypoint being present for each of 18 2D pose keypoints. We use the OpenPose feature provided by the organizers.

ResNetFace Feature: We employ a pre-trained ResNet-50 model [1] to extract deep visual representation. Specifically, the model is

first pre-trained on a large-scale face recognition dataset. Then, it's fine-tuned on a facial expression dataset, FER+ [5]. The accuracy of the fine-tuned model on the test set of FER+ is 87.4%. We apply spatial average pooling to the bottleneck layer of *conv*5_3, resulting in a 512-dimensional feature vector.

3.3 Textual Features

Global Word Vector: Word vector-based model represents each word with a real-value vector in the semantic vector space [35]. Compared to other representations, such as the one-hot vector, the word vector is more compact and can capture linguistic regularities in language [32]. Therefore, word vectors are widely used in modern natural language processing systems. We adopt two kinds of pre-trained word vector models, which are Word2Vec [2] and GloVe [3], to get the representation of each word in the transcription.

Contextual Word Embedding: However, the above mentioned word vectors are global (or static), i.e. they are fixed after training. In contrast, contextual word embeddings assign each word a dynamic representation based on its context [30], thus providing more flexibility. Recently, contextual word embeddings from Transformer-based architectures, such as BERT [13], have shown state-of-the-art performance on many downstream tasks in natural language processing. Thus, we employ a pre-trained BERT model to derive contextual word embedding. Specifically, an uncased base BERT model [4] is adopted. Then, the output of the last layer, the sum of outputs of the last two layers and the last four layers are extracted as contextual features, which are referred to as "BERT", "BERT-2" and "BERT-4", respectively.

4 EMOTION RECOGNITION MODELS

In this section, we introduce the emotion recognition model in detail. Our model consists of three modules: LSTM, self-attention mechanism and the regression layer. As the state-of-the-art temporal model, LSTM is employed as the main module for long-term contextual modeling. However, its ability is limited when encountering very long sequences, which is common in continuous dimensional emotion recognition. The self-attention mechanism can relate different positions in the sequence without regard to their distance by means of position-pair computation [24, 43]. Therefore, we propose to augment LSTM with self-attention mechanism. On the one hand, we expect that the self-attention mechanism could endow LSTM with the ability to capture longer temporal dependencies. On the other hand, since self-attention could not make use of the order information in the sequence [43], we believe that LSTM can implicitly guide it to learn that information.

Overall, as shown in Fig.1, the input features are first encoded by the self-attention mechanism (denoted by the dotted lines). Then, LSTM transforms the encoded sequence into context-dependent hidden states. Finally, a regression layer maps them to emotion predictions. In the following parts, we elaborate on the main modules in our model and introduce the loss function as well as fusion strategies in the end.

[2]https://code.google.com/archive/p/word2vec
[3]https://nlp.stanford.edu/projects/glove
[4]https://huggingface.co/transformers/pre-trained_models.html

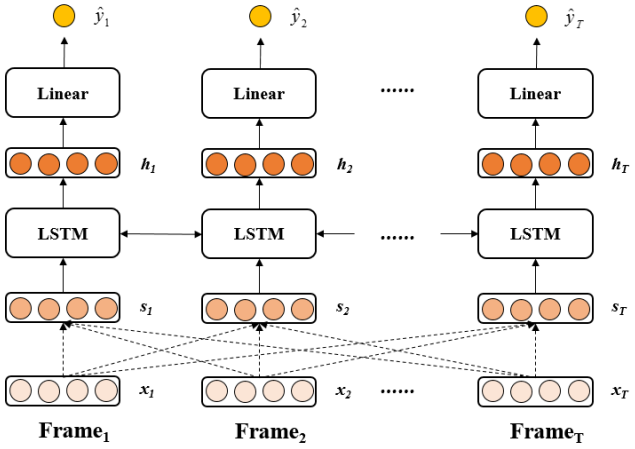

Figure 1: Overview of the proposed model. x refers to uni-modal or multi-modal features. s refers to the outputs of the self-attention layer. h refers to context-dependent hidden states output by the LSTM layer and $\hat{y}$ refers to final emotion predictions. Dotted lines denote the self-attention mechanism.

4.1 Self-attention mechanism

The self-attention mechanism utilizes multi-head scaled dot-product attention to transform the low-level input sequence into high-level and more abstract representations. In Fig.1, the dotted lines indicate the self-attention mechanism. Assume the input to self-attention layer is the feature sequence $X = \{x_1, x_2, ..., x_T\} \in \mathbb{R}^{T \times d}$, where $x_i \in \mathbb{R}^d$ is the frame-level feature at time step t and T is the max time step. To perform multi-head scaled dot-product attention on the input sequence X, we need to generate corresponding queries Q, keys K and values V. To this end, we project X for h times using different linear projection layers, which are computed as follows:

$$Q_i = XW_i^Q \tag{1}$$

$$K_i = XW_i^K \tag{2}$$

$$V_i = XW_i^V \tag{3}$$

where $Q_i \in \mathbb{R}^{d \times (d/h)}$, $K_i \in \mathbb{R}^{d \times (d/h)}$, $V_i \in \mathbb{R}^{d \times (d/h)}$, $i = 1, 2, ..., h$ and h is the number of heads.

For each head's query Q_i, key K_i and value V_i, we perform the scaled dot-product attention with the following equation:

$$Head_i = Softmax(Q_i K_i^T / \sqrt{d_k}) V_i \tag{4}$$

where $head_i \in \mathbb{R}^{T \times (d/h)}$, $d_k = d/h$ is the scale factor. Then, results of each head are concatenated together and linearly projected to obtain

$$R = Concat(Head_1, ..., Head_h) W_O \tag{5}$$

where $W_O \in \mathbb{R}^{d \times d}$ is the projection matrix. Following [43], we add a residual connection [19] and layer normalization [2] to get the final encoded sequence S.

$$S = LayerNorm(X + R) \tag{6}$$

4.2 LSTM

LSTM is one of the most famous variants of RNN. Compared with vanilla RNN, LSTM employs a memory cell to store information and additional gates (including an input gate, an output gate and a forget gate) to control the information flow, which alleviates the notorious problem of gradient vanishing in model training.

To obtain context-dependent representations $H = \{h_1, h_2, ..., h_T\}$, we employ LSTM to compute the hidden vector sequence H from S with the following equations:

$$f_t = \sigma_g(W_f s_t + U_f c_{t-1} + b_f) \tag{7}$$

$$i_t = \sigma_g(W_i s_t + U_i c_{t-1} + b_i) \tag{8}$$

$$o_t = \sigma_g(W_o s_t + U_o c_{t-1} + b_o) \tag{9}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_h(W_c s_t + b_c) \tag{10}$$

$$h_t = \sigma_h(o_t \circ c_t) \tag{11}$$

where σ_g is the sigmoid function and σ_h is the hyperbolic tangent function. $\circ$ is element-wise multiplication. f, i and o are the forget gate vectors, input gate vectors and output gate vectors, respectively. W, U and b are weight matrices and bias vectors for each gate.

Then, a fully connected layer following the LSTM layer makes the final emotion prediction $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$.

4.3 CCC Loss

In comparison to previous works which use either mean squared error [9, 48] or ϵ-insensitive mean absolute error [22] as the loss function, we utilize the Concordance Correlation Coefficient (CCC) [27] loss, which is introduced in [41], to train our model. As CCC is chosen as the evaluation metric in MuSe-Wild sub-challenge, we believe that optimizing CCC loss directly can improve the model's performance. The CCC loss is calculated as follows:

$$\mathcal{L} = 1 - CCC \tag{12}$$

$$CCC = \frac{2\rho\sigma_{\hat{Y}}\sigma_Y}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 + (\mu_{\hat{Y}} - \mu_Y)^2} \tag{13}$$

where $\mu_{\hat{Y}}$ and μ_Y are the mean of the prediction $\hat{Y}$ and the label Y, respectively. $\sigma_{\hat{Y}}$ and σ_Y are the corresponding standard deviations. ρ is the Pearson Correlation Coefficient (PCC) between $\hat{Y}$ and Y.

4.4 Fusion Strategies

We adopt both early fusion and late fusion for multi-modal emotion recognition in this paper. For early fusion, we simply concatenate multi-modal features and feed them into the model. For late fusion, we employ a second-level LSTM model to fuse the predictions from several unimodal features.

5 EXPERIMENTS

5.1 Dataset

The dataset used in MuSe 2020 is MuSe-CaR, which is a large multi-modal (audio, video, and text) dataset [40]. It consists of 36h:52m:08s of video data from 291 videos and 70 host speakers collected from YouTube. The topic of these videos is limited to reviews of cars with premium brands (such as BMW, Audi, and Mercedes-Benz). In contrast to the datasets used in the previous AVECs, there are several "in-the-wild" characteristics in MuSe-CaR. For example, 1) different shot size, dynamic face angles, and highly varying backgrounds in the video; 2) ambient noises in the audio; 3) usage of colloquialisms and domain-specific terms in the text.

For the MuSe-Wild sub-challenge, a total 35h:08m:01s of video data is annotated per 0.25 s on two emotion dimensions (i.e. arousal and valence). The numbers of videos in the training, validation, and test sets are 165, 62, and 64, respectively. The evaluation metric for this sub-challenge is CCC, which is defined in equation (13).

5.2 Experimental Setup

Data Preprocessing: We notice that there are too many frames in a video and the longest video has more than 6000 frames. As suggested in Huang et al. [22], cutting the video into multiple segments not only enriches the training samples but also contributes to model convergence during training. Thus, we segment each video in the training set with a window size of 200 frames (50s) and a hop size of 100 frames (25s). Besides, we apply standardization to handcrafted acoustic features (i.e., eGeMAPS, PyAudio and IS13 feature) and OpenPose feature before feeding them into the model. We empirically find that the meta feature *segment_id* in the feature files provided by the organizers can improve the performance on the validation set. Therefore, we also add it to the input features.

Model Training: We implement our models within the PyTorch framework [34]. Specifically, for unimodal and early fusion model, the model consists of a self-attention layer, a bidirectional LSTM layer and a fully connected layer. The number of hidden neurons in the model is 64, 128, or 256, which depends on the size of input features. The number of heads in the self-attention layer is 4 or 8. To train the model, we use Adam optimizer [26] with varied learning rate (0.002, 0.003, or 0.005) and batch size (256, 512, or 1024), which is also dependent on the size of the input features. Once the training loss does not decrease in 5 consecutive epochs, we halve the learning rate. The maximum number of epochs for model training is 100. Other hyper-parameters, such as dropout rate, are chosen based on the model's performance on the validation set. For the late fusion model, we employ a bidirectional LSTM layer with 32 cells to fuse the predictions from several unimodal features. We train the late fusion model at most 20 epochs using Adam optimizer with a learning rate of 0.001 and a batch size of 64.

5.3 Ablation Studies

We first conduct several experiments to verify the effectiveness of our model and loss function. The results of different models on the validation set are shown in Table 1. For a fair comparison, we use two bidirectional LSTM layers for the "LSTM" model and two self-attention layers for the "Self-Attn" model. We observe that: 1) the performance of the "Self-Attn" model usually is worst due to the loss of position information; 2) generally, the "LSTM" model performs better than the "Self-Attn" model; 3) the proposed model ("LSTM+Self-Attn") achieves the best performance in most cases except when predicting valence using the "BERT" feature. These results verify that the combination of LSTM and self-attention mechanism can capture longer temporal dependencies in the sequence and is more suitable for continuous dimensional emotion recognition. Therefore, the "LSTM+Self-Attn" model is utilized in

Table 1: CCC performance of different models on the validation set.

Feature	Model	Arousal	Valence
VGGish	Self-Attn	0.3357	0.1371
VGGish	LSTM	0.4590	0.1228
VGGish	LSTM+Self-Attn	**0.4996**	**0.1423**
ResNetFace	Self-Attn	0.3341	0.0775
ResNetFace	LSTM	0.3832	0.0896
ResNetFace	LSTM+Self-Attn	**0.4157**	**0.0999**
BERT	Self-Attn	0.2796	0.1225
BERT	LSTM	0.3570	**0.4385**
BERT	LSTM+Self-Attn	**0.3984**	0.4375

Table 2: CCC performance of different loss functions on the validation set.

Feature	Loss	Arousal	Valence
VGGish	MSE	0.2485	0.0953
VGGish	L1	0.2965	0.0998
VGGish	CCC	**0.4996**	**0.1423**
ResNetFace	MSE	0.2847	0.0702
ResNetFace	L1	0.2251	0.0572
ResNetFace	CCC	**0.4157**	**0.0999**
BERT	MSE	0.2160	0.3107
BERT	L1	0.2147	0.3593
BERT	CCC	**0.3984**	**0.4375**

the following experiments. The results of different loss functions on the validation set are shown in Table 2. We can notice that "CCC" loss outperforms "MSE" and "L1" loss. We suggest that the reason is that either "MSE" or "L1" loss is sensitive to local error and can't see the big picture when performing long-term emotion predictions. However, "CCC" loss is dependent on both local error and the correlation between prediction and label, which can force the model to learn the global trend and local variations simultaneously.

5.4 Unimodal Results

In this part, we evaluate the performance of handcrafted features and deep representations extracted from each modality. The results are shown in Table 3. When comparing the results within each modality, we observe that deep representations achieve better performance than handcrafted features in general. Besides, as expected, contextual word representations (i.e., "BERT" and its variants) perform better than the two global word vectors (i.e., "Word2Vec" and "GloVe"). When comparing the results from different modalities, we can find that: 1) on the arousal dimension, features extracted from the audio modality usually are more effective than those extracted from visual and textual modalities. 2) on the valence dimension, textual modality performs much better than the other two modalities. These findings are consistent with the results in the baseline paper [40]. We suggest that the reason behind it is that the perception of arousal is mainly dependent on how people speak while valence is mostly reflected by the speech content. However, it's noted that the performance of visual modality is the worst in general, which

Table 3: CCC performance of unimodal features on the validation set. "A", "V", and "T" denote audio, visual, and textual modality, respectively.

Feature	Modality	Dimension	Arousal	Valence
eGeMAPS	A	88	0.3903	0.1179
PyAudio	A	34	0.4150	0.1721
IS13	A	130	0.4248	0.1169
VGGish	A	128	**0.4996**	0.1423
FAUs	V	35	0.3718	0.1264
OpenPose	V	54	0.4101	0.0825
ResNetFace	V	512	0.4157	0.0999
GloVe	T	300	0.3676	0.3685
Word2Vec	T	300	0.3756	0.3486
BERT	T	768	0.3984	0.4375
BERT-2	T	768	0.3610	0.4443
BERT-4	T	768	0.3438	**0.4469**

disaccords with the findings in previous AVECs [9, 47]. We believe that the different datasets used in two challenges might lead to this phenomenon. As stated above, the MuSe-CaR dataset in this challenge has several "in the wild" characteristics, especially for visual modality. For example, the videos have highly varying backgrounds and the faces of car reviewers in the videos are often not frontal. What's more, in several videos, there are even no faces. Therefore, we think that these factors make predicting emotions from the visual modality more challenging. Finally, the best unimodal results for arousal and valence are 0.4996 and 0.4469 respectively, which outperform both unimodal (0.3078 and 0.1273) and multi-modal (0.2587 and 0.1506) baseline results [40].

5.5 Multi-modal Results

Two strategies (early fusion and late fusion) for multi-modal fusion are explored in this subsection. Unlike the previous study [23] which uses the greedy method to perform feature selection, we simply select several top-performing features from each modality to investigate the effectiveness of multi-modal fusion. The results of these two strategies on the arousal dimension on the validation set are shown in Table 4. We can observe that: 1) multi-modal fusion can boost the model's performance significantly; 2) tri-modal fusion performs better than bi-modal fusion; 3) too many unimodal features involved in the fusion process may hurt the performance; 4) generally, early fusion and late fusion achieve comparable results. For valence, the results of two fusion strategies are shown in Table 5. Observations 2) and 3) of arousal also hold. Different from arousal, the improvement of performance on the valence dimension is limited. We believe that contextual word representations play a major role in valence prediction and no more complementary information from other modalities can be exploited. Besides, late fusion achieves consistently better performance than early fusion. Finally, the best multi-modal results for arousal and valence are 0.5616 and 0.4704, respectively, which outperform both the best unimodal results (0.4996 and 0.4469) and the corresponding baseline results (0.3078 and 0.1506) [40].

Table 4: CCC performance of multi-modal features on the arousal dimension on the validation set. "A", "V", and "T" denote audio, visual, and textual modality, respectively.

Features	Modalities	Early Fusion	Late Fusion
IS13+ResNetFace	A+V	0.4844	0.4898
ResNetFace+BERT	V+T	0.4716	0.4605
IS13+BERT	A+T	0.4554	0.4931
IS13+ResNetFace+BERT	A+V+T	0.5224	0.5113
IS13+VGGish+ResNetFace+BERT	A+V+T	**0.5480**	**0.5616**
IS13+VGGish+OpenPose+ResNetFace+GloVe+BERT	A+V+T	0.5336	0.5372

Table 5: CCC performance of multi-modal features on the valence dimension on the validation set. "A", "V", and "T" denote audio, visual, and textual modality, respectively.

Features	Modalities	Early Fusion	Late Fusion
PyAudio+FAUs	A+V	0.1494	0.1930
FAUs+BERT-4	V+T	0.4408	0.4567
PyAudio+BERT-4	A+T	0.4515	0.4633
PyAudio+FAUs+BERT-4	A+V+T	0.4590	0.4605
PyAudio+FAUs+GloVe+BERT-4	A+V+T	**0.4635**	**0.4704**
PyAudio+VGGish+FAUs+ResNetFace+GloVe+BERT-4	A+V+T	0.4609	0.4676

Table 6: The best submission results of the proposed method on validation set and test set.

Emotion	Partition	Baseline	Proposed
Arousal	Val	0.3078	**0.5616**
Valence	Val	0.1506	**0.4876**
Arousal	Test	0.2834	**0.4726**
Valence	Test	0.2431	**0.5996**

5.6 Submission Results

The best submission results are shown in Table 6. Our proposed method outperforms the baseline system with the arousal of 0.4726 versus 0.2834 and valence of 0.5996 versus 0.2431. It's noted that considering the performance gap between the validation set and test set on the arousal dimension, our proposed method might overfit the validation set. The interesting thing is that, on the valence dimension, the best result of our proposed method on the validation set is 0.4876, while it achieves 0.5996 on the test set. We conjecture that the distribution of two sets might be different.

6 CONCLUSIONS

In this paper, we present our contributions to the MuSe-Wild sub-challenge of MuSe 2020. Various handcrafted features and deep representations from three common modalities (i.e., audio, video, and text) are explored. To capture temporal dependencies, LSTM is adopted as the main module of our emotion recognition model. To further enhance LSTM's ability of long-term contextual modeling, we propose to augment LSTM with the self-attention mechanism. The CCC loss is utilized to guide the model to capture both local variations and the global trend of emotion. Moreover, both early fusion and late fusion are adopted to boost the model's performance. Experimental results show that our proposed model outperforms the baseline system by a large margin.

There are several limitations in this work. First, we find that the label timestamps are not evenly spaced due to the exclusion of irrelevant video segments [40]. However, we don't consider this during model training and inference. Second, we simply concatenate several unimodal features and predictions to perform early fusion and late fusion, respectively. More advanced fusion methods such as attention-based fusion [11, 29] or tensor fusion [46] can be explored. Besides, Transformer-like [42] architectures can be employed to model the complex temporal dynamics of emotion in the future.

ACKNOWLEDGMENTS

This work is supported by the National Key Research & Development Plan of China (No. 2017YFB1002804), the National Natural Science Foundation of China (NSFC) (No. 61831022, No. 61771472, No. 61773379, No. 61901473) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).

REFERENCES

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*. 292–301.
[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.
[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 279–283.
[6] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[7] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2014. Multi-scale temporal modeling for dimensional emotion recognition in video. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 11–18.

[8] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 65–72.

[9] Haifeng Chen, Yifan Deng, Shiwen Cheng, Yixuan Wang, Dongmei Jiang, and Hichem Sahli. 2019. Efficient spatial temporal convolutional features for audio-visual continuous affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 19–26.

[10] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 508–513.

[11] Shizhe Chen and Qin Jin. 2016. Multi-modal conditional attention fusion for dimensional emotion prediction. In *Proceedings of the 24th ACM international conference on Multimedia*. 571–575.

[12] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 19–26.

[13] 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).

[14] Zhengyin Du, Suowei Wu, Di Huang, Weixin Li, and Yunhong Wang. 2019. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *IEEE Transactions on Affective Computing* (2019).

[15] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202.

[16] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[18] Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one* 10, 12 (2015), e0144610.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[22] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. 2018. Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 57–64.

[23] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. 2017. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 11–18.

[24] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu. 2019. Efficient Modeling of Long Temporal Contexts for Continuous Emotion Recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 185–191.

[25] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. 2020. Multimodal Transformer Fusion for Continuous Emotion Recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3507–3511.

[26] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[27] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[29] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. 2019. Conversational Emotion Analysis via Attention Mechanisms. *Proc. Interspeech 2019* (2019), 1936–1940.

[30] Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A Survey on Contextual Embeddings. *arXiv preprint arXiv:2003.07278* (2020).

[31] Danqing Luo, Yuexian Zou, and Dongyan Huang. 2018. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. *Proc. Interspeech 2018* (2018), 152–156.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[33] Mingyue Niu, Jianhua Tao, Bin Liu, and Cunhang Fan. 2019. Automatic Depression Level Detection via Lp-Norm Pooling. In *Proc. Interspeech 2019*. 4559–4563.

[34] 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. In *Advances in neural information processing systems*. 8026–8037.

[35] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[36] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[37] Enrique Sánchez-Lozano, Paula Lopez-Otero, Laura Docio-Fernandez, Enrique Argones-Rúa, and José Luis Alba-Castro. 2013. 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 (AVEC '13)*. New York, NY, USA, 31–40.

[38] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

[39] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[40] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media. In *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, co-located with the 28th ACM International Conference on Multimedia (ACM MM)*. ACM.

[41] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5200–5204.

[42] 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 57th Annual Meeting of the Association for Computational Linguistics*. 6558–6569.

[43] 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.

[44] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31, 2 (2013), 153–163.

[45] Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. 2019. Affective computing in education: A systematic review and future research. *Computers & Education* 142 (2019), 103649.

[46] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.

[47] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. 2018. Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 65–72.

[48] Jinming Zhao, Ruichen Li, Jingjun Liang, Shizhe Chen, and Qin Jin. 2019. Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 37–45.

[49] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. 2019. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–32.