Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230918

Multi-modal sentiment analysis based on graph neural network

Boao Li

School of Computer Science and Technology, Soochow University, Suzhou, China, 215000

Liboao3524006397@gmail.com

Abstract. Thanks to popularity of social media, people are witnessing the rapid proliferation of posts with various modalities. It is worth noting that these multi-modal expressions share certain characteristics, including the interdependence of objects in the posted images, which is sometimes overlooked in previous researches as they focused on single image-text posts and pay little attention on obtaining the global features. In this paper, a neural network with multiple channels for image-text sentiment detection is proposed. The first step is to encode text and images to capture implicit tendencies. Then the introduction of this model obtains multi-modal expressions by collecting the shared characteristics of the dataset. Finally, the attention mechanism provides reliable predictions of the sentiment tendencies of the given pairs of image-text data. The results of experiments conducted on two publicly available datasets crawled from Twitter prove the reliability of the model on multi-modal sentiment detection, since the model precedes previously proposed models in the main evaluating criteria.

Keywords: Sentiment Detection, Multi-modal Sentiment Analysis, Deep Learning, Graph Neural Networks, Attention Mechanism

1. Introduction

Considering the popularity of social networks and the potential benefits that the unbelievable amount of data might provide, considerable attention has been paid on the tasks of sentiment detection on social media datasets from both academic and industrial communities [1]. In this paper, a model that predicts sentiment tendencies based on image-text pairs, which is normal in social media posts is introduced. The very challenges of this task is the unavoidable flaws of data collected from social networks, including the mismatching between texts and the posted images, and the frequent occurrence of slangs, abbreviations and lack of punctuation marks. To tackle these challenges, various different networks have been introduced in recent years. Xu et al. proposed HSAN and MDSN [2-3]. Yang et al. constructed the Co-Mem network and MVAN model [4-5]. These models introduce memory networks to realize the interaction between image and text modalities. However, these approaches treat each image-text post as an isolated instance, while shared feature dependencies beyond different instances are sometimes neglected. How to effectively analyze the co-occurring features across instances and capture the global characteristics of the dataset remains to be a great challenge.

The model for multi-modal sentiment detection consists mainly of three phases. For texts, the first step is text encoding and the construction of dictionaries; and for images, the model extracts objects and

DOI: 10.54254/2755-2721/6/20230918

scenes to capture the semantic features. Then, the model introduces a Graph Neural Network based on the co-occurrence matrix. For images, considering that certain objects of an image stand for representations of different emotions, two graphs are built for scenes and objects respectively based on their co-occurrences. Then a Graph Convolutional Network (GCN) is introduced over the two graphs to obtain the features of the image [3]. To provide complementary results from different data pairs, different graphs are needed for multiple modalities, and by combining different GNN channels, the Multi-channel Graph Neural Networks module is built to collect the implicit common features of different text-image pairs. It is proved that the module can provide reliable representation from multiple datasets [6]. Finally, through the introduction of the attention module, the model realizes the interaction between modalities from different channels and get a better representation.

The work can be summarized as follows:

(i) The paper proposes a Multi-channel Graph Neural Network to provide global sentiment characteristics.

(ii) The paper applies the attention mechanism from different channels to realize multi-modal interactions.

2. Related work

2.1. Multi-modal sentiment detection

Traditional machine learning methods have long been used in multi-modal emotion detection tasks. In recent years, the introduction and studies on deep learning have also achieved satisfying outcomes. For visual datasets, a method to capture multi-modal features with end-to-end translation modules called TransModality was proposed by Wang et al., and Hazarika et al. constructed another framework, MISA, which projects each modality to two subspaces: modality-invariant and modality-specific [7-8]. For text modality, pre-trained models based on Transformer, especially exemplified as GPT, BERT and RoBERTa, prove to be efficient for analyzing linguistic context representations [9-12]. Considering the massive amount of image-text data available on social platforms, the study of graphical and literal multi-modal emotion detection has aroused the attention of many researchers. Deep learning and neural networks are popular among recent multi-modal sentiment analysis approaches. Classical deep learning models of multi-modal sentiment detection include convolutional neural networks (CNNs) and support vector machines (SVMs) [13-14]. Besides, there exist several multi-modal datasets with sentiment annotations that are frequently used in experiments, including the newly-introduced CMU-MOSI dataset, as well as other datasets including TumEmo and MOUD [4, 15-16].

2.2. Graph neural networks

GNN has achieved encouraging results for text sentiment analysis and multi-modal tasks. It has been proved that GNN perform better than other traditional method like Text GCN and Tensor GCN thanks to the rapid development of it as well as its variants [17-18]. For the image-text dataset that the paper lay experiments on, it is found that emotional tendencies can be indicated through the simultaneous occurrence of certain texts as well as the re-occurring objects in the image. The model respectively obtains these features of different pairs through the multi-channel GNN module.

2.3. Multi-modal attention module

Inspired by the prototype of Transformer, a Multi-model Attention Interaction module is built to acquire the relationship and influence of text modality and image modality in order to provide more reliable results.

3. Proposed model

The overall process of the proposed model consists of three phases:

First, the input data of text and images are encoded into implicit expressions for the model to understand. Then, Graph Neural Networks are introduced with different channels, each of the channels

DOI: 10.54254/2755-2721/6/20230918

capturing emotional expressions from a modal. For instance, the Text-GNN (TG) for texts, and the Scene-GCN (SG) and Object-GCN (OG) for scenes and objects in the images. Finally, the model endows certain weights for different modalities to obtain the implicit emphasis of different modalities and provide more reliable predictions through the multi-modal attention module.

P stands for a set of multi-modal posts collected from the social media which includes text and image pairs, P = {(T_1, V_1), ..., (T_N , V_N)}. T_i and V_i stand for the information of text and visual modalities, and N represents the number of pairs in the set. The function $f : P \rightarrow C$ needs to be acquired to classify the posts while C stands for the predefined categories of emotions. Then the pairs (T_i, V_i) are casted into certain categories C_i. For polarity classification, $C_i \in$ {Positive, Neutral, Negative}; and for emotion detection, $C_i \in$ {Angry, Bored, Calm, Fear, Happy, Love, Sad}.

3.1. Encoding

For textual messages, the model encodes the text using GloVe for the embedding vector, and then retrieves the repository through BiGRU [19-20].

$$R^t = f_{BiGRU}\big(Embedding(T)\big), R^t \in R^{L^t * 2d^t} \quad (1)$$

T is the text sequence obtained through encoding, L^t is the length of padded text sequence, and d^t is the dimension of implicit units in the BiGRU layer.

For visual messages, the model captures features of images from both objects and scenes for intact information. To explicitly model the interdependence between the objects and scenes in the image, the module extracts objects O by YOLOv and scenes S by VGG-place [21-22].[22] Then the object and scene memory banks can be obtained from a pretrained ResNet [23].[23]

$$B^x = f^x_{ResNet}(V), B^x \in R^{L^X * d^x} \quad (2)$$

Where x $\in$ {Object, Scene}.

3.2. Multi-channel graph neural networks

As is mentioned in the previous part, the multi-GNN consists of the Text-GNN module, the Scene-GCN module and the Object-GCN module.

Text-GNN: For text T with l^t words, T = {ω_1, ..., ω_{l^t}}, where the k^{th} word ω_k is obtained through the GloVe embedding $r^t_k \in R^d$, d = 300. Then a text-based dictionary graph N^t is acquired;

$$N^t = \{\omega_k | k \in [1, l^t]\} \quad (3)$$

In the graph, the edge between w_k and w_j is to be built if the two words co-occur for no less than 2 times. In this way, the graph of edges in N^t can be defined;

$$E^t = \{e^t_{k,j} \mid \omega_k \in [\omega_1, \omega_{l_t}]; \omega_j \in [\omega_{k-ws}, \omega_{k+ws}]\} \quad (4)$$

The representations of words in N^t and the weights of the edges in E^t can be obtained from a general matrix created based on the vocabulary bank and edge set of the data. In this case, the representation of some nodes and weights is accessible globally. $e^t_{k,j}$ is initialized and learned during training [24]. ws is the size of the mobile window in the layer, which indicates the number of nodes that is connected to a single word in the text. Then, by the message passing mechanism, the representation of nodes based on the original representation can now be reloaded, as well as that of their neighbor nodes [25].

$$A^t_k = \max_{j \in N^{ws}_k} e^t_{kj} r^t_k \quad (5)$$

$$r^{t'}_k = \alpha r^t_k + (1 - \alpha) A^t_k \quad (6)$$

Where A^t_k is the information that is to be collected from neighboring nodes $k - ws$ to $k + ws$. α indicates how much original information of the nodes is to be preserved, $r^{t'}_k$ is the new representation

DOI: 10.54254/2755-2721/6/20230918

of the node k. In this way, the new representation of the text T can be calculated:

$$T' = \sum_{k=1}^{l^t} r_k^{t'} \tag{7}$$

Image-GCN: For the images, the paper explicitly model objects and scenes of the images through the previously introduced SG and OG module. In this way, the graph can be built as follows:

$$N^x = \{x_p \mid p \in [1, l^x]\} \tag{8}$$

Where l^x is the number of objects or scenes in the image and N^x is the set of nodes in the graph. Then, to build the set of edges, the first step is to build the dataset of co-occurrence matrix E^x:

$$E^x = \{e_{p,q}^x \mid p \in [1, l^x], q \in [1, l^x]\} \tag{9}$$

Where $e_{p,q}^x$ indicates the times $x_p$and x_q co-occurred.

Then the weight of the edge (p, q) can be calculated as follows:

$$P_{p,q}^x = e_{p,q}^x / N_p^x \tag{10}$$

Where N_p^x indicates the times of occurrences of x_p in the dataset.

Finally, the module input the set N^x and P^x into the graph convolutional network. Each layer of the network is calculated as follows:

$$H_{L+1}^x = h(\hat{R}^x H_L^x W_L^x) \tag{11}$$

Where $\hat{R}_x$ is the normalized expression of R^x, and $h()$ stands for a non-linear operation.

Then, by stacking the GCN layers, the interdependence of the nodes is modeled. In this way, the representation of objects or scenes dependencies can be obtained:

$$I^x = MaxPooling\ (M^x)(H_{L+1}^x)^T \tag{12}$$

3.3. Multi-modal attention module

The MMA module is introduced to capture the representations explicitly when guided by texts and images. Note that the image-guided attention module consists of both objects-guided and scene-guided attention. For the text-guided attention:

$$O_{N+1}^{TgX} = LN(MH\big(Q = H_N^{TgX}, K = V = M^x\big) + H_N^{TgX} \tag{13}$$

$$H_{N+1}^{TgX} = LN(FFN(O_{N+1}^{TgX}) + O_{N+1}^{TgX}) \tag{14}$$

For the image-guided attention:

$$O_{N+1}^{XgT} = LN(MH\big(Q = H_N^{XgT}, K = V = M^t\big) + H_N^{XgT} \tag{15}$$

$$H_{N+1}^{XgT} = LN(FFN(O_{N+1}^{XgT}) + O_{N+1}^{XgT}) \tag{16}$$

Where $LN()$ stands for layer normalization operation, and $FFN()$ is the feed-forward network. X $\in$ {Object, Scene}.

The fused multi-modal representation is obtained as follows:

$$R^m = [H_N^{TgO} \oplus H_N^{TgS} \oplus H_N^{OgT} \oplus H_N^{SgT}] \tag{17}$$

Where $\oplus$ serves the function of concatenation.

3.4. Sentiment analysis

Finally, the model input the fused representation R^m, which is calculated previously, into the top fully connected layer and use the softmax function to obtain the final results of sentimental analysis.

DOI: 10.54254/2755-2721/6/20230918

$$L^m = softmax\ (w^s R^m + b^s) \quad (18)$$

Where w^s and b^s are parameters of the top fully connected layer.

4. Experiments and analysis

4.1. Datasets

The experiments are conducted on two publicly available multi-modal datasets: MVSA-Single and MVSA-Multiple [26]. The two datasets which is collected from the popular social media Twitter contains emotional text-image data of different scales. Some figures of the two datasets are shown in Table 1.

Table 1. Some figures of the datasets.

Dataset	Train	Val	Test	Total
MSVA-Single	3608	451	452	4511
MSVA-Multiple	13618	1703	1703	17024

4.2. Setup

Table 2. Parameter settings of the dataset.

Parameters	
Learning rate	$4e-5$
ws	4
L^x	2
N	1

In the process of extracting the representation images, the model discards the objects with the frequency lower than 0.5 and preserves the 5 highest ranking scenes. Table 2 shows how some of the parameters in the experiments are set. Noting that the precise selection of the hyperparameter ws should vary when it comes to different datasets due to the different average length of the text. The GNN cannot obtain sufficient information from neighboring nodes if the ws value is too small, while setting the value of ws too large also degrades the performance as neighboring nodes fail to provide abundant information required.

4.3. Baselines

By comparing the model is this paper with the previously introduced multi-modal sentiment models with the same modalities, the reliability of the proposed model can be proved. HSAN is a delaminated sentimental attentional network based on image titles. MDSN is a deep learning network with attention mechanism for multi-modal sentiment analysis, and Co-Mem is a globally shared network for iterative modeling of interactions between multiple modalities.

4.4. Results

Accuracy (Acc) and F1-score (F1) are most frequently used as evaluation metrics. The results of experiments are shown in Table 3.

Table 3. Experiment results on the two datasets.

Models	MVSA-Single		MVSA-Multiple	
	Acc	F1	Acc	F1
HSAN	0.6988	0.6690	0.6796	0.6776
MDSN	0.6984	0.6963	0.6886	0.6811
Co-Mem	0.7051	0.7001	0.6992	0.6983
MGNN	0.7232	0.7127	0.7112	0.6839

DOI: 10.54254/2755-2721/6/20230918

The proposed model achieved 0.7232/0.7112 in accuracy and 0.7127/0.6839 in F1, out-performing the baseline models on both datasets, which proves that the idea of achieving the interaction between different modalities benefits the prediction of sentiment tendencies.

5. Conclusion

This paper proposes the Multi-channel Graph Neural Network model which is built according to the global features of the dataset for multi-modal sentiment detection tasks. The experimental results on publicly available datasets demonstrated that the proposed model is competitive with strong baseline models. To make the sentiment detection more reliable, we can make further calculation on the GCN layers, where the simple correlation we used may suffer certain drawbacks, such as the influence of noisy edges and the different importance between nodes. In the future, further experiments on the transferability of the model as well as the necessity of each module can be conducted to have a better understanding of this model.

References

[1] Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. Knowledge and Information Systems, 60(2), 617-663.

[2] Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), pages 152–154. IEEE.

[3] Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 2399–2402. ACM.

[4] Yang, X., Feng, S., Wang, D., & Zhang, Y. (2020). Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia, 23, 4014-4026.

[5] Nan Xu, Wenji Mao, and Guandan Chen. 2018. A comemory network for multimodal sentiment analysis. In The 41st international ACM SIGIR conference on research & development in information retrieval, pages 929–932.

[6] Anjith George and Sebastien Marcel. 2021. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. IEEE Transactions on Information Forensics and Security, 16:361–375

[7] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020b. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In Proceedings of The Web Conference 2020, pages 2514–2520.

[8] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, pages 1122–1131.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R.Garnett, editors, Advances in Neural Information Processing Systems 30 (NIPS), pages 5998–6008.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

DOI: 10.54254/2755-2721/6/20230918

[13] Soujanya Poria, Erik Cambria, and Alexander F Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In EMNLP. pages 2539–2544.
[14] Amir Zadeh. 2015. Micro-opinion sentiment intensity analysis and summarization in online videos. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, pages 587– 591.
[15] Amir Zadeh, Rowan Zellers, Eli Pincus, and LouisPhilippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems 31(6):82–88.
[16] Veronica P ´ erez-Rosas, Rada Mihalcea, and Louis-´ Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In ACL (1). pages 973– 982
[17] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 7370–7377.
[18] Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. 2020. Tensor graph convolutional networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8409–8416.
[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.
[20] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724– 1734.
[21] Redmon, J., & Farhadi, A. (2018). YOLOv3: an incremental improvement, computer vision and pattern recognition. arXiv 2018. arXiv preprint arXiv:1804.02767.
[22] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6): 1452–1464.
[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770– 778.
[24] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020a. Fake news detection via knowledge-driven multimodal graph convolutional networks. In Proceedings of the 2020 International Conference on Multimedia Retrieval, pages 540–547.
[25] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, pages 1263–1272.
[26] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In International Conference on Multi-media Modeling, pages 15–27. Springer.