Multimodal Emotion Recognition with Factorized Bilinear Pooling and Adversarial Learning

Haotian Miao*
Computer Science and Engineering School, Northeastern University, Shenyang, China
huhumht@gmail.com

Yifei Zhang
Computer Science and Engineering School, Northeastern University, Shenyang, China
zhangyifei@cse.neu.edu.cn

Daling Wang
Computer Science and Engineering School, Northeastern University, Shenyang, China
wangdaling@cse.neu.edu.cn

Shi Feng
Computer Science and Engineering School, Northeastern University, Shenyang, China
fengshi@cse.neu.edu.cn

ABSTRACT

With the fast development of social networks, the massive growth of the number of multimodal data such as images and texts allows people have higher demands for information processing from an emotional perspective. Emotion recognition requires a higher ability for the computer to simulate high-level visual perception understanding. However, existing methods often focus on the single-modality investigation. In this work, we propose a multimodal model based on factorized bilinear pooling (FBP) and adversarial learning for emotion recognition. In our model, a multimodal feature fusion network is proposed to encode the inter-modality features under the guidance of the FBP to help the visual and textual feature representation learn from each other interactively. Beyond that, we propose an adversarial network by introducing two discriminative classification tasks, emotion recognition and multimodal fusion prediction. Our entire method can be implemented end-to-end by using a deep neural network framework. Experimental results indicate that our proposed model achieves competitive performance on the extended FI dataset. Progressive results prove the ability of our model for emotion recognition against other single- and multi-modality works respectively.

CCS CONCEPTS

• **Information systems**; • **Information retrieval**; • **Retrieval tasks and goals**; • **Sentiment analysis**;

KEYWORDS

Multimodal emotion recognition, Factorized bilinear pooling, Adversarial learning

ACM Reference Format:
Haotian Miao*, Yifei Zhang, Daling Wang, and Shi Feng. 2021. Multimodal Emotion Recognition with Factorized Bilinear Pooling and Adversarial Learning. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487164

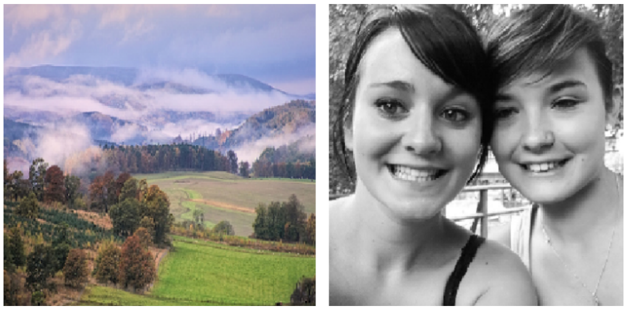

Figure 1: Two Example Images with Textual Descriptions. The Left Image Has a Textual Description: "I Love Here. Might Want a Stop and Enjoy the Beauty of Nature!" The Textual Description for the Right Image Is: "Happy to See You on Here! Hope Things Are Going Good for You My Friend".

1 INTRODUCTION

Emotion recognition plays a crucial role in natural human-machine interactions, which benefits a wide variety of applications [1]. People often would like to share their emotional state to others by visual and textual content in social networks [2]. Emotion is the abstract high-level semantic information that contains lots of subjectivity [3]. Therefore, it is necessary to analyze the multimedia data on social networks for emotion recognition.

Figure 1 shows two examples of how text and visual contents can carry different information and enrich each other. In the left figure, apparent affective mood "love" and "enjoy" in the textual description cannot be captured from the image directly, nor can the text draw out the beautiful and comfortable environment of the picture in detail, which shows the complementarity between them. In the right figure, "happy" and "good" in the text can reflect users' emotion more than other words, and smiley expression in the image region contains more emotional information, which shows their bidirectional semantic relevance.

Many early works have mostly focused on emotion classification considering only the single-modality feature. They try to predict

CSAE 2021, October 19–21, 2021, Sanya, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487164

the aroused human moods when given a particular piece of visual or textual content [4, 5]. For visual feature extraction, previous works often use manually crafted methods that rely on various pixel-level feature representations such as color histograms [6] and texture descriptors [7]. In recent years, many deep neural network models, especially CNNs, have achieved state-of-the-art performance on many computer vision related work [8, 9].

Multimodal affective analysis according to the images and the texts on the internet has been receiving increasing attention and related investigations have proved their effectiveness [10]. For multimodal contents analysis, combination or fusion strategies are commonly used, including feature fusion [11] and decision fusion [12].

Motivated by all of this, in this paper, we propose a multimodal emotion recognition model based on factorized bilinear pooling and adversarial learning. The main contributions of our work are as follows:

- We propose an end-to-end trainable framework that contains two modules, factorized bilinear pooling module (FBP) and adversarial learning module (AL) for multimodal emotion recognition.
- In our model, FBP captures complex interactions between visual and textual features to achieve effective multimodal feature fusion. AL plays a minimax game by introducing an adversarial loss between emotion recognition and multimodal feature fusion to obtain the discriminative representations for multimodal emotion recognition.
- We evaluate our method on the extended FI datasets, and our proposed method achieves superior performance over previous approaches in the task of emotion recognition.

The remainder of this work is organized as follows: in Section 2, we discuss previous relevant works on emotion recognition. Section 3 details our proposed multimodal emotion recognition model based on FBP and AL. In Section 4, we describe the datasets and implementation details in our experiments and then discuss the experimental results. Finally, in Section 5, we conclude our work.

2 RELATED WORKS

Emotion recognition is an important task that has many applications, including movie box-office prediction, product evaluation and political election prediction. In this section, we discuss some related work of visual textual and multimodal methods.

Before the popularity of CNN, most visual emotion recognition studies has been dominated by traditional methods using manually crafted features such as SIFT [13] or shallow classifiers such as local binary patterns (LBPs) [14]. In the last few years, deep neural networks have contributed a lot to performance improvements concerning other popular learning algorithms, such as SVM [15]. Beyond that, a classical method is to build visual sentiment ontology [16] by detecting adjective–noun pairs. You et al. [17] produce a automatically attributes detection model based on attention mechanism for image sentiment analysis.

Correspondingly, LSTM [18, 19] is widely utilized for many textual emotion recognition tasks, such as Twitter [20]. Alsaeedi and Zubair [21] analyze different kinds of sentiment analysis that is applied on to Twitter dataset and provide a summary.

Researches have shown that analyzing affective expression in a multimodal way is effective [22]. Taking both textual and visual based approaches into account, Luo et al. [23] propose a binary multimodal sentiment regression modal and find that the performance can be boosted by jointly considering the inter-modality contents. Huang et al. [24] propose an image-text attentive fusion model for sentiment analysis. Ji et al. [25] proposed a novel bi-layer multimodal hypergraph learning (Bi-MHG) for robust sentiment prediction of multimodal tweets.

However, these previous works have some limitations. Most methods focus on image-text sentiment analysis, which only consider binary labels (positive and negative) and that is not adequate to bridge the complex affective semantic gap between humans and machines. Efforts on image-text emotion recognition have been limited by the relative paucity of the datasets. Corchs et al. [26] make the first attempt to adopt multimodal data for a fine-grained emotion classification task within eight emotional categories , present some ensemble approaches that combine five classifiers for multimodal eight emotion categories prediction. Xu et al [27] propose a hierarchical deep fusion model to combine visual content with textual by their links. Inspired by their work, we crawl titles and image descriptions as textual information to combine with visual content for multimodal emotion recognition.

Besides that, past multimodal approaches based on feature fusion often concatenate features of different modality as the input for the prediction models. There is a lack of adequately capturing complex association between image and text for their integrated features since the feature distributions of different modalities vary dramatically.

Inspired by the success of the application of bilinear pooling in visual question answering [27] and generative adversarial learning in computer vision [28], in this paper, we utilize FBP to fuse intermodality features and design an adversarial network for multimodal emotion recognition.

3 PROPOSED MODEL

In this section, we propose a multimodal emotion recognition model assisted by factorized bilinear pooling and adversarial learning. The overall architecture of our framework is shown in Figure 2. Both visual and textual information are concerned in our model, where the interactions between them are conducted based on factorized bilinear pooling (FBP). Beyond that, pairwise multimodal inputs that have positive and negative sentiment are fed into the generator to generate multiple features. We define two classifiers – the emotion classifier (EC) and the multimodal fusion classifier (MFC). Therefore, we attempt to train the adversarial networks with the emotion classification loss and the multimodal fusion classification loss as the adversarial loss. In the following sections, we will provide a detailed illustration of our model.

3.1 Multimodal Feature Fusion with Factorized Bilinear Pooling

Emotion recognition profits from multimodal contents as they can provide more vivid and adequate information. Formally, let I be the image and T be the text. For the image, the visual feature$X \in \mathbb{R}^m$is extracted by pretrained Residual Network (ResNet) [29]. For the

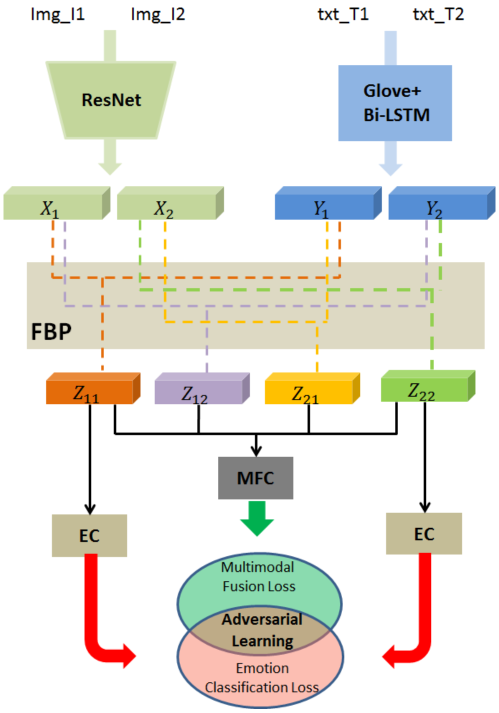

Figure 2: Illustration of Our Proposed Framework.

text, it is first tokenized into a word vector via the GloVe word embedding [30] model pretrained on large corpus, and then passed through a bi-LSTM [19] to generate the textual feature$Y \in \mathbb{R}^n$.The simplest multimodal bilinear pooling model is defined as follows:

$$z_i = X^T W_i Y \tag{1}$$

where $W_i \in \mathbb{R}^{m \times n}$ is a projection matrix and $z_i \in \mathbb{R}$ is the output of the bilinear pooling model. To obtain the output$Z = [z_1, \cdot s, z_o] \in \mathbb{R}^o$, $W = [W_1, \cdot s, W_o] \in \mathbb{R}^{m \times n \times o}$ is needed to be learned. Although bilinear pooling can availably catch the pairwise interactions between the multimodal features, it also increases a mass of parameters which result in a high computational cost and a risk of overfitting [31].

Inspired by the matrix factorization [32], the projection matrix W_i in Eq. 1) can be factorized into two low-rank matrices:

$$\begin{aligned} z_i &= X^T U_i V_i^T Y \\ &= \sum_{d=1}^{k} X^T u_d v_d^T Y \\ &= \mathcal{I}^T \left(U_i^T X \otimes \ V_i^T Y \right) \end{aligned} \tag{2}$$

where k is the dimension of the factorized matrices $U_i = [u_1, \cdot s, u_k] \in \mathbb{R}^{m \times k}$ and$V_i = [v_1, \cdot s, v_k] \in \mathbb{R}^{n \times k}$, $\otimes$ denotes the element-wise product of two vectors, and $\in \mathbb{R}^k$ is an all-1 vector.

To obtain the output feature vector Z by Eq. 2), two 3-D low-rank matrices U = $[U_1, \cdot s, U_o] \in \mathbb{R}^{m \times k \times o}$ and V = $[V_1, \cdot s, V_o] \in \mathbb{R}^{n \times k \times o}$ would be learned. The 3-D matrices U and V can be reformulated as 2-D matrices $\hat{U} \in \mathbb{R}^{m \times ko}$ and $\hat{V} \in \mathbb{R}^{n \times ko}$ respectively. Therefore, the multimodal fusion feature Z can be rewritten as follows:

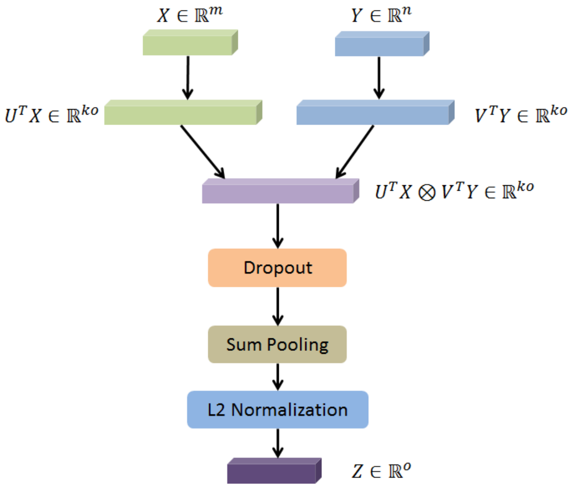
$X \in \mathbb{R}^m$
$Y \in \mathbb{R}^n$
$U^T X \in \mathbb{R}^{ko}$
$V^T Y \in \mathbb{R}^{ko}$
$U^T X \otimes V^T Y \in \mathbb{R}^{ko}$
Dropout
Sum Pooling
L2 Normalization
$Z \in \mathbb{R}^o$

Figure 3: Factorized Bilinear Pooling (FBP) for Multimodal Feature Fusion.

$$Z = SumPooling \left(\hat{U}^T X \otimes \hat{V}^T Y, k \right) \tag{3}$$

where the function SumPooling(x, k) performs sum pooling with a 1-D non-overlapped window with the size k over x.

The factorized bilinear pooling module (FBP) is shown in Figure 3. A dropout layer is used for preventing over-fitting. After L2-normalization, we obtain the multimodal fusion feature Z as follows:

$$Z = \frac{Z}{\|Z\|} \tag{4}$$

3.2 Adversarial Learning for Multimodal Emotion Recognition

Assume $M_1 = \{[I_1^i, T_1^i]\}_{i=1}^{N_1}$ and $M_2 = \{[I_2^i, T_2^i]\}_{i=1}^{N_2}$ are two subsets of the multimodal emotion dataset with positive and negative sentiment labels respectively, where N is the dataset size. We omit the subscript i for notation simplicity. First, we extract visual features X_1, X_2 from I_1, I_2 and textual features Y_1, Y_2 from T_1, T_2. Then, we fed these features into FBP and generate inter-modality pairwise fusion features Z_{11}, Z_{12}, Z_{21}, and Z_{22}. That is, the feature generator (G) transforms the single-modality features X and Y to the multimodal feature Z. We define two classifiers – the emotion classifier (EC) and the multimodal fusion classifier (MFC). EC uses Z_{11} and Z_{22} for emotion classification, and MFC uses Z_{11}, Z_{12}, Z_{21}, and Z_{22} for multimodal fusion classification.

EC is the classifier to predict emotion probability $\hat{I}_s$. The emotion classification is optimized by the cross-entropy loss as follows:

$$L_{EC} = -\sum\nolimits_{s=1}^{S} I_s \log \left(\hat{I}_s \right) \tag{5}$$

where the label I_s is either from M_1 or M_2 , and S is the number of the emotional classes.

The multimodal fusion category label t $\in$ {0, 1, 2} is self-supervised. For Z_{11} the label t is 0, and for Z_{12} and Z_{21}, t is 1, and for Z_{22}, t is 2. This requires no extra manual annotation. We take another classifier MFC to predict the probability $\hat{t}$. We then compute the loss function as:

$$L_{MFC} = -\sum\nolimits_{r=1}^{3} t_r \log\left(\hat{t}_r\right) \tag{6}$$

where $\hat{t}_r$ is one of the predictions for Z_{11}, Z_{12}, Z_{21}, and Z_{22}, and t_r is the corresponding label.

Since adversarial learning (AL) runs as a minimax game that the targets at beating each other. To unite two objective functions together, we flip the sign of L_{MFC} . The final objective loss function is as follows with a hyper-parameter $\alpha \in [0,\ 1]$ to balance the two losses:

$$L = L_{EC} - \alpha L_{MFC} \tag{7}$$

4 EXPERIMENTS AND RESULTS

In this section, we conduct experiments to evaluate the performance of our multimodal emotion recognition model with FBP and adversarial learning. First, we introduce the datasets and the experimental settings in our work. Then the superiority of our proposed method is discussed against baselines. Finally, we perform several ablation studies.

4.1 Datasets and Experimental Settings

We conduct experiments on a large-scale emotion dataset FI [33], where images are manually divided into eight emotion categories, i.e., amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. We downloaded more than 22,000 images that receive at least three votes from their assigned 5 Amazon Mechanical Turk workers. Each category consists of more than 1,000 images. We enrich the FI dataset by crawling textual descriptions using the image url provided in the dataset. Besides, we use two subsets of the extended FI dataset with positive (i.e., amusement, excitement, awe, contentment) and negative (i.e., disgust, anger, fear, sadness) sentiment labels for adversarial learning. Since the size of the negative subset (more than 6,000) is much smaller than the positive subset (more than 15,000), we upsample the negative subset during training. Data partition used in our work is 80% and 20% for training and testing, and a 5-fold cross-validation strategy is adopted to help learning optimal model and parameters.

ResNet101 is utilized for our visual feature extraction. The input raw images are resized to 224 × 224, and the output feature dimension is 2048 after the *pool*5 layer (with $L2$-normalization). The texts are first tokenized into words with a maximum of 30 and further transformed into feature vectors using 100-D GloVe word embedding. Then, the feature vectors are sent into a bi-LSTM with 1024 hidden units to form 2048-D feature vectors. For network optimization, SGD is used as the optimizer. The base learning rate is 10^{-4} and reduced by a factor of 10 every 10 epochs. The momentum, weight decay, dropout ratio, total epoch, and batch size are set to 0.9, 10^{-5}, 0.1, 100, and 32 respectively. Our networks are implemented based on PyTorch.

4.2 Performance Evaluation

In this section, we first present a comparative evaluation on the extended FI dataset to demonstrate the effectiveness of our method. Then we evaluate the performance of our model in the 8-category emotion prediction task on the extended FI dataset.

Table 1: Comparison on the Extended FI Dataset for Emotion Recognition Performance

Algorithms	ACC (%)	MAP (%)
Visual (Fine-tuned-CNN)	58.3	50.4
Visual (Hand-crafted + RandomSubSpace DT)	51.9	42.5
Visual (deep feature + SVM)	57.3	49.9
Text (SVM)	71.0	65.5
Multimodal (hand-crafted + bagging SVM)	64.8	59.7
Multimodal (deep feature + Late-BMA)	76.0	69.6
Multimodal FBP-AL (Ours)	**80.7**	**76.3**

Table 1 shows the performance of our method on the extended FI dataset along with different results of others. You et al. [33] present a Fine-tuned-CNN method on their FI image dataset. Corchs et al. [26] adopt ensemble learning approaches on single- and multi-modality dataset, i.e., hand-crafted feature with RandomSubSpace DT and deep feature with SVM for the visual emotion recognition, textual feature with SVM, and hand-crafted feature with bagging SVM and deep feature with Late-BMA for the multimodal prediction. The experimental results in Table 1 reveal that our proposed multimodal FBP-AL model outperforms other methods. Compared with single-modality prediction, our method outperforms both the best visual result (Fine-tuned-CNN) and the result on textual data, by a large margin. This proves that emotion recognition could profit from multimodal contents as they provide more vivid and adequate information. Beyond that, compared with the multimodal approaches, our method gets at least 4.7% higher accuracy (ACC) and 6.7% higher macro average precision (MAP). The competitive result demonstrates the effectiveness of our multimodal BFP-AL model in emotion recognition.

Figure 4 demonstrates the normalized confusion matrix of 8-category emotion on the extended FI datasets to analyze the performance of our proposed method in each emotion category. The first figure is the baseline confusion matrix (late-BMA with deep features) and the second is ours. They both contain eight emotion categories, i.e., amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. For our confusion matrix in the second figure, results of Amusement, Awe, Contentment, and Disgust categories are competitive, while Anger and Fear categories have lower accuracy than others. A large number of Anger and Fear categories are misclassified as Sadness categories. Besides that, Excitement category can be easily mistaken for Amusement category, which may be attributed to the similar emotional features of the two categories. Compared with the baseline, our method has an improvement in Awe, Disgust, Anger and Fear categories, while similar in the rest.

4.3 Ablation Studies

In this section, we present a comparative evaluation on different modules in our method. We replace FBP with a concatenate fusion (Concat) and replace AL with the fully connected layers (FC) respectively. Figure 5 summarizes the results in different methods on the

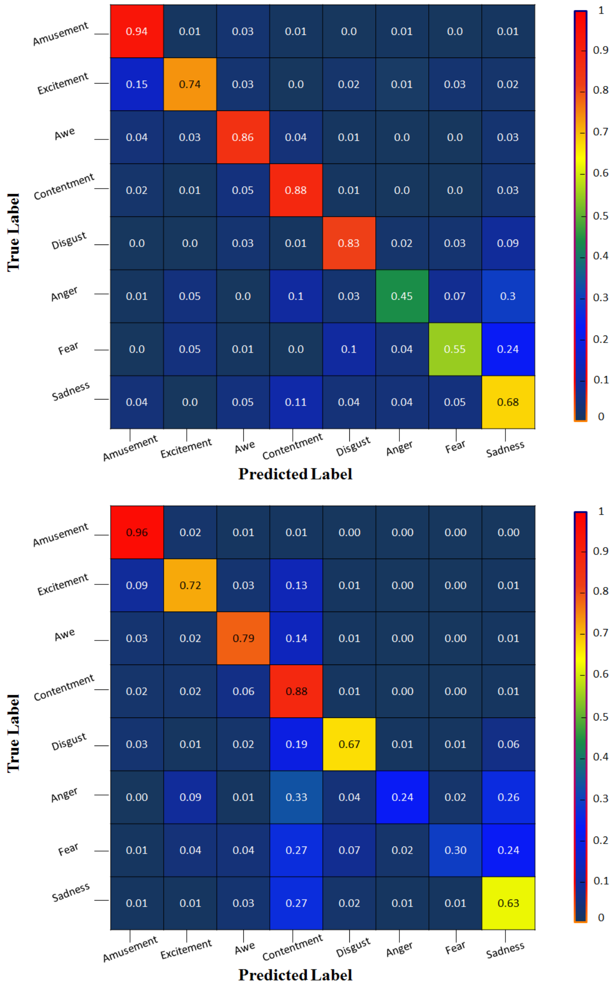

Figure 4: Normalized Confusion Matrix of 8-Category Emotion on the Extended FI Datasets.

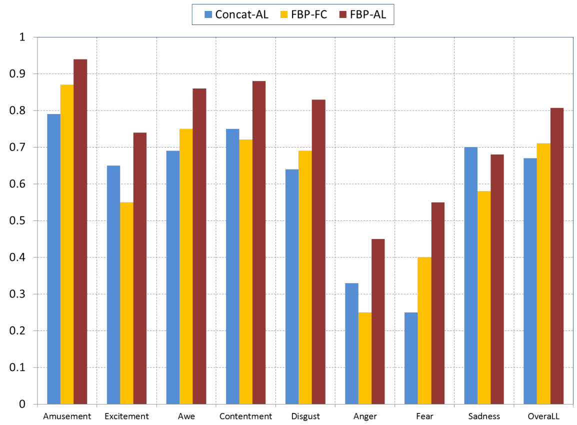

Figure 5: Comparison in Different Methods on Extended FI Dataset.

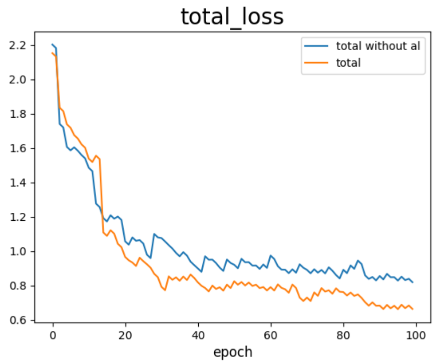

Figure 6: The Trend of Total Loss and Total Loss without Adversarial Loss during the Training Process.

extended FI dataset. It is shown that FBP-AL overcome Concat-FC and FBP-AL in total accuracy and most emotion categories, like Awe, Contentment, Disgust and Anger.

For multimodal emotion recognition, both FBP and AL in our model can boost the overall performance, and also performs well in enhancing the prediction of most categories.

Figure 6 is a visualization of our loss trend during the training process. We can see that the loss of our model decreases more slowly than without adversarial loss (multimodal fusion loss), but converges to a smaller value. It proves that adversarial loss affects our model as a directional guide and helps improve the performance of multimodal feature fusion and emotion classification.

5 CONCLUSIONS

In this paper, we propose a multimodal emotion recognition model based on factorized bilinear pooling and adversarial learning. We first extract the visual features from the images and the textual features from the descriptions. Then, we send these features into FBP to capture more complex interactions and achieve more effective fusion. Beyond that, we take the classification loss of different sentiment and inter-modality combination as multimodal fusion loss. Through the adversarial competition of the emotion recognition loss and multimodal fusion loss, the modules play a minimax game by beating each other and the discriminative representations are generated by AL for emotion recognition. Experimental results on the extended FI datasets demonstrate the performance of our proposed method for multimodal emotion recognition.

ACKNOWLEDGMENTS

The work was supported by the National Key R&D Program of China under grant 2018YFB1004700, and National Natural Science Foundation of China (61772122, 61872074).

REFERENCES

[1] Zhang D, Wu L, Sun C, *et al.* (2019). Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations [C]. International Joint Conference on Artificial Intelligence (IJCAI), 415-5421.

[2] Zhao S, Yao H, Gao Y, *et al.* (2016). Continuous probability distribution prediction of image emotions via multi-task shared sparse regression [J]. IEEE Transactions on Multimedia, 19(3), 532-645.

[3] Datta R, Li J and Wang J Z (2008). Algorithmic inferencing of aesthetics and emotion in natural images: an exposition [C]. IEEE International Conference on Image Processing (ICIP), 105-108.

[4] Lee J, Kim Seungryong, Kim Sunok, Park J and Sohn K (2019). Context-Aware Emotion Recognition Networks [C]. IEEE/CVF International Conference on Computer Vision (ICCV), 10142-10151.

[5] Wilson T, Wiebe J and Zubair M (2005). Recognizing contextual polarity in phrase-level sentiment analysis [C]. HLTEMNLP, 347-354.

[6] Shan C, Gong S and McOwan P W (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study [J]. Image and Vision Computing, 27(6), 803-816.

[7] Zhao G and Pietikainen M (2007). Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), 915-928.

[8] Liu F, Zhou X, Yan X, *et al.* (2021). Image Steganalysis via Diverse Filters and Squeeze-and-Excitation Convolutional Neural Network [J]. Mathematics, 9(2), 189.

[9] Darabant A S, Borza D and Danescu R (2021). Recognizing Human Races through Machine Learning-A Multi-Network, Multi-Features Study [J]. Mathematics, 9(2), 195.

[10] Fersini E, Pozzi FA and Messina E (2017). Approval network: a novel approach for sentiment analysis in social networks [J]. World Wide Web, 20(4), 831-854.

[11] Morency L, Mihalcea R and Doshi P (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web [C]. ICMI 169-176.

[12] Wöllmer M, Weninger F, Knaup T, *et al.* (2013). YouTube movie reviews: Sentiment analysis in an audio-visual context [J]. IEEE Intell. Syst. 28(3) 46-53.

[13] Zhong L, Liu Q, Yang P, *et al.* (2015). Learning Multiscale Active Facial Patches for Expression Analysis [J]. IEEE Transactions on Cybernetics, 45(8), 1499-1510.

[14] Zhao G and Pietikainen M (2007). Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), 915-928.

[15] Yu Z and Zhang C (2015). Image based Static Facial Expression Recognition with Multiple Deep Network Learning [C]. International Conference on Multimodal Interaction (ICMI), 435-442.

[16] Borth D, Ji R, Chen T *et al.* (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs [C]. ACM MM 223-232.

[17] You Q, Jin H and Luo J (2017). Visual sentiment analysis by attending on local image regions [C]. AAAI 231-237.

[18] Sahni T, Chandak C, Chedeti N R and Singh M (2017). Efficient Twitter sentiment classification using subjective distant supervision [C]. International Conference on Communication Systems and Networks (COMSNETS), 548-553.

[19] Sainath T N, Vinyals O, Senior A W and Sak H (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks [C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4580-4584.

[20] Nakov P, Ritter A, Rosenthal S, *et al.* (2016). SemEval-2016 Task 4: Sentiment Analysis in Twitter. International Workshop on Semantic Evaluation (SemEval@NAACL-HLT), 1-18.

[21] Alsaeedi A and Zubair M (2019). A study on sentiment analysis techniques of Twitter data [J]. Int. J. Adv. Comput. Sci. Appl. 10(2), 361-374.

[22] Niu T, Zhu S, Pang L and El-Saddik A (2016). Sentiment analysis on multi-view social data [C]. International conference on MultiMedia Modeling, 15–27.

[23] You Q, Luo J, Jin H and Yang J (2016). Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia [C]. ACM international conference on web search and data mining, 13–22.

[24] Huang F, Zhang X, Zhao Z, *et al.* (2019). Image–text sentiment analysis via deep multimodal attentive fusion [J]. Knowl. Based Syst. 167,26-37.

[25] Ji R, Chen F, Cao L and Gao Y (2019). Cross-Modality Microblog Sentiment Prediction via Bi-Layer Multimodal Hypergraph Learning [J]. IEEE Trans. Multim. 21(4), 1062-1075.

[26] Corchs S, Fersini E and Gasparini F (2019). Ensemble learning on visual and textual data for social image emotion classification [J]. International Journal of Machine Learning and Cybernetics, 10(8), 2057-2070.

[27] Xu J, Huang F, Zhang X, *et al.* (2019). Sentiment analysis of social images via hierarchical deep fusion of content and links [J]. Appl. Soft Comput. 80, 387-399.

[28] Dai P, Ji R, Wang H, *et al.* (2018). Cross-Modality Person Re-Identification with Generative Adversarial Training [C]. International Joint Conference on Artificial Intelligence, 677-683.

[29] He K, Zhang X, Ren S and Sun J (2016). Deep Residual Learning for Image Recognition [C]. IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[30] Pennington J, Socher R and Manning C D (2014). Glove: Global vectors for word representation [C]. Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[31] Yu Zhou, Yu Jun, Fan J and Tao D (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering [C]. IEEE International Conference on Computer Vision, 1839-1848.

[32] Rendle S (2010). Factorization machines [C]. IEEE International Conference on Data Mining (ICDM), 995–1000.

[33] You Q, Luo J, Jin H and Yang J (2016). Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and the Benchmark [C]. AAAI Conference on Artificial Intelligence, 308-314.