Video Emotion Analysis via Deep Temporal Convolution

Zhipeng Bao
Tsinghua University
bzp15@mails.tsinghua.edu.cn

Hua Xu
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University
xuhua@tsinghua.edu.cn

Abstract

With the surge of social websites, hundreds of online videos are generated every day, which makes multimodal sentiment analysis an increasingly popular research hot spot. There are three main characteristics of outstanding multimodal sentiment analysis models: real-time analysis of sentiment, accurate extraction of multimodal features, and efficient feature fusion strategy. The current well-performing models still have room for improvement in terms of these characteristics. On the basis of this fact, this work proposes improvement measures for each of these three aspects. We have designed a deep architecture to extract the features of the multimodal data, a temporal convolution neural network to improve the operational efficiency and also proposed an effective weighted fusion strategy. An entire model for multimodal sentiment analysis combining these three architectures is also proposed. Experimental results on the real-world dataset indicate that the proposed model outperforms the other state-of-the-art models in terms of accuracy, efficiency and robustness. Targeted experiments also prove the superiority of the proposed deep feature-extracting network, temporal convolution network and weighted fusion strategy.

1. Introduction

With the explosive growth of the Internet, a rich number of self-recorded videos are generated and posted each day. In such case, video analysis has become vitally important for both companies and researches [1]. Video emotion analysis, also named multimodal sentiment analysis, is one research hot spot in the field of video analysis [18, 21]. The main task of it is to predict the sentiment polarity of a given video on the basis of visual (gestures and facial expressions), acoustic (voice and tone) and textual (spoken language) information [20].

The most advanced characteristic, and also the biggest challenge, of multimodal sentiment analysis is that it requires the classifier to comprehensively analyze and efficiently utilize the information of three modalities.

In order to achieve this goal, a lot of previous researches have been done. The early approaches for multimodal sentiment analysis are based on classic machine learning approaches, such as support vector machine (SVM) [12], and extracted linguistic, acoustic and visual features [24]. These methods are able to basically understand the human emotion from a video, but the performance is not stable. Then, with the development of hardware and the improvement of neural computing theory, neural network (NN)-based methods are introduced to this area, and further improvements on the feature processing are also proposed. Cambria *et al*. apply an entire whole end-to-end convolutional neural network (CNN) architecture to abstract the multimodal features [4]. The work of Poria *et al*. considers the relationship among the three modality and designed a context-based recurrent network[22]. [29] proposes a tensor fusion strategy to integrate multimodal features. A novel multi-stage fusion strategy has been proposed by Liang, which considers the sequential characteristic of the videos [17].

These popular models generally share a similar architecture, which contains two main parts: feature-extracting module and feature fusion module [29, 4]. Figure 1 illustrates this architecture. In our opinion, the feature-extracting module is more essential since the subsequent feature fusion process is based on the extracted features. Once the multimodal features gain improvements, many existing well-performing models can achieve better performances. However, current researches have not settled this matter well. Two types of features, classic features and learning-based features, are used in the current models. The classic features model human being's audiovisual charac-

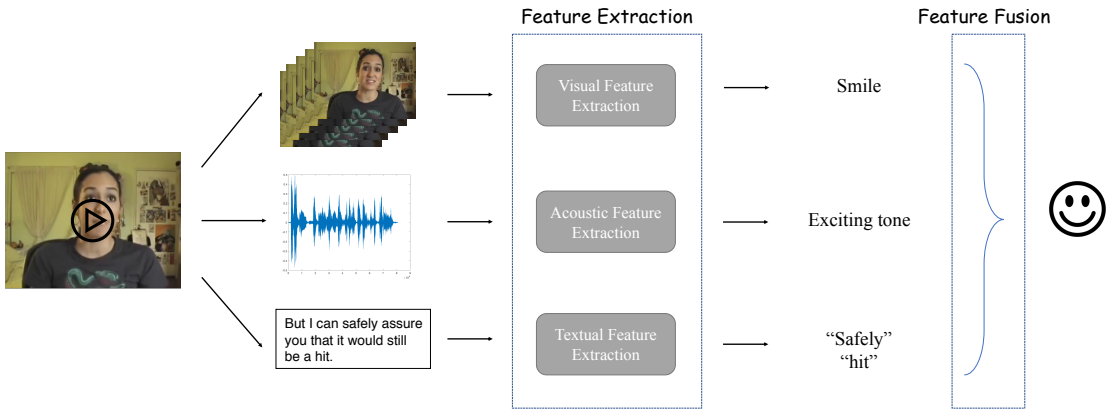

Figure 1. A typical framework for multimodal sentiment analysis.

teristics and extract facial texture features and temporal-spectral features based on the raw video utterances [3, 6]. However, this kind of methods has low robustness and generation ability for different types of videos. The other models extract the related features by designing and training neural networks [4, 20]. Compared with the classic methods, this kind of approaches is more accurate and robust. However, the current methods mainly take use of shallow CNN architectures, which cannot represent the videos well. In our opinion, deeper and well-designed neural network can be more effective for feature extraction thus benefits the multimodal sentiment analysis. Based on this understanding, we deepen the feature-extracting network and adapt residual modules to help abstract the features [11].

Another concern for video emotion analysis is the time consumption since the video emotion analysis is a real-time task in the real world most of the time. A great majority of the current models adopts recurrent neural network (RNN) or similar architectures to predict the sentiment polarity as they are capable to handle time-related tasks. However, recurrent models require much more time compared with equal-sized CNN models. Thus they have relatively poor performance in real-time emotion analysis. In order to tackle this issue, we adapt some revision to classic CNN model so that it is able to conduct temporal convolution as well. Compared with the recurrent models, the temporal convolutional network (TCN) can obtain consideration results with much less time consumption [2].

Finally, this study also proposes an entire end-to-end system for multimodal sentiment analysis that combines the deep feature-extracting network and temporal convolution. We have also introduced a dynamic weighting strategy for multimodal feature fusion. In order to verify the proposed approach, we have designed various experiments on a standard real-world multimodal dataset— CMU-MOSI dataset [30]. The comparable results indicate that the proposed deep feature extraction and temporal convolution work well in multimodal sentiment analysis and the whole model outperforms the other state-of-the-art methods in terms of accuracy, efficiency and robustness.

In summary, three major contributions of this study are listed below:

- We adapt a novel deep feature-extracting neural network for multimodal sentiment analysis. Features extracted by this architecture can characterize the videos more precisely.
- Temporal convolution network is introduced to improve the efficiency of the model. Compared with current structures, the proposed architecture can save a lot of calculating time with a comparable result at the same time.
- We also propose a weighted feature fusion strategy for multimodal sentiment analysis. Compared with the previous fusion strategies, the proposed approach can be more effective.
- This study further proposes an end-to-end deep framework for video emotion analysis. Various experiments on a real-world database indicate that the proposed model achieves the best performance among other state-of-the-art approaches in accuracy, efficiency and robustness.

2. Related Work

2.1. Deep Neural Network for Unimodal Feature Extraction

Deep neural networks (DNNs), especially deep convolutional neural networks (DCNNs) have gained great achievements with the role of feature-extracting architecture in object detection, audio analysis and text classification tasks.

For the task of image detection, deep architectures have gained great development since the first relatively

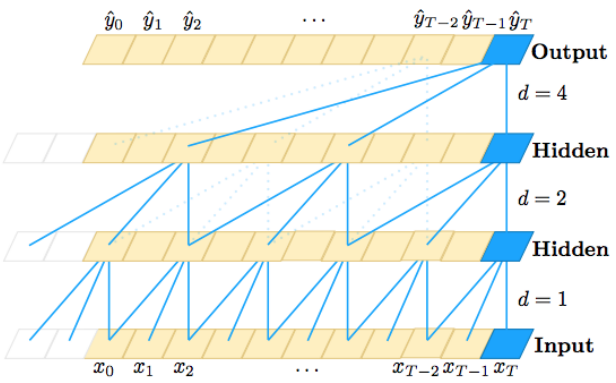

Figure 2. Example of causal convolution and dilated convolution in TCN architecture.

DCNN was introduced to image classification tasks [16]. Then well-designed networks, such as VGG [25] and ResNet [11], further improve the capability of CNN structure, especially for large-scale databases. Moreover, for the smaller datasets, the deep architectures can still be helpful by fine-tuning the pre-trained deep networks with transfer learning methods [14, 16].

For audio content analysis, deep architectures also achieve considerable results. VGGish model obtains pretty good achievements on large-scale audio analysis tasks [13]. For a long time, DCNN architectures are considered not suitable for text classification tasks since the texts usually do not contain high-level information. However, Conneau *et al*. prove that very deep architectures are also powerful for text classification tasks and their model outperforms a lot of popular models in several tasks [5].

These models provide reliable references and also inspirations for this work.

2.2. Temporal Convolutional Network

A lot of researches have committed to solving the problem of running time for RNN based architectures. TCN is one of the noticeable attempts [2].

The basic idea of TCN is to treat the time information as one dimension. To adapt CNN architecture to sequential tasks, causal convolution and dilated convolution are applied in the convolution process [26, 28]. For causal convolutions, the output at time step t is only related to the inputs at the same step t or previous steps. Dilated convolutions expand the receptive field of the filters by convolving the data block by block instead of step by step. Therefore, they can take use of a wider range of data, thus retrospect longer history, without increasing the quantity of the parameters. Combining these two specific structure, TCN can deal with long-history tasks. Figure 2 demonstrates a simple TCN structure that combines causal convolution and dilated convolution.

Compared with sequence-based neural networks, such as RNN and LSTM, TCN has a lot of advantages. It enables parallel computing in the network as the data do not need to be processed sequentially, which saves a lot of running time. It also requires fewer memories for building and training, and it can also deal with unfixed-length inputs. Based on these advantages, TCN architecture has been applied to several NLP tasks and obtained considerable results [19]

3. Proposed Method

In this part, we depict the proposed model in details. The proposed model generally follows the architecture demonstrated in Figure 1, which contains two main parts: feature extraction module and feature fusion module. However, as the feature extraction network for each modality has significant differences, we describe them separately in this section.

3.1. Visual Feature-extracting Channel

The visual feature-extracting sub-network takes the frames of the videos as the input and generates the features of the whole video. This process is modeled as a two-step framework: visual feature extraction from single images and the following time-based feature regeneration. For the single image feature-extracting module, We have proposed a well-designed deep architecture to abstract the textural features, and it is followed by a TCN architecture to further utilize the sequential features. The whole architecture of this module is demonstrated by Figure 3.

3.1.1 Single Image Deep Feature Extracting Network

Deep architectures have played more and more important roles in image recognition tasks since AlexNet was proposed [16]. Deeper architectures allow the model to represent the images more precisely. Further modifications are proposed to deal with the gradient vanishing problem in the deep networks [11].

In this sub-architecture, we adapt the architecture of ResNet [11] and apply the identical blocks and convolution blocks in our network. The structure of these two blocks are illuminated by Figure 4.

Specifically, we design a residual convolution neural network with one convolution layer, three identity blocks and three conv blocks. The whole number of convolution layers is 19. Besides, between each two blocks (including the first convolution layer), we also add a dropout layer, a batch-normalization [15] layer and a max-pooling layer. Thus, the output of the feature is re-scaled to $\frac{1}{64}$ in size compared with the original image. To obtain the feature output, we add a global maxpooling layer to generate the feature vector. The final output of this sub-module has the dimension of $[1, 512]$.

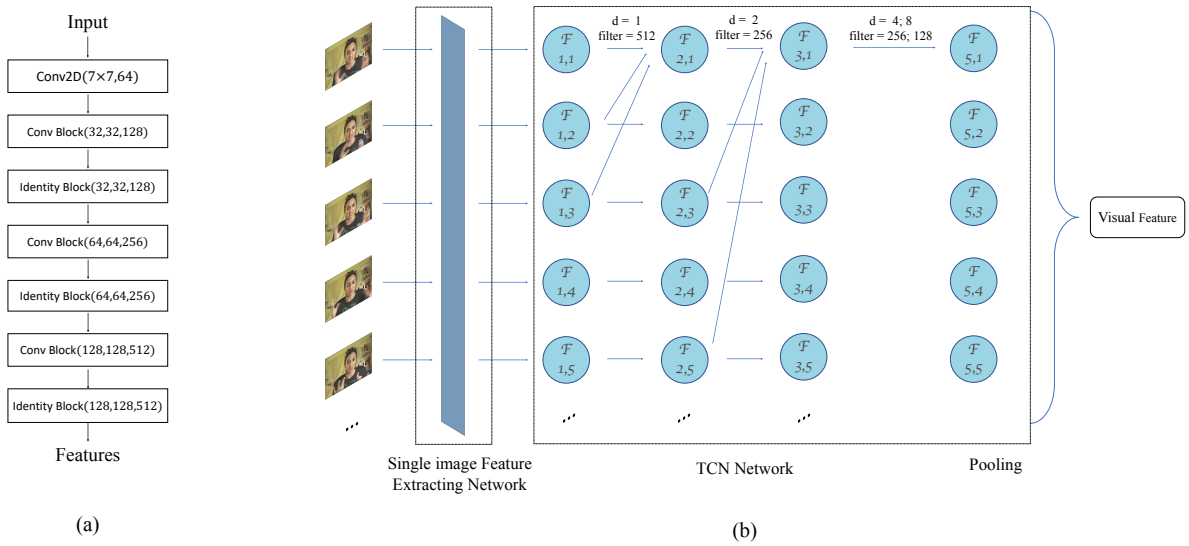

Figure 3. Architecture of the visual feature extraction network. (a) structure of deep feature-extracting network; (b) overall architecture of visual sub-network.

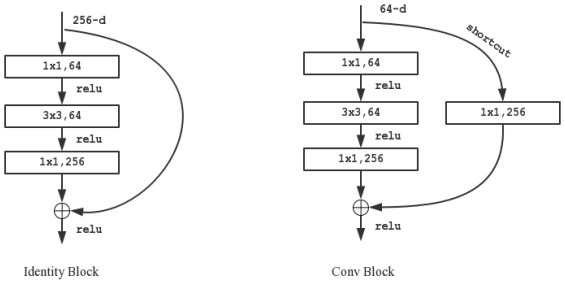

Figure 4. An example of identity block and conv block.

3.1.2 Temporal Convolution Network

On the basis of the extracted features from image frames, we have designed a TCN architecture to further process the features.

The basic unit of the TCN architecture is the temporal convolution filter. For one temporal convolution filter with dilation factor d, the output is obtained by Eq. (1)

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i}. \qquad (1)$$

In the formula, d is the dilation factor, k is the filter size, and $s - d \cdot i$ accounts for the direction of the past. Thus, dilation can be treated as introducing a fixed step between every two adjacent filter taps. If $d = 1$, the dilated convolution reduces to a regular convolution. Applying larger dilation factor enables an output at the top level to represent a wider range of inputs, thus effectively expanding the receptive field.

In our experiments, we apply 4 temporal convolution layers which have the dilation factors of 1, 2, 4, and 8 respectively, and the numbers of filters are 512, 256, 256, and 128. The final output of this channel is the global maximum value of the output of the last temporal convolution layer, which has the dimension of $[1, 128]$.

3.2. Acoustic Feature Extraction

For the acoustic sub-channel, we also build a deep convolution structure to extract the features. The audio is first divided into segments of one second. Then these raw utterances are sent to the network as the input. Figure 5 illustrates the architecture of acoustic feature-extracting networks. The extracted feature of each second has the dimension of $[1, 128]$ and the whole feature of the given audio utterance has the following form:

$$F_{acoustic} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,128} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,128} \\ \cdots & & & \\ f_{l,1} & f_{l_2} & \cdots & f_{l,128} \end{bmatrix}. \qquad (2)$$

where l is the length of the audio and $f_{i,j}$ is the j^{th} feature extracted from the i^{th} second of an audio utterance.

As most of the audios are only 3 to 4 seconds, timing information has little effect for this this channel. So we use the global maximum value of all the features to represent the most prominent features. Then we conduct a network with 3 fully connected layers to obtain the final feature output. The hidden dimensions of the layers are set to 128, 64, and 64. Therefore, the shape of the output of this channel is $[1, 64]$.

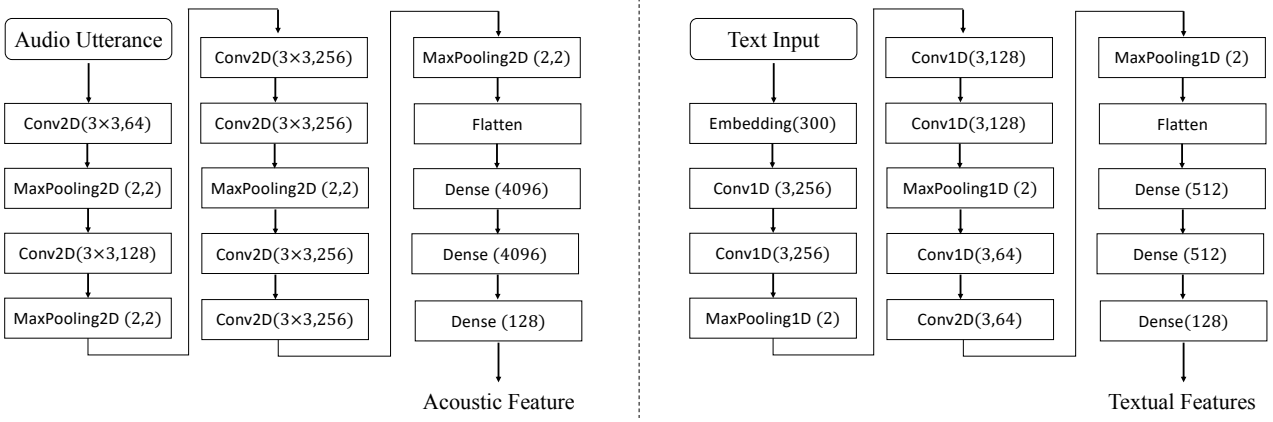

Figure 5. Architecture of audio feature extractor and textual sub-network. Left: audio feature extractor, it will extract acoustic features from the audio utterance instead of the whole audio slide; right: text feature extracting network.

3.3. Textual Feature Extraction

As the input texts cannot be processed directly, we first apply an embedding network for this channel to transfer the words to word vectors [9]. The embedding layer retrieves a sequence of word vectors for the input text, which has the following form:

$$\mathbf{V} = [v_1^T, v_2^T, v_3^T, \cdots, v_n^T]. \tag{3}$$

in which $\mathbf{V}$ is output of this layer and v_i represents the corresponding word vector of the i^{th} word in the text. During our experiments, the dimension of the embedding vectors is set to 300.

Then, we also designed a deep convolution network with 6 convolutional layers to generate the textual features based on the supposition that deeper architectures can also benefit the textual feature extraction. Then 3 fully connectly layers are followed to attain the final output of this channel. The overall architecture of this channel is also illustrated in Figure 5. The output of this channel has the shape of $[1, 128]$.

3.4. Feature Fusion

The fusion network is another significant part for multimodal sentiment analysis. The basic fusion approach is the concatenation of all the features. However, this approach does not work well for multimodal sentiment analysis due to the insufficient utilization of the features. [8] concludes some effective fusion approaches. Inspired by this work, we propose to use a weighted fusion strategy.

First, we concatenate the three types of features and include a bias value in the end of it. Then, we form a feature vector which has the following form:

$$F_{feature} = [F_v,\ F_a,\ F_t,\ Bias], \tag{4}$$

where $F_v,\ F_a,\ F_t$ represent the extracted feature from the visual, acoustic and textual network and $Bias$ is set to 1 to further revise the feature vector. As these three features do not share the same contribution to the final result and each unit in the unimodal feature vector counts differently, we introduce a weight vector W to measure their contributions and the final weighted feature is obtained by the following equation:

$$F_{wt} = W \odot F_{feature} = [w_1 \cdot f_1,\ w_2 \cdot f_2,\ \cdots,\ w_k \cdot f_k] \tag{5}$$

In the formula, F_{wt} is the weighted feature, and k is the length of the feature vector, which is equal to 321 for our model. w_i is the i^{th} element in the weight vector W and f_i represents the i^{th} element of the concatenated feature vector. The weight vector is also trainable in the architecture.

Since the weighted feature and the original feature vector are consistent in form, the weighted feature vector can also be obtained through training the parameters of the former network theoretically. However, the weight vector W has a larger update rate than the feature vector $F_{feature}$ if the whole network share the same learning rate, as the depth of the feature-extracting network is quite large. So by introducing the weight vector, the weighted feature are updated at two different rates, which is more flexible. We envision that it can benefit the model better compared with training the original feature vector $F_{feature}$ directly.

Once we obtain the weighted feature vector, two fully connected layers are followed to attain the final prediction result. The numbers of hidden units of these two layer are 64 and 64.

4. Experiments

4.1. Experiments Setting

In order to verify the effect of the proposed model, we design different kinds of experiments on CMU-MOSI dataset [30] which is a widely used dataset for multimodal sentiment analysis tasks. This dataset contains video opinions from YouTube movie reviews. It is annotated on a seven-step Likert scale from very negative to very positive by five different Mechanical Turk annotators. The final label of each video is the mean value of annotation results which has the form of a float number. Thus, the final label of MOSI dataset is a float number varying from -3 to 3, which makes it convenient to conduct both regression and classification experiments on it. The dataset is segmented by opinion utterances and there are 2199 opinion utterances with 93 distinct speakers in total. Each utterance has an average length of 4.2 seconds. In our experiments, we remove the relatively long utterances and control the length of the videos shorter than 20 seconds. At last, 2150 utterances are selected and used in our experiments.

During the experiments, we separate the MOSI dataset randomly and select 80% as the training set and 20% as the test set. For the training set, we randomly split 20% as the validation set. Finally there are 1476 samples in the training set, 344 samples in the validation set and 430 samples in the test set.

For the comparing models, we choose five novel and well-performing multimodal sentiment analysis models as the comparative approaches to evaluate the overall effect of the proposed method. The details of these models are listed as the following:

SVM-MD [30] is a SVM model trained with multimodal features. It is one of the earliest machine learning models for multimodal sentiment analysis.

SAL-CNN [27] proposes a novel "select-additive learning method" which attempts to prevent identity-dependent information from being learned in a deep neural network.

BC-LSTM [22] is a context-dependent model which considered the connection between the three modality.

MV-LSTM [23] adopts special regions inside one LSTM model to extract and group the multimodal features.

TFN [29] proposes a tensor-based fusion strategy for multimodal feature fusion that is better than concatenation.

DTCN is the model proposed in this paper.

4.2. Experiments Design

To further verify the other capabilities of the proposed method, we conduct three tasks in MOSI dataset: 1) Binary Sentiment Classification 2) Five-Class Sentiment Classification and 3) Sentiment Regression in range [-3,3]. In the first task, the metrics we use are the accuracy and F-1 score. For the second tasks, we report accuracy to evaluate the performance and for the last task, mean absolute error (MAE) is used to compare the performances. These three tasks and four metrics are used in all the experiments and they will help us to compare and analyze the performance of all the compared models.

There are four experiments in total, and the details and purposes are enumerated below:

1. We performed the three tasks on the MOSI dataset to comprehensively compare the performance of our model and other state-of-the-art approaches. All the 6 models are included in this experiment.

2. In order to verify the superiority of the proposed feature-extracting network, we provide two types of features for the same architecture: deep features and classic features. Only two well-performing models, DTCN and TFN, are included in this experiment.

3. As we introduce TCN to multimodal sentiment analysis tasks, it is essential to compare the effect of TCN. To achieve this goal, we replace the TCN architecture in our model by LSTM and compare the performance and time consumption. Only the proposed DTCN model is included in this experiment. The performance of the visual channel and the whole model are monitored and compared in this experiment.

4. To analyze the performance of each modality and the fusion strategy, we have also made ablation study in our experiments.

In each experiment, we use the opinion level annotation directly as the output label. For the classification tasks, we project the outputs to different classes depending on the range of the outputs. As this classification standard may differ from the original paper of the previous models, we build and retrain all the compared models by us according to the original articles or the released codes.

4.3. Network Training

During our experiments, some adjustable parameters are set based on our experience and attempts. For all the models, the $batch\ size$ is set to 128 and the optimizer we use is $Adam$ with $learning\_rate$ initialized to 0.01. The $dropout\_rate$ is 0.5 and we apply an early stopping method with training patience as 10.

To attain a more robust result, we introduce Gaussian noise in the training process. We add them to the original data and the variance of Gaussian noise is set to 0.3. In order to obtain a more precise network, we pre-trained the visual and acoustic deep feature-extracting network on large-scale datasets and fine-tune them in MOSI dataset during our experiments. The visual deep network is pre-trained on ImageNet [7], an object classification database with more than

Model	Binary		5-Class	Regression
	Acc(%)	F-1(%)	Acc(%)	MAE
SVM-MD	69.07	68.88	32.09	1.19
SAL-CNN	70.93	70.64	34.65	1.11
MV-LSTM	72.09	72.11	35.81	1.02
BC-LSTM	72.33	72.07	36.98	1.08
TFN	73.95	73.83	37.67	1.04
DTCN	**78.14**	**77.44**	**41.16**	**0.91**

Table 1. Main results on MOSI dataset.

Model	V	A	V+A+T
TFN	65.35	58.60	73.95
DF-TFN	66.04	60.93	74.18
CF-TCN	64.42	58.83	74.42
DTCN	69.30	61.39	78.14

Table 2. Binary accuracy of models with different features. "V" stands for the visual channel, "A" stands for the acoustic channel and "T" represents the textual channel.

150 million samples, and the acoustic network is pre-trained on AudioSet [10], an audio classification database contains more than 2 million samples.

5. Result Analysis

5.1. Main Results

We first compare the performances of the proposed model and some other state-of-the-art models. Table 1 lists the comparative results of them.

From the results, we can see that the proposed DTCN model outperforms the other compared models significantly in terms of all the four metrics, which indicates that the proposed model is effective for various multimodal sentiment analysis tasks. The excellent performance of the model is the combination of feature extraction and feature fusion. For the binary classification task, we check the data in the test set and find there are 220 positive samples and 210 negative samples, which is relatively balanced. Under such condition, the models have equally effect on both positive and negative samples. Therefore, $F-1$ score and accuracy share the same trend for all the models, but both of them can represent the capability of the models.

We can also find that for more complicated tasks, our model is more advantageous. For the five-category classification problem, we have a 9.26% relatively improvement over the second-place model and our model has also greatly reduced the MAE by 12.5% compared with the second-place model for the regression task.

5.2. Effect of Deep Feature Extraction

One main contribution of this work is that we introduce an entire deep architecture to extract the features for video emotion analysis tasks. To verify the advantage of the architecture, we compare the performance of four specific models based on two kinds of features.

TNF uses the classic features in their experiments. The visual features are extracted by Openface [3] which captures the facial geometric features, and the acoustic features are extracted by COVAREP framework [6], which combines the statistical and frequency characteristics of the audio.

DF-TNF takes the proposed image-feature extraction structure and audio-feature extraction structure as feature-extracting network .

CF-TCN feeds the classic visual features to TCN and applies the fully connected layers on the mean value of the classic acoustic features.

DTCN is the proposed model using deep feature-extracting networks.

We only record the binary accuracy for this task. The results are listed in Table 2.

As can be seen, both DF-TFN and DTCN obtain a better performance when using the deep feature-extracting architecture. These results indicate that the proposed deep feature-extracting architecture can extract the features more precisely so that can provide more reliable information than the classic features.

The deep architecture does not bring a significant improvement for TFN model but the classic features pull down the performance of DTCN clearly. The reason is that TFN model is designed with the classic features and DTCN is designed based on the deep architectures. If we change the features or the architectures, the model should also be adapted. So the improvement is not significant but the drop is obvious. However, the deep features improve the performance of TFN without adaptation, which indicates the deep feature-extracting architecture is robust and can provide more precise information about the image frames or audio utterances.

5.3. Superiority Analysis of Introducing TCN

Another contribution of this study is that we design a TCN architecture to optimize the network. In order to verify the capability of TCN architecture, we replace the position of TCN by a LSTM model with the same output shape and compare the performances. For the models, we only conduct the experiments on our DTCN model. We also select the visual sub-channel to make the comparison as TCN influences the visual sub-channel directly. Besides, in the experiments, we set the deep architecture of visual feature-extracting network and acoustic training network as untrainable since these deep architectures will greatly reduce the speed of model training and thus make the comparison results indistinct.

The metrics in this experiment are the time consumption and the binary accuracy. We train all the models equally

Model	Runtime(s)	Binary Acc(%)
D-TCN	390.54	74.88
D-LSTM	665.03	73.72
D-TCN(V)	202.22	67.21
D-LSTM(V)	476.35	63.26

Table 3. Comparison of running time between TCN and LSTM.

with one Titan X GPU for 100 epochs. The batch size we choose is 128. The results of this experiment are listed in Table 3.

It is clear to see that TCN models save about half of the running time with an even higher accuracy compared with the similar LSTM architecture. This experiment reveals that the proposed TCN architecture is a fast and precise model, and it is suitable for multimodal sentiment analysis.

5.4. Ablation Study

The last part of our experiments is the ablation study. We evaluate the performance of each sub-channel and their influence on one another. For the whole multimodal model, we have also test the performance with different fusion strategies to verify the advantage of the weighted fusion strategy. In the experiment, we use concatenation fusion strategy (CF), tensor fusion strategy (TF) [29] and weighted fusion strategy (WF) to make the comparison. The related results are listed in Table 4.

For the three unimodal models, the experimental results are consistent. Textual model has the best performance, then followed by the visual model, and the acoustic model has the worst performance. We analyze and think the reason is that when people express emotions, the change in tone is not clear compared to facial expressions and it is difficult to conduct the semantic understanding based on the audio directly. Audio sentiment analysis is therefore relatively difficult for researchers. We can also see that when combining the other features, the performance generally improved except for "A+T" model on the binary classification task. We analyze the reason and find the gap between the "A+T" model and "T" model is small and the size of the dataset is not big. Thus, we believe the fluctuation of the models and it cannot indicate the capability of the models.

Then, for the fusion strategies, the proposed weighted fusion strategy outperforms the other strategies significantly among all the tasks, which indicates that the weighted fusion strategy is proper for multimodal sentiment analysis. For the other two fusion strategies, they do not share consistent performances. We think this also indicates that these two fusion strategies are not significant for video emotion analysis.

	Binary		5-Class	Regression
Model	Acc(%)	F-1	Acc(%)	MAE
V	69.30	68.67	35.58	1.15
A	61.39	61.25	26.28	1.26
T	74.88	74.14	39.07	0.99
V+T	75.58	75.53	39.53	0.95
V+A	70.23	70.16	36.74	1.12
A+T	74.65	73.95	39.30	0.99
A+V+T(CF)	76.28	76.31	40.23	0.94
A+V+T(TF)	76.74	75.38	39.77	0.93
A+V+T(WF)	**78.14**	**77.44**	**41.16**	**0.91**

Table 4. Results of some derived models and the influence of fusion strategies. "V" is the visual sub-channel, "A" is the acoustic sub-channel and "T" is the textual sub-channel.

6. Conclusions and Limitations

This paper adapts a novel deep temporal convolutional neural network for multimodal sentiment analysis. By introducing well-designed deep architecture and temporal convolution, the proposed architecture can greatly benefit multimodal sentiment analysis tasks. By combining these architectures and a dynamic weighting strategy, the proposed end-to-end deep neural network achieves the best performance in a real-world database compared with other popular models.

However, due to the size of CMU-MOSI database, the proposed model does not display its full capability, and also does not expose its potential problems. More experiments on large-scale datasets, such as the new released CMU-MOSEI dataset [31], are necessary to further revise the proposed method.

References

[1] L. Anthony, Y. Kim, and L. Findlater. Analyzing user-generated youtube videos to understand touchscreen use by people with motor impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1223–1232. ACM, 2013. 1

[2] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 2, 3

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016. 2, 7

[4] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramaanyam. Benchmarking multimodal sentiment analysis. *arXiv preprint arXiv:1707.09538*, 2017. 1, 2

[5] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2017. 3

[6] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. Covarepa collaborative voice analysis repository for speech

technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964, 2014. 2, 7

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255, 2009. 6

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941, 2016. 5

[9] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1019–1027, 2016. 5

[10] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 7

[11] 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 (CVPR)*, pages 770–778, 2016. 2, 3

[12] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. 1

[13] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017. 3

[14] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285, 2016. 3

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012. 3

[17] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*, 2018. 1

[18] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces (ICMI)*, pages 169–176, 2011. 1

[19] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur. Low latency acoustic modeling using temporal convolution and lstms. *IEEE Signal Processing Letters*, 25(3):373–377, 2018. 3

[20] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. 1, 2

[21] S. Poria, E. Cambria, and A. Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)*, pages 2539–2544, 2015. 1

[22] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, volume 1, pages 873–883, 2017. 1, 6

[23] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision (ECCV)*, pages 338–353. Springer, 2016. 6

[24] V. P. Rosas, R. Mihalcea, and L.-P. Morency. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, 2013. 1

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[26] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016. 3

[27] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*, 2016. 6

[28] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 1, 6, 8

[30] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 2, 6

[31] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, volume 1, pages 2236–2246, 2018. 8