Available online at https://www.cnki.net

June 2022, 29(3): 25 – 33

The Journal of China Universities of Posts and Telecommunications

https://jcupt.bupt.edu.cn

Multi-level fusion with deep neural networks for multimodal sentiment classification

Zhang Guangwei[1], Zhao Bing[2], Li Ruifan[3] (✉)

1. School of Computer Sciences, Beijing University of Posts and Communications, Beijing 100876, China
2. School of Science, Yanshan University, Qinhuangdao 066004, China
3. School of Artificial Intelligence, Beijing University of Posts and Communications, Beijing 100876, China

Abstract

The task of multimodal sentiment classification aims to associate multimodal information, such as images and texts with appropriate sentiment polarities. There are various levels that can affect human sentiment in visual and textual modalities. However, most existing methods treat various levels of features independently without having effective method for feature fusion. In this paper, we propose a multi-level fusion classification (MFC) model to predict the sentiment polarity based on the fusing features from different levels by exploiting the dependency among them. The proposed architecture leverages convolutional neural networks (CNNs) with multiple layers to extract levels of features in image and text modalities. Considering the dependencies within the low-level and high-level features, a bi-directional (Bi) recurrent neural network (RNN) is adopted to integrate the learned features from different layers in CNNs. In addition, a conflict detection module is incorporated to address the conflict between modalities. Experiments on the Flickr dataset demonstrate that the MFC method achieves comparable performance compared with strong baseline methods.

Keywords multimodal fusion, sentiment analysis, deep learning

1 Introduction

Online social networks provide multiple forms of being available to their users. For instance, people can post a tweet attached with images or videos. Social networks play a more and more important role in our daily life for acquiring information and sharing experiences[1]. Meanwhile, online users love to express their opinions on subjects they are interested in, because of the free expression of speech on social networks. Sentiment analysis on online user generated data in social networks can be helpful to understand users' behavior and to improve applications[2].

Among the large amount of data, we are particularly interested in analyzing sentiment of tweets containing both texts and images towards specific events and topics. Currently, most of the work on emotion analysis are conducted in the field of natural language processing. Therefore, the main datasets and resources available are limited to text-based opinion mining. However, with the development of social media and mobile networks, people begin to widely use multimodal information to express their opinions on

Received date: 28-02-2022

Corresponding author: Li Ruifan, E-mail: rfli@bupt.edu.cn

DOI: 10.19682/j.cnki.1005-8885.2022.1013

social media platforms, such as video, pictures, and audio. Interestingly, statistics have shown that the use of pictures in a tweet can increase the click rate, forwarding rate and collection rate of the tweet[3], which further encourages users to publish more visual content.

In recent years, deep neural networks have achieved remarkable performance in various fields, especially in compute vision[4-5] and natural language processing[6]. Inspired by the enormous success of deep learning, researches on sentiment analysis have applied deep learning algorithms. However, most of these mainly focused on one single modality of user content instead of the closely-related modalities. In fact, a large number of posted images do not contain any sentiment words in text at all; or the text sentiment is obvious but the image sentiment is inconspicuous.

In this paper, we focus on the task of sentiment prediction using the joint textual and visual information of online posts. Deep neural networks employed in previous work[7-10] have been shown effective in solving the tasks of image or text sentiment analysis. In addition, the fusion of multimodal data cannot provide additional information with an increase in overall accuracy[11]. Therefore, to solve the challenging problem, we propose a feature fusion method based on multiple neural networks. Moreover, to further improve MFC method, we design a conflict detection module to extend our model, i. e., rectified multi-level fusion classification (R-MFC). Experimental results demonstrate the effectiveness of the proposed method for the task of joint vision and text sentiment analysis.

2 Related work

Previous researches on feature fusion in the area of multimodal sentiment analysis can be roughly grouped into several categories. Feature-level fusion[12-15] fuses the features extracted from various modalities such as visual features, text features, audio features, as a general feature vector and the combined features are sent for analysis. Decision-level fusion[16-19] fuses the features of each modality that are examined and classified independently and the results are fused as a decision vector to obtain the final decision. Hybrid multimodal fusion[20-21] is the combination of both feature-level and decision-level fusion methods. Model-level fusion[22-23] is a technique that uses the correlation between data observed under different modalities, with a relaxed fusion of the data. Classification-based fusion[24-25] uses a range of classification algorithms to classify the multimodal information into pre-defined classes. In the rule-based fusion methods[26-27], multimodal information is fused by statistical rule-based methods such as linear weighted fusion, majority voting and custom-defined rules. Estimation-based fusion methods[28-29] are usually employed to estimate the state of moving object using multimodal information, especially audio and video.

Previous work only utilized the high-level features for fusion, such as using the last outputs of different modal model layers[17]. Due to the complexity of emotions and the differences between text sentiment and visual sentiment, more and more work were proposed to exploit multiple levels features fusion. CNN have achieved good results in many areas, including visual sentiment[10, 30] and textual sentiment[31]. RNNs can effectively model the long-term dependency in sequential data. Rao et al. [32] utilized three different networks to capture different levels of visual features with high expense of parameters. Inspired by this work, we try to explicitly exploit the dependency between low-level features and high-level features to improve the sentiment classification model.

3 MFC model

In this paper, MFC model combining CNN and RNN is proposed. In addition, the rectification of conflict detection module R-MFC is incorporated to improve the model. In the dataset, the model and its deformation are compared with the popular cross-modal emotion classification model of image and text. The results show that the model and its variation have high accuracy.

As shown in Fig. 1, the proposed MFC mainly consists of three components: image feature extractor, text feature extractor and bi-directional gated recurrent unit (Bi-GRU) feature fusion. The input image and text are first fed to the two different CNN models to extract multiple levels of features at different branches. We expect that those features from different layers represent different parts of modals, such as line, color, texture, and object in image model, word, subject in text model, which characterize different levels of features from the local view to global view. Bi-GRU fusion aims to integrate the different levels of features by exploiting the dependency between low-level and high-level features. The integrated features from Bi-GRU model are concatenated for textual and visual sentiment classification.

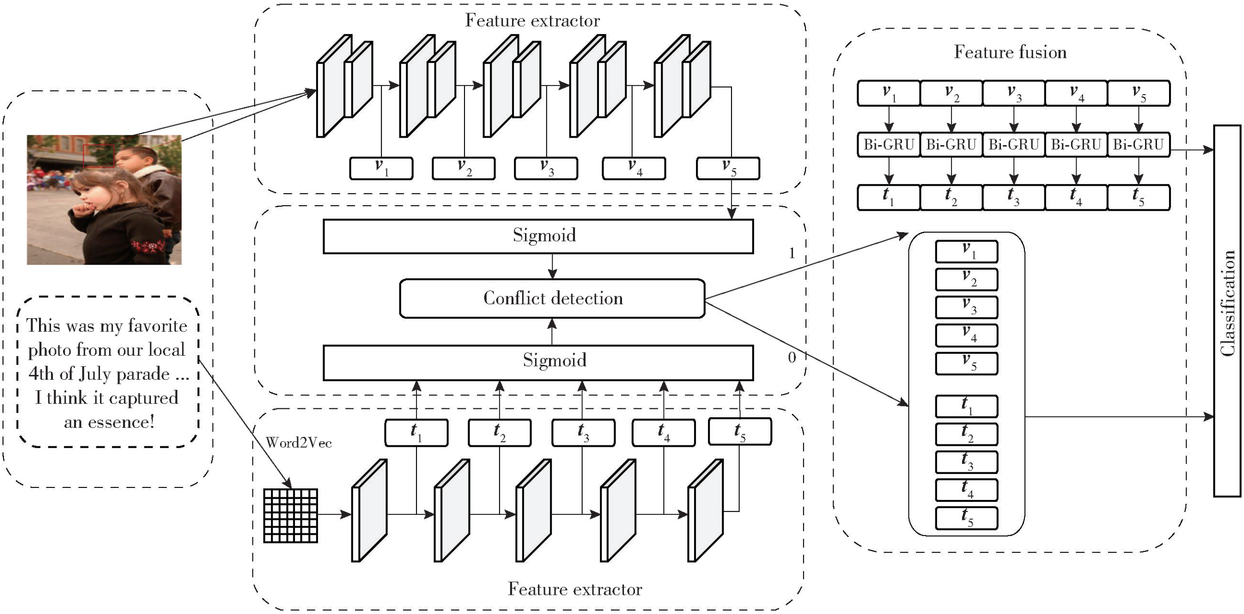

Fig. 1 Structure of the proposed MFC model

3.1 Image CNN with multiple branches

CNN is a structure of deep neural network, which can be regarded as a locally connected network. Compared with fully-connected network, its significant characteristics are local connectivity and weight sharing. The image CNN in this paper adds multiple intermediate branches. In a CNN, as the number of layers increases, the activation receptive field of the pooling layer of different convolutional (Conv) layers becomes more meaningful and easier to be recognized by humans. Thus, the proposed multi-branch image CNN adopts a multi-branch design to obtain different levels of image features from low to high when extracting image features, not just the overall image features for the next feature fusion.

Compared with the top layer, the bottom layer and the intermediate layer can provide complementary information, including low-level features and middle-level features. The image CNN used to extract image features in this paper has multiple branches, which are used to extract a variety of different levels of image features. The pre-processed image is sent to the pre-trained image CNN to obtain image features $\{\boldsymbol{v}_m\}_{m=1}^{M}$, where M represents the number of branches and m is the branch index. As shown in Fig. 1, we use five branches to learn different levels of features from the local view to global view. The detailed architecture of image CNN is summarized in Table 1. We incorporate different branches from each pooling layer by inserting 1 × 1 convolution layer and fully connected layer. The 1 × 1 convolution layer with 128 filters is used for dimension reduction and rectified linear activation. Those five branches from the pooling layers and different levels of features extracted from multiple branches will be fed into the Bi-GRU fusion module. We pre-train the image CNN on the task of image sentiment classification.

Table 1 Details of image CNN

Layer	Channel	Kernel size/Stride	Respective field
Conv 1	64	11 × 11/4	11
Pooling 1	64	3 × 3/2	19
Branch 1	128	1 × 1/1	19
Conv 2	128	5 × 5/1	51
Pooling 2	128	3 × 3/2	67
Branch 2	128	1 × 1/1	67
Conv 3	256	3 × 3/1	99
Branch 3	128	1 × 1/1	99
Conv 4	512	3 × 3/1	131
Branch 4	128	1 × 1/1	131
Conv 5	512	3 × 3/1	163
Pool 5	512	3 × 3/2	195
Branch 5	128	1 × 1/1	195

3.2 Text CNN with multiple branches

In the field of natural language processing, CNNs have been successfully applied. The advantage of CNN is that it can easily use filters to extract local features in sentences, and can reduce the influence of irrelevant words on the final result. The multi-branch CNN can also extract key features in different ranges. And the size of the convolution kernel can be arbitrarily changed as needed, which is more flexible. This paper uses CNN to extract text features, and also uses a multi-branch design.

Specifically, MFC uses the pre-processed dictionary to index the text to obtain $\{\boldsymbol{w}_k\}_{k=1}^{K}$ (k is the word index), and then input these words into the embedding module to convert it into a word vector representation $\{\boldsymbol{e}_k\}_{k=1}^{K}$. Then the word vector representation of each textual description is sent to the text CNN, and the text feature vector $\{\boldsymbol{t}_m\}_{m=1}^{M}$ is obtained. First, we use word embedding method (Word2Vec) to vectorize the input text, then we also use five branches to learn different levels of features from the local view to global view. The detailed architecture of text CNN is summarized in Table 2. We add different branches from each convolutional layer by inserting two pooling layers with 1 × 1 convolution layer and fully-connected layer. The 1 × 1 convolution layer with 128 filters is used for dimension transformation and rectified linear activation. Those five branches from the convolutional layers and different levels of features extracted from multiple branches will also be fed into the Bi-GRU fusion module. We pre-train the text CNN on the task of text sentiment classification.

Table 2 Details of text CNN

Layer	Channel	Kernel size/Stride
Conv 1	64	3 × 3/1
Max-pooling in Branch 1	64	3 × 3/1
Mean-pooling in Branch 1	64	3 × 3/1
Conv 2	128	3 × 3/1
Max-pooling in Branch 2	128	3 × 3/1
Mean-pooling in Branch 2	128	3 × 3/1
Conv 3	256	3 × 3/1
Max-pooling in Branch 3	256	3 × 3/1
Mean-pooling in Branch 3	256	3 × 3/1
Conv 4	512	3 × 3/1
Max-pooling in Branch 4	512	3 × 3/1
Mean-pooling in Branch 4	512	3 × 3/1
Conv 5	512	3 × 3/1
Max-pooling in Branch 5	512	3 × 3/1
Mean-pooling in Branch 5	512	3 × 3/1

3.3 Multiple layer fusion

On the basis of the extracted features of image and text, the information interaction of image and text features is carried out through a Bi-GRU and the features are further extracted. We first use pre-trained image CNN and text CNN to extract different levels features, i. e., image features $\{\boldsymbol{v}_m\}_{m=1}^{M}$ for image, text features $\{\boldsymbol{t}_m\}_{m=1}^{M}$ for text. Here, M represents the branch number, and its default value is 5. Then we concatenate them in sequence, and feed them into Bi-GRU fusion module as the inputs of different time steps. This process is formulated in Eq. (1). We use G to denote the gated recurrent unit (GRU), in which the arrow over $\underline{G}$ denotes the direction.

$$\left.\begin{aligned} \boldsymbol{y}^{\mathrm{L}} &= \overrightarrow{G}([\boldsymbol{v}_m;\boldsymbol{t}_m]_{m=1}^{M}) \\ \boldsymbol{y}^{\mathrm{R}} &= \overleftarrow{G}([\boldsymbol{v}_m;\boldsymbol{t}_m]_{m=1}^{M}) \\ \boldsymbol{y}^{\mathrm{B}} &= [\boldsymbol{y}^{\mathrm{L}};\boldsymbol{y}^{\mathrm{R}}] \end{aligned}\right\} \tag{1}$$

in which, outputs $\boldsymbol{y}^{\mathrm{L}}$ and $\boldsymbol{y}^{\mathrm{R}}$ of Bi-GRU are concatenated with the concatenation operation [· ; ·], obtaining the final representation $\boldsymbol{y}^{\mathrm{B}}$. Then we train the Bi-GRU module with a logistic regression classifier.

3.4 R-MFC module

In popular social networks, due to the randomness and subjectivity of blog posts, there could be sentimental conflicts between modalities. To solve this problem, we incorporate a conflict detection in the MFC model obtaining R-MFC. Specifically, before feature fusion, we firstly calculate the sentimental polarity of a single modality, and then, according to whether the conflict exits we choose different fusion strategies. We suppose that emotional conflicts between visual and textual modalities may be more complex. This situation needs to be addressed by the fusion of complex network.

The rectified conflict detection first uses a Sigmoid activation function to calculate the probability distribution. For the visual feature $\{\boldsymbol{v}_m\}_{m=1}^{M}$, we use $p_{\theta_v}(\cdot)$ under the parameter θ_v. For the textual feature $\{\boldsymbol{t}_m\}_{m=1}^{M}$, we use $p_{\theta_t}(t)$ under the parameter θ_t. Then we choose the fusion strategy according to whether distributions are consistent. This process is given as

$$\left.\begin{array}{ll}\psi(\cdot); & \text{if } (p_{\theta_v}(\cdot)-0.5)(p_{\theta_t}(\cdot)-0.5)>0 \\ \phi(\cdot); & \text{else}\end{array}\right\} \tag{2}$$

where $\psi(\cdot)$ represents the feature fusion using the feature concatenation method, $\phi(\cdot)$ represents the feature fusion using Bi-GRU. Through such a design, the relatively easy situation where the image and text have the same emotional tendency, is assigned to a simple feature splicing combined with a single-layer fully connected neural network. The complex task situation in which images and texts have conflicting sentiment across modalities, is assigned to a complex RNN fusion network.

4 Experiments and results

We evaluate the proposed multi-level fusion neural network on a largest dataset and compare the performance of MFC with four baseline methods.

4.1 Dataset and metric

We adopt the largest datasets which can provide texts and images for evaluation. The details of these datasets are described as follows.

Flickr dataset The Flickr dataset is a part of multilingual visual sentiment ontology (MVSO)[33]. MVSO consists of 15 600 concepts in 12 types of languages that are strongly related to emotions and sentiments expressed in images. All of these concepts are defined in the form of adjective and noun pairs (ANPs), which are crawled from the Flickr website. In our settings, we use only the English dataset in all our experiments. In order to avoid the problem of excessive noisy data and the category imbalance, we choose ANPs with sentiment scores more than 0.16 and less than −0.1. We choose images with description words in the range [1, 100]. In addition, in order to increase the label reliability, we use the probability sampling to further clean the data and obtain the experimental dataset. In total, there are 286 307 positive images and text pairs and 278 003 negative images and text pairs. We then split them with 8∶2 into the training and testing subsets.

Human labeled Flicker dataset The human labeled Flicker dataset is also a part of MVSO. With the help of a task called Human Intelligent Tasks on Amazon Mechanical Turk crowd-sourcing site, a significant amount of manual annotation data is collected. There are 5 categories for each instance, including very positive, positive, neutral, negative, and very negative. Each instance is scored by 5 persons. In order to match the training set in this paper, very positive and positive are combined into one category, and very negative and negative are combined into one category. Regardless of neutrality, the dataset agreed by at least two persons is retained. Finally, we obtain a dataset including 5 744 positive instances and 2 502 negative instances. The dataset is used for testing but not for training.

We adopt the popular evaluation metrics, including Accuracy, Recall, and F1 score for our model evaluation.

4.2 Implementation details

For the textual part, we employ the classic pre-trained Word2Vec model to obtain the distributed word representations. The Word2Vec is pre-trained on the Stanford Twitter Sentiment corpus, and has a fixed size of 200 dimensions. Words not in the pre-training dataset are initialized randomly. The size of embedded word matrix is 100 × 200, due to the maximum length of description is 100 words. For the visual part, the input images are first resized to 224 × 224 before feeding into the image CNN. The mini-batch size of 32 is adopted. In addition, adaptive weight decay (AdamW) optimization is adopted. The MFC model is trained on a workstation with i7-8700 central processing unit (CPU), 16 GB random access memory (RAM), and NVDIA 1060 graphic processing unit (GPU).

4.3 Experimental baselines

To demonstrate the effectiveness of the proposed MFC, we compare it with baseline methods and different variants of MFC. The baseline methods are briefly described as follows.

1) Firstly, two single modality methods, i. e., Image CNN and Text CNN are compared.

2) Cross-modality consistent regression (CCR)[34]. The main idea of consistency regression fusion model is that different modality should express the identical sentiment when describing the same thing.

3) Deep convolutional network (DCN)[35]. The deep neural network is used to extract the features of image and text, and then the splicing operation is carried out before the classifier training the model.

4) Deep fusion convolution (DFC)[36]. The DFC model utilizes the end-to-end architecture with fully convolution network.

5) Deep multimodal attentive fusion (DMAF)[37]. The attention-based feature fusion is combined with a decision level fusion.

6) Attention-based modality-gated network (AMGN)[38]. AMGN is a visual-semantic attention model which designs a modality-gated long short-term memory (LSTM) to learn the multimodal features by adaptively selecting the modality having stronger sentiment information.

7) Image-text interaction graph neural network (ITIGNN)[39]. ITIGNN is a graph neural network for image-text sentiment analysis. A text graph neural network of the text features and a pre-trained CNN of image features are used for image-text interaction graph network.

Furthermore, to verify the contributions of different components in MFC, we design different variants as follows.

1) MFC (Concat). This method simply concatenates five branch outputs from image CNN and text CNN.

2) MFC (GRU). This method puts the five branch outputs of image CNN and text CNN into a uni-directional GRU as the inputs of different time steps.

3) MFC (3 branches). This variant puts the first, the third, and the last branch outputs of image CNN and text CNN to a Bi-GRU as the inputs of different time steps.

4) MFC (4 branches). This variant puts the first, the third, the fourth, and the last branch outputs of image CNN and text CNN to a Bi-GRU as the inputs of different time steps.

5) MFC (Bi-LSTM). This method puts the five branch outputs of image CNN and text CNN to a Bi-LSTM as the inputs of different time steps.

6) MFC. This method puts the five branch outputs of image CNN and text CNN to a Bi-GRU as the inputs of different time steps.

7) R-MFC. This method first performs the sentiment classification of image and text independently, then puts the five branch outputs of image CNN and text CNN to a Bi-GRU as the inputs of different time steps, if the two sentiments are different. Or else, we take the five branch outputs of image CNN and text CNN as inputs of a fully-connected neural network.

4.4 Results and analysis

Tables 3 and 4 report the performances of MFC and comparison methods on the two datasets, respectively. For these two tables, the top-half part is the comparison with other baselines, the bottom-half part is the results of ablation studies. First, we observe that MFC outperforms DCN, CCR, DFC, DMAF, and AMGN approaches. This shows that the multi-level fusion strategy with multimodal information could improve the model's performance. However, ITIGNN method achieves the best performance.

Table 3 Experimental results on Flickr dataset

Method	Accuracy	Recall	F1 score
Image CNN	0.783	0.799	0.790
Text CNN	0.712	0.722	0.715
DCN	0.805	0.817	0.811
CCR	0.810	0.836	0.820
DFC	0.824	0.851	0.840
DMAF	0.859	0.845	0.850
AMGN	0.870	0.862	0.868
ITIGNN	0.919	0.917	0.922
MFC (Concat)	0.846	0.857	0.850
MFC (GRU)	0.851	0.870	0.862
MFC (3 branches)	0.870	0.886	0.877
MFC (4 branches)	0.871	0.890	0.879
MFC (Bi-LSTM)	0.873	0.890	0.880
R-MFC	0.884	0.890	0.888
MFC	0.885	0.897	0.891

Table 4 Experimental results on human labeled Flickr dataset

Method	Accuracy	Recall	F1 score
Image CNN	0.675	0.710	0.682
Text CNN	0.633	0.660	0.647
DCN	0.705	0.721	0.811
CCR	0.691	0.710	0.696
DFC	0.720	0.732	0.728
DMAF	0.753	0.751	0.759
AMGN	0.781	0.785	0.779
ITIGNN	0.828	0.820	0.829
MFC (Concat)	0.730	0.800	0.741
MFC (GRU)	0.777	0.793	0.780
MFC (3 branches)	0.730	0.764	0.739
MFC (4 branches)	0.735	0.729	0.733
MFC (Bi-LSTM)	0.773	0.780	0.775
R-MFC	0.801	0.799	0.800
MFC	0.818	0.801	0.811

We suppose that ITIGNN uses a pre-trained CNN with graph neural networks for textual and visual analysis. In comparison, our MFC method has fewer scale in parameters.

Furthermore, in these six variant models, the one with five branches and Bi-GRU fusion approach achieves the best performance. This verifies that better multi-level fusion could improve the multimodal sentiment classification performance. Furthermore, compared with MFC with 3 and 4 branches, the variant of 5 branches has a significant increase due to the non-linear effect of fusion information and local-to-global features of high-level CNNs. MFC shows better performance compared with R-MFC. This could be caused by underestimated trade-off between the conflict detection and without detection in MFC. In addition, this also indicates that the fusing of multiple middle layers of two modalities improves the performance of sentiment prediction.

5 Conclusions and future work

In this paper, we present a new fusion method for visual and textual sentiment analysis. Our MFC leverages different levels of features from multiple branches in image CNN and text CNN, and effectively integrates these features by exploiting the dependencies among them with the Bi-GRU approach. Extensive experimental results demonstrate that the proposed multi-level fusion method achieves comparable performance on multimodal sentiment analysis compared with strong baseline approaches. In the future, we will explore the effect of trade-off between conflict detection and without detection.

Acknowledgements

This work was supported in part by the National Key Research and Development (R&D) Program of China (2018YFB1403003).

References

[1] REN F J, WU Y. Predicting user-topic opinions in twitter with social and topical context. IEEE Transactions on Affective Computing, 2013, 4(4): 412 - 424.

[2] PENG L, CUI G, ZHUANG M Z, et al. What do seller manipulations of online product reviews mean to consumers. HKIBS/WPS/070 - 1314. Hong Kong, China: Hong Kong Institute of Business Studies (HKIBS), 2014.

[3] ASUR S, HUBERMAN B A. Predicting the future with social media. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology: Vol. 1, 2010, Aug 31 - Sept 3, Toronto, Canada. Piscataway, NJ, USA: IEEE, 2010: 492 - 499.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25(2): 1097 - 1105.

[5] LECUN Y, BOSER B, DENKE J S, et al. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1(4): 541 - 551.

[6] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13), 2013, May 26 - 31, Vancouver, Canada. Piscataway, NJ, USA: IEEE, 2013: 6645 - 6649.

[7] SANTOS C D, GATTI M. Deep convolutional neural networks for sentiment analysis of short texts. Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING'14), 2014, Aug 23 - 29, Dublin, Ireland. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014: 69 - 78.

[8] KIM Y. Convolutional neural networks for sentence classification. Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP'14), 2014, Oct 25 - 29, Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1746 - 1751.

[9] XU C, CETINTAS S, LEE K C, et al. Visual sentiment prediction with deep convolutional neural networks. arXiv Preprint, arXiv:1411.5731, 2014.

[10] YOU Q Z, LUO J B, JIN H L, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks. Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15), 2015, Jan 25 - 30, Austin, TX, USA. Menlo Park, CA, USA: American Association for Artificial Intelligence (AAAI), 2015: 381 - 388.

[11] D'MELLO S K, KORY J. A review and meta-analysis of multimodal affect detection systems. ACM Computing Survey, 2015, 47(3): 1 - 36.

[12] MONKARESI H, HUSSAIN M S, CALVO R A. Classification of affects using head movement, skin color features and physiological signals. Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC'12), 2012, Oct 14 - 17, Seoul, Republic of Korea. Piscataway, NJ, USA: IEEE, 2012: 2664 - 2669.

[13] PORIA S, CAMBRIA E, HUSSAIN A, et al. Towards an intelligent framework for multimodal affective data analysis. Neural Networks, 2015, 63: 104 - 116.

[14] SARKAR C, BHATIA S, AGARWAL A, et al. Feature analysis for computational personality recognition using youtube personality data set. Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition (WCPR'14), 2014, Nov 7, Orlando, FL, USA. New York, NY, USA: ACM, 2014: 11 - 14.

[15] WANG S F, ZHU Y C, WU G B, et al. Hybrid video emotional tagging using users' EEG and video content. Multimedia Tools and Applications, 2014, 72(2): 1257 - 1283.

[16] ALAM F, RICCARDI G. Predicting personality traits using multimodal information. Proceedings of the 2014 Workshop on Computational Personality Recognition (WCPR'14), 2014, Nov 7, Orlando, FL, USA. New York, NY, USA: ACM, 2014: 15 - 18.

[17] CAI G Y, XIA B B. Convolutional neural networks for multimedia sentiment analysis. Natural Language Processing and Chinese Computing: Proceedings of the 4th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC'15), 2015, Oct 9 - 13, Nanchang, China. LNCS 9362. Berlin, Germany: Springer, 2015: 159 - 167.

[18] DOBRIŠEK S, GAJŠEK R, MIHELI Č F, et al. Towards efficient multi-modal emotion recognition. International Journal of Advanced Robotic Systems, 2013, 10(1): 1 - 10.

[19] GLODEK M, REUTER S, SCHELS M, et al. Kalman filter based classifier fusion for affective state recognition. Multiple Classifier Systems: Proceedings of the 11th International Workshop on Multiple Classifier Systems (MCS'13), 2013, May 15 - 17, Nanjing, China. LNIP 7872. Berlin, Germany: Springer, 2013: 85 - 94.

[20] PORIA S, CAMBRIA E, GELBUKH A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, Sept 17 - 21, Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 2539 - 2544.

[21] PORIA S, CAMBRIA E, HOWARD N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 2016, 174 (Part A): 50 - 59.

[22] BALTRUŠAITIS T, BANDA N, ROBINSON P. Dimensional affect recognition using continuous conditional random fields. Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG'13), 2013, Apr 22 - 26, Shanghai, China. Piscataway, NJ, USA: IEEE, 2013: 1 - 8.

[23] METALLINOU A, WOLLMER M, KATSAMANIS A, et al. Context-sensitive learning for enhanced audiovisual emotion classification. IEEE Transactions on Affective Computing 2012, 3(2): 184 - 198.

[24] ADAMS W H, IYENGAR G, LIN C Y, et al. Semantic indexing of multimedia content using visual, audio, and text cues. EURASIP Journal on Advances in Signal Processing, 2003: 1 - 16.

[25] NEFIAN A V, LIANG L H , PI X B, et al. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP Journal on Advances in Signal Processing, 2002: 1 - 15.

[26] CORRADINI A, MEHTA M, BERNSEN N O, et al. Multimodal input fusion in human-computer interaction. NATO Science Series Sub Series III: Computer and Systems Sciences 198. Odense, Denmark: University of Southern Denmark, 2005.

[27] IYENGAR G, NOCK H J, NETI C. Audio-visual synchrony for detection of monologues in video archives. Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03): Vol 5, 2003, Apr 6 - 10, Hong Kong, China. Piscataway, NJ, USA: IEEE, 2003: V -

772.

[28] NICKEL K, GEHRIG T, STIEFELHAGEN R, et al. A joint particle filter for audio-visual speaker tracking. Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI'05), 2005, Oct 4 – 6, Torento, Italy. New York, NY, USA: ACM, 2005: 61 – 68.

[29] POTAMITIS I, CHEN H M, TREMOULIS G. Tracking of multiple moving speakers with multiple microphone arrays. IEEE Transactions on Speech and Audio Processing, 2004, 12(5): 520 – 529.

[30] CAMPOS V, JOU B, GIRÓ-I-NIETO X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. Image and Vision Computing, 2017, 65: 15 – 22.

[31] WANG J, YU L C, LAI K R, et al. Dimensional sentiment analysis using a regional CNN-LSTM model. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Vol 2, Short Papers, 2016, Aug 7 – 12, Berlin, Germany. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 225 – 230.

[32] RAO T R, LI X X, XU M. Learning multi-level deep representations for image emotion classification. Neural Processing Letters, 2020, 51: 2043 – 2061.

[33] JOU B, CHEN T, PAPPAS N, et al. Visual affect around the world: a large-scale multilingual visual sentiment ontology. Proceedings of the 23rd ACM International Conference on Multimedia (MM'15), 2015, Oct 26 – 30, Brisbane, Australia. New York, NY, USA: ACM, 2015: 159 – 168.

[34] YOU Q Z, LUO J B, JIN H L, et al. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16), 2016, Feb 22 – 25, San Francisco, CA, USA. New York, NY, USA: ACM, 2016: 13 – 22.

[35] YU Y H, LIN H F, MENG J N, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms, 2016, 9(2): 1 – 11.

[36] CHEN X Y, WANG Y H, LIU Q J. Visual and textual sentiment analysis using deep fusion convolutional neural networks. Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP'17), 2017, Sept 17 – 20, Beijing, China. Piscataway, NJ, USA: IEEE, 2017: 1557 – 1561.

[37] HUANG F R, ZHANG X M, ZHAO Z H, et al. Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 2019, 167: 26 – 37.

[38] HUANG F R, WEI K M, WENG J, et al. Attention-based modality-gated networks for image-text sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16(3): 1 – 19.

[39] LIAO W X, ZENG B, LIU J Q, et al. Image-text interaction graph neural network for image-text sentiment analysis. Applied Intelligence, 2022, DOI:10.1007/s10489 – 021 – 02936 – 9.

(Editor: Luo Lang)

From p. 24

[23] RAGHU A, RAGHU M, BENGIO S, et al. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. Proceeding of the 8th International Conference on Learning Representations (ICLR'20), 2020, Apr 27 – 30, Addis Ababa, Ethiopia. 2020: 1 – 6.

[24] OH J, YOO H, KIM C H, et al. Boil: towards representation change for few-shot learning. Proceeding of the 9th International Conference on Learning Representations (ICLR'21), 2021, May 3 – 7, Vienna, Austria. 2021: 1 – 15.

[25] LUO Y D, HUANG Z, ZHANG Z, et al. Learning from the past: continual meta-learning with Bayesian graph neural networks. Proceeding of the 34th AAAI Conference on Artificial Intelligence (AAAI'20), 2020, Feb 7 – 12, New York, NY, USA. Menlo Park, CA, USA: American Association for Artificial Intelligence (AAAI), 2020: 5021 – 5028.

[26] SANTORO A, BARTUNOV S, BOTVINICK M, et al. Meta-learning with memory-augmented neural networks. Proceeding of the 33rd International Conference on Machine Learning (ICML'16), 2016, Jun 19 – 24, New York, NY, USA. 2016: 1842 – 1850.

[27] HUANG H X, ZHANG J J, ZHANG J, et al. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 2020, 23: 1666 – 1680.

[28] WELINDER P, BRANSON S, MITA T, et al. Caltech-UCSD Birds 200. CNS-TR-2010-001. Pasadena, CA, USA: California Institute of Technology, 2010.

[29] KRAUSE J, STARK M, DENG J, et al. 3D object representations for fine-grained categorization. Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops (ICCV'13), 2013, Dec 2 – 8, Sydney, Australia. Piscataway, NJ, USA: IEEE, 2013: 554 – 561.

[30] KHOSLA A, JAYADEVAPRAKASH N, YAO B P, et al. Novel dataset for fine-grained image categorization: Stanford Dogs. https://people. csail. mit. edu/khosla/papers/fgvc2011. pdf. 2011.

[31] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115: 211 – 252.

[32] ZHANG X T, QIANG Y T, SUNG F, et al. RelationNet2: deep comparison columns for few-shot learning. Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN'20), 2020, Jul 19 – 24, Glasgow, UK. Piscataway, NJ, USA: IEEE, 2020: 1 – 15.

[33] ZHANG H G, LI H D, KONIUSZ P. Multi-level second-order few-shot learning. IEEE Transactions on Multimedia, 2022, Early Access Article, DOI:10.1109/TMM.2022.3142955.

(Editor: Luo Lang)