Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition

Shizhe Chen
Renmin University of China
cszhe1@ruc.edu.cn

Qin Jin*
Renmin University of China
qjin@ruc.edu.cn

Jinming Zhao
Renmin University of China
zhaojinming@bjfu.edu.cn

Shuai Wang
Renmin University of China
shuaiwang@ruc.edu.cn

ABSTRACT

Automatic emotion recognition is a challenging task which can make great impact on improving natural human computer interactions. In this paper, we present our effort for the Affect Subtask in the Audio/Visual Emotion Challenge (AVEC) 2017, which requires participants to perform continuous emotion prediction on three affective dimensions: Arousal, Valence and Likability based on the audiovisual signals. We highlight three aspects of our solutions: 1) we explore and fuse different hand-crafted and deep learned features from all available modalities including acoustic, visual, and textual modalities, and we further consider the interlocutor influence for the acoustic features; 2) we compare the effectiveness of non-temporal model SVR and temporal model LSTM-RNN and show that the LSTM-RNN can not only alleviate the feature engineering efforts such as construction of contextual features and feature delay, but also improve the recognition performance significantly; 3) we apply multi-task learning strategy for collaborative prediction of multiple emotion dimensions with shared representations according to the fact that different emotion dimensions are correlated with each other. Our solutions achieve the CCC of 0.675, 0.756 and 0.509 on arousal, valence, and likability respectively on the challenge testing set, which outperforms the baseline system with corresponding CCC of 0.375, 0.466, and 0.246 on arousal, valence, and likability.

KEYWORDS

Dimensional Emotion; Multimodal Features; LSTM; Multi-task Learning

1 INTRODUCTION

Automatic emotion recognition is a crucial component to improve natural human-computer interactions. It has a wide range of applications ranging from computer tutoring to mental health diagnoses.

Dimensional emotion theory is one of the most popular computing models for emotion recognition [1]. It considers an emotion state as a point in a continuous space described by the dimensions of arousal (a measure of affective activation), valence (a measure of pleasure) and dominance (a measure of control). Therefore, dimensional theory can model subtle, complicated, and continuous affective behaviors.

The Audio-Visual Emotion Challenge (AVEC 2017) [2] Emotion Subtask provides an audiovisual dataset captured in the real-life conditions to compare different methods for the dimensional emotion recognition. Previous research works [3–8] for this task have explored different multimodal features, fusion methods, and regression models.

Our contributions to the challenge in this paper are from three aspects:

First, we investigate different deep learned features from acoustic, visual and textual modalities. For the acoustic features, besides the expert-knowledge based features, we employ the Soundnet convolutional neural networks (CNN) pretrained on unsupervised video data to extract low-level acoustic features. Since there are speeches from the two participants in the conversation audio, we adopt three different strategies to construct acoustic features in order to explore the influence of interlocutor for emotion recognition. For the visual features, we compare the performance of facial features extracted from two kinds of deep CNNs, which are pretrained on the other facial expression dataset. And for the textual features, we utilize the word embedding features on the speech transcriptions in original language German and translated language English in addition to the traditional bag-of-words textual representation. Our results show that the fusion of different modality features can benefit the arousal and valence prediction, but only the textual features perform and generalize well for the likability prediction.

Second, we compare the recognition performance between the non-temporal model SVR and the temporal model LSTM. For arousal and valence emotion dimensions, the temporal LSTM model significantly outperforms the non-temporal model, which benefits from its ability to capture long-range

[*] Qin Jin is the corresponding author.

AVEC'17, October 23, 2017, Mountain View, CA, USA
© 2017 ACM. ISBN 978-1-4503-5502-5/17/10. . . $15.00
DOI: https://doi.org/10.1145/3133944.3133949

dependencies with the inputs of detailed short-time features. But the performance gap for the likability prediction is small for the temporal and non-temporal models, because the likability is mainly related to the semantic meanings of the speech content which is not continuous in time.

Finally, since the arousal and valence are highly correlated with each other, we propose to learn the arousal and valence prediction simultaneously in the multi-task learning framework. For finetuning the deep learned facial features, the multi-task learning improves the performances of the finetuned features compared with those finetuned by single task. For training the temporal LSTM-RNN model, our experiments suggest that the multi-task learning is beneficial when the capacity of the model is increased over that in the single task learning. Our approaches significantly improve the baseline [2] on the challenge testing set, which achieve CCC of 0.675 on arousal, 0.756 on valence and 0.509 on likability.

The paper is organized as follows. Section 2 introduces the related works. Section 3 and Section 4 describe our proposed solutions in multimodal features and recognition models respectively. Experimental results and analysis are presented in Section 5. Finally, Section 6 draws some conclusions.

2 RELATED WORKS

Multimodal Features and Fusions: Previous works in the series of the AVEC challenge have explored a variety of multimodal features. Brady et al. [5], the winner of the AVEC2016 challenge, derive high-level acoustic, visual and physiological features from the low-level descriptors using sparse coding and deep learning. Povolny et al. [6] focus on features extracted from audio including bottleneck acoustic features and text-based features. There are mainly three strategies to fuse the different modalities, namely early-fusion, late fusion and model-level fusion [9]. Early fusion concatenate multimodal features as the input for the prediction models, which is easy and can improve performance successfully in the previous work [10]. But the early fusion method suffers from the high dimensionality of the features and asynchronization of different features. Late fusion [11] combines the predictions from the different modalities by weighted sum or a second level model, but ignores the interactions between different features. As a compromise for early and late fusion, model-level fusion is proposed which fuses the intermediate representations of the features and dependent on the specific models such as neural networks, HMMs and kernel models [12]. Recently, Chen et al. [13] propose a temporal fusion model to dynamically pay attention to relevant modality features through time, which shows improvements over the traditional fusion strategy.

Emotion Recognition Models: Models for the dimensional emotion recognition can be divided into two categories. The first is non-temporal models which usually are fed with contextual features. Huang et al. [11] investigate the RML model with the annotation delay compensation and output-associative fusion. Brady et al. [5] use the linear support vector machine (SVM) to perform the regression task. The second place winner [3] in AVEC2014 challenge use Deep Belief Network (DBN) with temporal pooling and multimodal-temporal fusion, which demonstrates that high level temporal fusion can improve performance. In recent years, more and more researchers started to focus on the second type of models, temporal models, which can emphasizes the temporal dynamic information directly in the model. Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) [14] is one of the state-of-art sequence modeling techniques, and has also been successfully applied in dimensional emotion recognition. In AVEC2015 challenge, three out of the four top systems [4, 7, 8] utilize the LSTM model to predict dimensional emotions, which explore different structures of the LSTM and loss functions for the task.

Multi-task Learning: Xia et al. [15] treat the categorical emotion recognition and the dimensional emotion recognition as the multi-task learning based on the Deep Belief Network (DBN). Ringeval et al. [16] explore two types of multi-task learning for dimension emotion recognition. The first is to learn the predictions of different annotators and the other is to learn different emotion dimensions simultaneously. The multi-task learning framework improves the visual feature based systems but suffers in the audio feature based system. Chang et al. [17] use a deep convolutional generative adversarial network to learn features with unlabelled data and then apply multi-tasking learning to predict arousal and valence emotions. Besides using the target highly-correlated tasks as the multi-task, Garnin et al. [18] introduce the domain-adversarial and Saon et al. [19] propose the speaker-adversarial neural network to train a domain or speaker classifier respectively in parallel with the main task speech recognition, which shows improvements for both tasks.

3 MULTIMODAL FEATURES

In this section, we introduce the features from acoustic, visual and textual modalities respectively. As suggested in previous work [7], non-temporal models require long-time features to capture contextual information, while the temporal models prefer the short-time features since the model can capture the contexts and short-time features can reveal more details. So for each modality, we extract two groups of feature: long-time features and short-time features.

3.1 Acoustic Features

We use the challenge provided feature [2], the bag-of-audio-words (`BoAW`) feature with window length of 6s and shift of 100ms, as the long-time acoustic feature. We mainly describe the short-time acoustic features as below.

IS10 Feature: We utilize the OpenSMILE toolkit [20] to extract the expert-knowledge based acoustic features with `IS10` configuration [21], which includes 76 low-level descriptors such as MFCCs, loudness, F0, jitter and shimmer. The window and shift length are set to be 100ms to match with

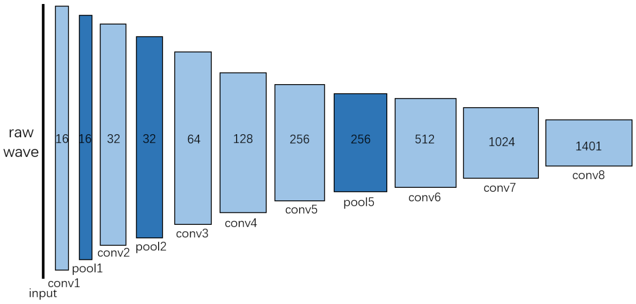


Figure 1: The structure of the soundnet. The number inside of the layer is the number of filters.

the groundtruth emotion labels. We then scale the features into $[-1, +1]$ for each dimension as a post-precessing step.

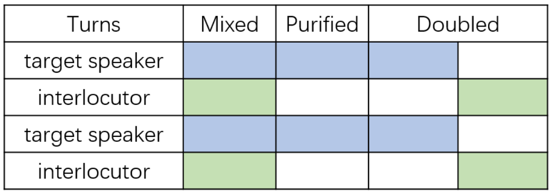
Soundnet Feature: Soundnet [22] is an 1-dimensional convolution network which directly learns from the raw audio waveforms. It consists of full convolutional layers and pooling layers as shown in Figure 1, so it could deal with variable length audio waves. In order to train such a network from scratch, Aytar et al. [22] transfer the knowledge (concepts probability distributions) learned from visual models into the sound on massive video dataset. The features extracted from middle layers of the soundnet achieve significant improvement over the traditional MFCC features on the audio event recognition task [22]. Since convolution networks have great generalization ability, in this work, we extract features from the `conv5` layer of soundnet directly as the short-time acoustic features. We set the sampling rate of the raw waveform to be 20,480 so that the shift of the `conv5` features is 100ms.

Interlocutor Influence: The interlocutor influence (a person's influence on the interacting partner's behaviors) is shown to be important in dyadic human interactions [23]. Since the original waveforms in the dataset contain the speeches from the two interacting speakers, we consider to explore the interlocutor influence when constructing the acoustic feature sequences. The proposed three strategies are illustrated in Figure 2. In the `Mixed` strategy, no information about the interlocutor speaker is given, which means that the acoustic features of the target speaker and interlocutor are mixed in the sequence. The `Purified` strategy takes into account the speech turn information, which only keeps the acoustic features of the target speaker in the acoustic sequence and pads zeros for the interlocutor's turns. Our third strategy `Doubled`, doubles the dimensionality of the acoustic features with one half to represent the features from the target speaker and the other half for the interlocutor.

3.2 Visual Features

We adopt the provided bag-of-video-words (`BoVW`) [2] as our long-time visual features, which covers facial information with window of 6s and shift of 100ms.

For the short-time visual features, we mainly use the facial appearance feature in each frame. Since the deep convolutional neural networks (CNN) are the state-of-the-art models in many vision tasks and have achieved superior performance for facial expression recognition [24], in this work we explore two kinds of CNN structures to extract the facial features. Both the CNNs are pretrained on the FER+ dataset [24] following the standard data split to classify 8 different facial expressions.



Turns	Mixed	Purified	Doubled	
target speaker	🟦	🟦	🟦	
interlocutor	🟩			🟩
target speaker	🟦	🟦	🟦	
interlocutor	🟩			🟩

Figure 2: The three strategies for acoustic feature sequence construction to consider interlocutor influence. The blue boxes denote the features from target speaker's speeches. The green boxes denote the features from interlocutor's speeches. The white boxes denote empty features which are padded with zeros.

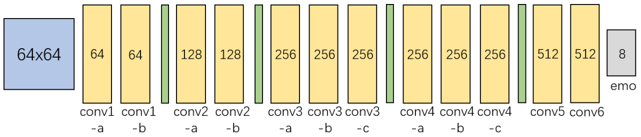


Figure 3: The structure of the VGGFace model. The input facial image is 64x64 pixels. Yellow boxes denote the convolution layer and the green boxes denote max pooling layers. The number inside of the box is the number of filters.

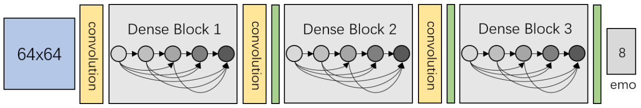


Figure 4: The structure of the DenseFace model.

VGG-style CNN (VGGFace): The characteristic of the VGG-style CNN is to construct the convolution layers with small convolutional filters, which are interleaved with max pooling and dropout layers, and when the input size reduces to its half, the number of convolutional filters doubles. Figure 3 shows the structure of our custom VGGFace network. Our VGGFace model achieves the accuracy of 82.8% on the testing set of FER+ dataset. Though the target is different between the VGGFace model and the AVEC affective task, the middle layers can contain useful features related to the facial emotions. So we extract facial features from the `conv5` and `conv6` layers respectively as the frame-level feature.

DenseNet-style CNN (DenseFace): The recent proposed dense convolution network (DenseNet) [25] has achieved the state-of-the-art performance in the image recognition task. It connects all the preceding layers as the input for a certain layer as shown in Figure 4, which can strengthen feature propagation and alleviate the gradient vanishing problem. Besides, due to the feature reuse, DenseNet only needs to learn a small set of new feature maps in each layer and thus requires fewer parameters than traditional CNNs, which is more suitable for small datasets. We follow the DenseNet-BC structure proposed in Huang et al. [25] with the growth rate as 12 and depth as 100, which consumes about 86% fewer parameters than our VGGFace model and achieves 2% higher accuracy on the FER+ testing set. We extract the activations from the last mean pooling layer and refer the feature as `denseface`.

To match with the shift of groundtruth labels, we apply mean pooling over consecutive 5 facial CNN frame features. The frames where no face is detected are filled with zeros.

3.3 Textual Features

The speech transcriptions in German and the corresponding start and end time for each speech turn are provided in the challenge [2]. Since the textual content is not continuous in time, we only extract the textual features at the turn-level. For the training of the non-temporal model, we only use the turn-level textual features, and the turn-level labels are set as the Gaussian weighted sum of all the frame-level groundtruth labels in the turn. Adjacent turn-level features can be concatenated to incorporate more contextual information. For the temporal model, we repeat the turn-level textual features within the turn duration and pad zeros for others. The turn-level textural features are described as follows.

Bag-of-Words: We remove the stop words in the training transcriptions and use the remained words as our dictionary. The word frequencies are used as the Bag-of-Words (`BoW`) features.

Word Vectors: Word vectors are distributional word representations learned from massive textual dataset [26], which are not only more compact than BoW representation but also are related to the semantic meanings of the words. We adopt an unofficial pretrained German word embedding model [27] with 300 embedding dimensions and mean pool the word embeddings in the turn as turn-level features. Since the quality of the German word embedding might not be very good, we also translate the German transcriptions into English by Google Translator and use the Google English word embedding model [28] to extract textual features.

4 EMOTION RECOGNITION MODELS

In this section, we introduce the emotion recognition models in details. We utilize the Support Vector Regression model [29] as the non-temporal model, which is widely used for dimensional emotion recognition; the temporal model and the multi-task learning framework are described as below.

4.1 Temporal Model

Long-short term memory (LSTM) recurrent neural network (RNN) [14] is the state-of-the-art temporal model for sequence prediction. It employs a memory cell to store information and additional gates to control the information flow to address the gradients vanishing problem.

We adopt a LSTM variant with peephole connections [30]. The memory cell c_t is the core of the LSTM which records the history of inputs before the current timestep. The three gates are functions of previous output state h_{t-1} and the current input x_t. The input gate i_t controls how much new information of current input x_t can be accepted. The forget gate f_t is used to control whether the LSTM should forget its previous memory c_{t-1}, and finally the output gate o_t determines the amount of information from memory cell c_t to output state h_t. In this way, the LSTM is capable of capturing complex and long-term dependencies in the sequences. The formulas of the LSTM at timestep t are given below:

$$\begin{aligned}
\text{input gate}: &\quad i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\
\text{forget gate}: &\quad f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \\
\text{output gate}: &\quad o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_{t-1} + b_o) \\
\text{cell input}: &\quad g_t = \phi(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \\
\text{cell state}: &\quad c_t = i_t \odot g_t + f_t \odot c_{t-1} \\
\text{cell output}: &\quad h_t = o_t \odot \phi(c_t)
\end{aligned} \tag{1}$$

where σ is sigmoid function, ϕ is tanh function, $\odot$ denotes element-wise production, and$\{W_{**}, b_*\}$ are parameters.

We use the mean square error (MSE) as our loss function, which minimizes

$$L = \frac{1}{2T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 \tag{2}$$

$$\hat{y}_t = W_y h_t + b_y \tag{3}$$

where W_y and b_y are parameters to be learned which are used to predict the groundtruth labels, and T is the total timesteps of the data.

4.2 Multi-task Learning

In Table 1, we investigate the correlations between different emotion dimensions with the Pearson Correlation Coefficient (PCC). We can see that the arousal and valence are highly correlated with each other, which might share representations and constrains the representation to be more emotion related. Therefore, we propose the multi-task learning framework to learn the arousal and valence simultaneously. We use the mean square error (MSE) as our loss function, which minimizes

$$L = \frac{1}{2T}\sum_{t=1}^{T}(\alpha(\hat{y}_t^{(a)} - y_t^{(a)})^2 + \beta(\hat{y}_t^{(v)} - y_t^{(v)})^2) \tag{4}$$

$$\hat{y}_t^{(a)} = W_y^{(a)} h_t + b_y^{(a)} \tag{5}$$

$$\hat{y}_t^{(v)} = W_y^{(v)} h_t + b_y^{(v)} \tag{6}$$

where $W_y^{(*)}$ and $b_y^{(*)}$ are parameters and α, β are hyper-parameters to balance the two tasks. We set the $\alpha = \beta = 1$ in our experiments which shows good performance.

We apply the multi-task learning strategy in two kinds of network training. The first is to finetune the DenseFace model. Since all the input layers are connected to the targets

Table 1: PCC of different emotion dimension pairs on the training set.

	Arousal	Valence	Likability
Arousal	-	0.625	-0.012
Valence	0.625	-	-0.181
Likability	-0.012	-0.181	-

in the DenseNet CNN structure, there are not explicit middle layers like the VGGFace model which might hurt the generalization ability of the DenseFace features. Therefore, we use the facial images with their dimensional emotion labels to finetune the DenseFace model. The second is to train the temporal LSTM-RNN model to predict arousal and valence simultaneously.

5 EXPERIMENTAL RESULTS

5.1 Dataset

The AVEC2017 dimensional emotion dataset is a subset of the Sentiment Analysis in the Wild (SEWA) database [2], which records audiovisual spontaneous human-human interactions in the wild. There are 64 German subjects in the dataset and are divided into training with 36 subjects, validation with 14 subjects and testing with 16 subjects. Subjects participated in pairs and were asked to discuss the commercial product they had just viewed. The duration of each conversation is at most 3 minutes. Besides the common emotional dimensions: Arousal for the emotion activation and Valence for the emotion positiveness, AVEC2017 introduces another dimension of likability, which presents the user's preference to the commercial product. All three emotion dimensions are annotated every 100ms and scaled into [-1, +1]. The concordance correlation coefficient (CCC) works as the evaluation metric for this challenge, which is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{7}$$

where μ_x and μ_y are the means of the sequence x and y, and σ_x and σ_y are the corresponding standard deviations. ρ is the Pearson Correlation Coefficient (PCC) between the x and y. PCC and Root Mean Square Error (RMSE) can also be used to discuss the prediction performance.

5.2 Experimental Setup

We implement our temporal model and multi-task strategy with the tensorflow [31] deep learning framework. The number of layers in the LSTM are set to be one and the hidden size is optimized for different input features. We use the truncated back propagation through time (BPTT) with max step of 100 timesteps to train our LSTM networks. Adam optimizer is applied and the learning rate is initialized from 0.01 and reduce half every 50 epochs. We train at most 200 epochs for each model. Dropout is adopted to avoid overfitting with dropout rate of 0.5. We optimize the feature delay timesteps for both the non-temporal and temporal models. The predictions of our models are smoothed by simply averaging the predictions within a fixed window.

Table 2: CCC performance using long-time audio and visual features and SVR model on the validation set. A, V, L denote the arousal, valence and likability respectively.

		A	V	L
audio	BoAW	0.2565	0.3161	0.0572
visual	BoVW	**0.4408**	0.3575	0.0757
BoAW-BoVW		0.4324	**0.4200**	-

Table 3: CCC performance of the likability using different textual features and contexts with SVR model on the validation set.

context	1	3	5	7
GermanBoW	**0.3340**	0.2908	0.2909	0.2886
GermanWordvec	0.3230	0.3250	**0.3620**	0.3384
EnglishWordvec	**0.3700**	0.3327	0.3241	0.3327
LateAverageFusion	**0.4006**			

5.3 Non-temporal Model Results

We use the SVR as our non-temporal model, and the long-time multimodal features as the models' inputs. As shown in Table 2, the visual BoVW feature consistently outperforms the audio feature and the multimodal fusion improves the valence prediction. However, the audio and visual features perform extremely bad for the likability dimension. Since the likability is what people think about the commercial product, we assume that the likability might be mainly reflected from peoples' comments from speech contents rather than the audiovisual signals. Therefore, we conduct a preliminary verification to check whether the likability labels is highly correlated with the speech turns. We set the likability as -1 when the subject speaks (because most of the likability is negative) and 0 otherwise as the baseline of speech turn prediction. The PCC of the speech turn prediction and the groundtruth likability is 0.2080 with 2s feature delay, which is already higher than the predictions from audio and visual features. This verification suggests that likability is closely related to speech turns. Therefore, we employ our turn-level textual features to predict the likability. Table 3 presents the CCC performance of different textual features with the SVR model on the validation set. The context denotes the number of turns we concatenated together with the reference turn in the center. We can see that for the `GermanWordvec` feature, contextual information significantly improves the performance, but for the `EnglishWordvec` feature, only using the current turn-level feature performs best which might relate to the error accumulations from the imperfect translation. The superior performance of the `EnglishWordvec` feature than the `GermanWordvec` feature also suggests that the English word embedding model we used is better than the German word embedding model. And for the GermanBoW feature, the concatenation of more context features results in high dimensionality and can easily cause overfitting.

5.4 Temporal Model and Multi-task Learning Results

Table 4 shows the CCC of the acoustic short-time features with different interlocutor influence strategies using the temporal model. For both the IS10 and soundnet features, the Purified strategy outperforms the Mixed strategy which suggests that the features of the interlocutor should not be treated as the same as the target speaker. The Doubled strategy that contains but separates the speeches of the target speaker and interlocutor achieves the best performance for all the emotion dimensions. This demonstrates that the interlocutor influence is beneficial for the emotion recognition of the target speaker if used properly. We can also see that the soundnet features without any finetuning on the emotion dataset are comparable with the hand-crafted IS10 features, which suggests that we may further improve the deep learned acoustic features for the emotion recognition task with careful finetuning in the future.

In Table 5, we show the performance of the visual short-time features. The vggface.conv5 feature generalizes better from the categorical to dimensional emotion recognition than the vggface.conv6 feature. Due to the structure of the DenseFace model, the denseface feature doesn't perform as well as the VGGFace features. However, finetuning the DenseFace model with each emotion dimension significantly improves the performance. The multi-task finetuning with arousal and valence targets together achieves further performance gain than the single task finetuning.

For acoustic and visual features, no feature delay is required to compensate the cognition delay of the annotators since the LSTM-RNN is able to capture the time dependencies of that range. But for the textual features, our experiments show that the best performance is achieved with 1s feature delay, which suggests that human processes for longer duration to analyze the texts than audiovisual signals. Table 6 presents the performance of the textual features with 1s feature delay using the LSTM-RNN. We can also see that the temporal model boosts the all the unimodal feature performance of the non-temporal model.

We use the early fusion strategy to explore the complementarity of the multimodal features and various feature combinations. Some of the top performance of the multimodal fusion for each emotion dimension is presented in Table 7, which all improve the performance of unimodal features. However, as shown in our testing performance[1], the multimodal fusion is beneficial for the arousal and valence prediction, but it hurts the performance on the testing set and is worse than the textual modality fusion system for likability prediction. We might also overfit the validation set for the likability prediction using the multimodal fusion.

Table 4: CCC performance of the acoustic short-time features with different interlocutor influence strategies using the temporal model LSTM-RNN.

feature	strategy	A	V	L
IS10	mixed	0.4575	0.4155	0.1281
	purified	0.4957	0.4527	0.1836
	doubled	**0.5236**	**0.5042**	**0.2726**
soundnet	mixed	0.4220	0.4020	0.1883
	purified	0.4853	0.3963	0.1992
	doubled	**0.5268**	**0.4469**	**0.3544**

Table 5: CCC performance of the visual short-time features using the temporal model LSTM-RNN. (*) denotes the emotion dimensions that we use to finetune the DenseFace model.

feature	A	V	L
vggface.conv5	0.6231	**0.7079**	**0.2348**
vggface.conv6	0.5565	0.6103	0.1206
denseface	0.5343	0.6270	0.1657
denseface.tune(AV)	**0.6747**	0.6934	-
denseface.tune(A)	0.6147	-	-
denseface.tune(V)	-	0.6740	-
denseface.tune(L)	-	-	0.2236

Table 6: CCC performance of the textual features using the temporal model LSTM-RNN.

	A	V	L
EnglishWordvec	0.4448	0.5196	**0.4662**
GermanWordvec	0.4775	0.5321	0.4168

Table 8 shows results using the multi-task learning strategy. We can see that the capacity of the model should be increased in the multi-task model to outperform the performance in the single task settings.

5.5 Run Submission

For arousal and valence prediction, we average multiple multimodal systems while for the likability prediction the best performance is achieved mostly with the textual features. We also scale and shift the predictions according to the statistics on the validation set. Table 9 presents our best results on the challenge validation and testing set.

6 CONCLUSION

In this paper, we present our approaches for AVEC 2017 emotion subtask, in which we predict continuous Arousal, Valence and Likability values based on the audiovisual data.

[1] Since we fuse multiple multimodal systems in the submission, it is inconvenient to present the details in our paper.

[2] IS10-soundnet-vggface.conv5-GermanWordvec

[3] IS10-soundnet-vggface.conv5-denseface.tune(AL)-EnglishWordvec-GermanWordvec

Table 7: Some of the top performance of multimodal fusion for each emotion dimension on the validation set.

	features	hidden	RMSE	PCC	CCC
A	IS10-vggface.fc5-GermanWordvec	200	0.0963	0.7474	0.7318
	IS10-soundnet-vggface.conv5-denseface.tune(AL)-EnglishWordvec-GermanWordvec	400	0.0922	0.7704	0.7503
V	IS10-soundnet-vggface.conv5-GermanWordvec	200	0.0941	0.7706	0.7601
	IS10-soundnet-vggface.conv5-denseface.tune(AL)-EnglishWordvec-GermanWordvec	400	0.0889	0.7854	0.7757
L	EnglishWordvec-GermanWordvec	200	0.1076	0.5023	0.4799
	IS10-vggface.conv5-denseface.tune(L)-EnglishWordvec-GermanWordvec	300	0.1012	0.6130	0.5786

Table 8: Performance comparison of the multi-task learning and uni-task learning on the validation set.

		Arousal			Valence		
feature	hidden	RMSE	PCC	CCC	RMSE	PCC	CCC
set1[2]	250	0.1042	0.7154	0.6040	0.0947	0.7543	0.7052
	300	0.0983	0.7633	0.6498	0.0921	0.7759	0.7178
	400	**0.0901**	**0.7900**	**0.7534**	**0.0884**	**0.7925**	**0.7806**
	best uni-task	0.0963	0.7474	0.7318	0.0941	0.7706	0.7601
set2[3]	500	**0.0844**	**0.8152**	**0.8014**	**0.0864**	**0.8001**	**0.7921**
	best uni-task	0.0922	0.7704	0.7503	0.0889	0.7854	0.7757

Table 9: Performance of our submission system on the validation and testing set.

	Arousal			Valence			Likability		
	RMSE	PCC	CCC	RMSE	PCC	CCC	RMSE	PCC	CCC
validation	0.086	0.823	0.823	0.091	0.806	0.796	0.113	0.546	0.545
testing	0.086	0.702	0.672	0.081	0.758	0.756	0.146	0.520	0.509
baseline testing [2]	-	-	0.375	-	-	0.466	-	-	0.246

We explore all available modalities including the acoustic, visual and textual modalities and multimodal fusion methods. We also explore the interlocutor influence when constructing the feature sequences. The fusion of the three modalities are beneficial for the arousal and valence prediction, while the textual features are most effective for the likability prediction. We then compare the non-temporal model SVR and temporal model LSTM-RNN. Our results show that the temporal model LSTM-RNN not only can alleviate the feature engineering efforts such as construction of contextual features and feature delay, but also improve the recognition performance significantly. Finally, since different emotion dimensions are correlated with each other, we apply the multi-task learning strategy to collaboratively predict multiple emotion dimensions with the shared representations. Our approaches significantly improve the baseline systems on the challenge testing set. In the future, we will extract more powerful deep learned audio and visual features and model the interlocutor influence.

7 ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Plan under Grant No. 2016YFB1001202. We appreciate the thoughtful discussion with Dr. Shilei Zhang and Dr. Yong Qin and the support from IBM China Research Lab.

REFERENCES

[1] Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2014.

[2] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schimitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2017.

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

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

[5] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S Huang. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 97–104. ACM, 2016.

[6] Filip Povolny, Pavel Matejka, Michal Hradis, Anna Popková, Lubomir Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 75–82. ACM, 2016.

[7] Shizhe Chen and Qin Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM, 2015.

[8] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM, 2015.

[9] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3, 2014.

[10] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad. Emotion recognition using acoustic and lexical features. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[11] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 41–48. ACM, 2015.

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

[13] Shizhe Chen and Qin Jin. Multi-modal conditional attention fusion for dimensional emotion prediction. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 571–575. ACM, 2016.

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

[15] Rui Xia and Yang Liu. Leveraging valence and activation information via multi-task learning for categorical emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5301–5305. IEEE, 2015.

[16] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015.

[17] Jonathan Chang and Stefan Scherer. Learning representations of emotional speech with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1705.02394*, 2017.

[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[19] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al. English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*, 2017.

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

[21] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[22] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[23] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Modeling Mutual Influence of Interlocutor Emotion States in Dyadic Spoken Interactions. In *Proceedings of Interspeech 2009*, Brighton, UK, September 2009.

[24] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283. ACM, 2016.

[25] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *CVPR*, 2017.

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

[27] German Word Embeddings. http://devmount.github.io/GermanWordEmbeddings/, 2017. [Online; accessed 19-July-2017].

[28] English Word Embeddings. https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing, 2017. [Online; accessed 19-July-2017].

[29] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[30] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[31] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.