Fusion-Extraction Network for Multimodal Sentiment Analysis

Tao Jiang, Jiahai Wang(✉), Zhiyue Liu, and Yingbiao Ling

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
wangjiah@mail.sysu.edu.cn

Abstract. Multiple modality data bring new challenges for sentiment analysis, as combining varieties of information in an effective manner is a rigorous task. Previous works do not effectively utilize the relationship and influence between texts and images. This paper proposes a fusion-extraction network model for multimodal sentiment analysis. First, our model uses an interactive information fusion mechanism to interactively learn the visual-specific textual representations and the textual-specific visual representations. Then, we propose an information extraction mechanism to extract valid information and filter redundant parts for the specific textual and visual representations. The experimental results on two public multimodal sentiment datasets show that our model outperforms existing state-of-the-art methods.

Keywords: Sentiment analysis · Multimodal · Fusion-Extraction Model

1 Introduction

With the prevalence of social media, social platforms like Twitter and Instagram, have become part of our daily lives and played an important role in people's communication. As a result of the increasing multimodality of social networks, there are more and more multimodal data which combine images and texts in social platforms. Though providing great conveniences for people communication, multimodal data bring growing challenges for social media analytics. In fact, it is often the case that the sentiment cannot be reflected with the support of single modality information. The motivation is to leverage the varieties of information from multiple sources for building an efficient model.

This paper studies the task of sentiment analysis for social media, which contains both visual and textual contents. Sentiment analysis is a core task of natural language processing, and aims to identify sentiment polarity towards opinions, emotions, and evaluations. Traditional methods [14,21] for text-only sentiment analysis are mainly statistical methods which highly rely on the quality of feature selection. With the rapid development of machine learning techniques and deep neural network, researchers introduce many dedicated methods [7, 13], which achieve significantly improved results. In contrast to single modality

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12085, pp. 785–797, 2020.
https://doi.org/10.1007/978-3-030-47436-2_59

based sentiment analysis, multimodal sentiment analysis attracts more and more attention in recent works [20,24,26,28].

However, most previous works cannot effectively utilize the relationship and influence between visual and textual information. Xu et al. [22] only take the single-direction influence of image to text into consideration and ignore interactive promotion between visual and textual information. A co-memory network [23] then is proposed to model the interactions between visual contents and textual words iteratively. Nevertheless, the co-memory network only applies a weighted textual/visual vector as the guide to learn attention weights on visual/textual representation. It can be seen as a coarse-grained attention mechanism and may cause information loss because attending multiple contents with one attention vector may hide the characteristic of each attended content. Further, the previous studies directly apply multimodal representations for final sentiment classification. However, there is partial redundancy information which may bring confusion and is not beneficial for final sentiment classification.

This paper proposes a new architecture, named **F**usion-**E**xtraction **N**etwork (FENet), to solve the above issues for the task of multimodal sentiment classification. First, a fine-grained attention mechanism is proposed to interactively learn cross-modality fused representation vectors for both visual and textual information. It can focus on the relevant parts of texts and images, and fuse the most useful information for both single modality. Second, a gated convolution mechanism is introduced to extract informative features and generate expressive representation vectors. The powerful capability of Convolution Neural Networks (CNNs) for image classification has been verified [8,19]. It is a common way that applying CNNs to extract relativeness of different regions of an image. For textual information, it deserves to be pointed out that CNNs also have strong ability to process [25]. CNNs have been observed that they are capable of extracting the informative n-gram features as sentence representations [10]. Thus, the convolution mechanism is quite suitable for the extraction task in the multimodal sentiment classification. Meanwhile, we argue that there should be a mechanism controlling how much part of each multimodal representation can flow to the final sentiment classification procedure. The proposed gate architecture mechanism plays the role to modulate the proportion of multimodal features. The experimental results on two public multimodal sentiment datasets show that FENet outperforms existing state-of-the-art methods.

The contributions of our work are as follows:

- We introduce an **I**nteractive **I**nformation **F**usion (IIF) mechanism to learn fine-grained fusion features. IIF is based on cross-modality attention mechanisms, aiming to generate the visual-specific textual representation and the textual-specific visual representation for both two modality contents.
- We propose a **S**pecific **I**nformation **E**xtraction (SIE) mechanism to extract the informative features for textual and visual information, and leverage the extracted visual and textual information for sentiment prediction. To the best of our knowledge, no CNN-gated extraction mechanism for both textual and visual information has been proposed in the field of multimodal sentiment analysis so far.

2 Related Work

Various approaches [1,4,5] have been proposed to model sentiment from text-only data. With the prevalence of multimodal user-generated contents in social network sites, multimodal sentiment analysis becomes an emerging research field which combines textual and non-textual information. Traditional methods adopt feature-based methods for multimodal sentiment classification. Borth et al. [2] firstly extract 1200 adjective-noun pairs as the middle-level features of images for classification, and then calculate the sentiment scores based on English grammar and spelling style of texts. However, these feature-based methods highly depend on the laborious feature engineering, and fail to model the relation between visual and textual information, which is critical for multimodal sentiment analysis.

With the development of deep learning, deep neural networks have been employed for multimodal sentiment classification. Cai et al. [3] and Yu et al. [27] use CNN-based networks to extract feature representations from texts and images, and achieve significant progress. In order to model the relatedness between text and image, Xu et al. [22] extract scene and object features from image, and absorb text words with these visual semantic features. However, they only consider the visual information for textual representation, and ignore the mutual promotion of text and image. Thus, Xu et al. [23] propose a co-memory attentional mechanism to interactively model the interaction between text and image. Though taking the mutual influence of text and image into consideration, Xu et al. [23] adopt a coarse-grained attention mechanism which may not have enough capacity to extract sufficient information. Furthermore, they simply concatenate the visual representation and the textual representation for final sentiment classification. Instead, our model applies a fine-grained information fusion layer, and introduces an information extraction layer to extract and leverage visual and textual information for sentiment prediction.

3 Our Model

Given a text-image pair (T, I), where $T = \{T_1, T_2, \ldots, T_M\}$ and I is a single image, the goal of our model is to predict the sentiment label $y \in \{positive, neutral, negative\}$ towards the text-image pair.

The overall architecture of the proposed FENet is shown in Fig. 1. The bottom layer includes a text encoding layer and an image encoding layer, which transforms the text $T = \{T_1, T_2, \ldots, T_M\}$ to $X = \{x_1, x_2, \ldots, x_M\} \in \mathbb{R}^{d_w \times M}$ and transforms image to a fixed size vector separately, where d_w denotes the dimensions of the word embeddings. The middle part of our model is an interactive information fusion (IIF) layer simultaneously used to interactively learn cross-modality fusion for text and image. The IIF layer contains a fine-grained attention mechanism and identity mapping [9], which allows fuse one modality information with another modality data and learns more specific features. The top part is a specific information extraction (SIE) layer, which consists of two gated convolution layers and a max-pooling layer. The SIE layer first utilizes

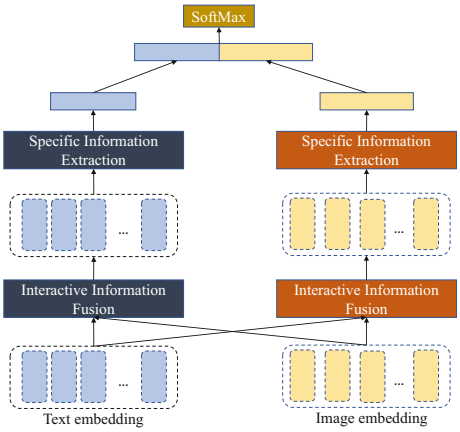


Fig. 1. The architecture of the proposed FENet.

convolution to extract informative features, and then selectively adjusts and generates expressive representations with gate mechanisms and a max-pooling layer. Finally, the visual-specific textual representation and the textual-specific visual representation from the SIE layer are concatenated for sentiment classification.

3.1 Text Encoding Layer

The function of the text encoding layer is mapping each word into a low dimensional, continuous and real-valued vector, also known as word embedding. Traditional word embedding can be treated as parameters of neural networks or pretrained from proper corpus via unsupervised methods such as Glove [17]. Further, a pretrained bidirectional transformer language model, also known as BERT [6], has shown its powerful capacity as word embedding. We applies Glove-based embedding for basic embedding and BERT-based embedding for extension embedding. The model variants are named FENet-Glove and FENet-BERT, respectively.

- **FENet-Glove.** It applies Glove as the basic embedding to obtain the word embedding of each word. Specifically, we employ a word embedding matrix $L \in \mathbb{R}^{d_w \times |V|}$ to preserve all the word vectors, where d_w is the dimension of word vector and $|V|$ is the vocabulary size. The word embedding of a word w_i can be notated as $l \in \mathbb{R}^{d_w}$, which is a column of the embedding matrix L.
- **FENet-BERT.** It uses BERT as the extension embedding to obtain the word representation of each word. Specifically, we use the last layer of BERT-base[1] to obtain a fixed-dimensional representation sequence of the input sequence.

[1] BERT-base contains 12 self-attention blocks, and its hidden dimension is 768.

3.2 Image Encoding Layer

Given an image I_p, where I_p indicates the image I rescaled to 224×224 pixels, we use Convolutional Neural Networks (CNNs) to obtain the representations of images. Specifically, the visual embedding V is obtained from the last convolutional layer of ResNet152[2] [8] pretrained on ImageNet [18] classification. This process can be described as follows:

$$V = ResNet152(I_p), \tag{1}$$

where the dimension of V is $2048 \times 7 \times 7$. 2048 denotes the number of feature maps, 7×7 means the shape of each feature maps. We then flatten each feature map into 1-D feature vector v_i corresponded to a part of an image.

$$V = \{v_1, v_2, \ldots, v_{2048}\}, v_i \in \mathbb{R}^{49}. \tag{2}$$

3.3 Interactive Information Fusion Layer

The above encoding representation only considers their single modality, and the attention mechanism is often applied to capture the interactions between different modality representations. However, previous works [22,23] adopt coarse-grained attention which may cause information loss, as the text contains multiple words and the image presentation contains multiple feature maps. In contrast, as shown in the middle part of Fig. 1, we adopt the IIF layer to solve this problem and the detail of the IIF mechanism is shown in Fig. 2(a).

Given two modality inputs, one of them is the target modality input which we fuse with another modality input named auxiliary input to generate the target modality output. Specifically, given a target input $S = \{S_1, S_2, \ldots, S_n\} \in \mathbb{R}^{d_s \times n}$ and an auxiliary input $A = \{A_1, A_2, \ldots, A_l\} \in \mathbb{R}^{d_a \times l}$, we first project the target input S and the auxiliary input A into the same shared space. The projecting process can be depicted as follows:

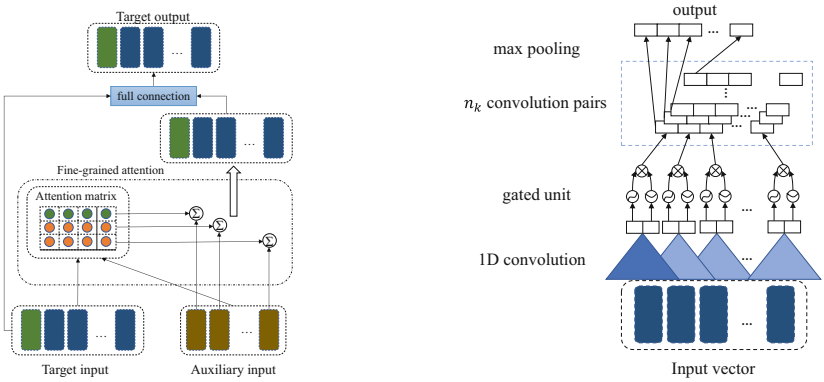
$$S_{emb_i} = tanh(W_{S_{emb}} S_i + b_{S_{emb}}), \tag{3}$$

$$A_{emb_i} = tanh(W_{A_{emb}} A_i + b_{A_{emb}}), \tag{4}$$

where $W_{S_{emb}} \in \mathbb{R}^{d_h \times d_s}$, $W_{A_{emb}} \in \mathbb{R}^{d_h \times d_a}$, $b_{S_{emb}}, b_{A_{emb}} \in \mathbb{R}^{d_h}$ are trainable parameters, and d_h denotes the dimension of shared space. Then, we use S_{emb} and A_{emb} to calculate the fine-grained attention matrix. Formally, we define the attention matrix as an alignment matrix $M \in \mathbb{R}^{n \times l}$, and M_{ij} indicates the relatedness between the i-th content of target input and the j-th content of auxiliary input. The alignment matrix M is computed by

$$M_{ij} = {S_{emb_i}}^T A_{emb_j}. \tag{5}$$

[2] ResNet152 indicates residual nets with a depth of up to 152 layers.



(a) (b)

Fig. 2. Details of IIF and SIE layer. (a) IIF layer. (b) SIE layer.

For each row of M, a softmax function is applied for quantifying the importance of each piece of auxiliary input to a specific piece of target input as follows:

$$M_{ij} = \frac{exp(M_{ij})}{\sum_{j=1}^{l} exp(M_{ij})}. \tag{6}$$

Then, the fine-grained attention output F is formulated as follows:

$$F = A \cdot M^T, \tag{7}$$

where $F \in \mathbb{R}^{d_a \times n}$ and "$\cdot$" denotes matrix multiplication. Finally, the concatenation of the target input S and the fine-grained attention output F is fed into a full connection layer to obtain the specific representation $G = \{G_1, G_2, \ldots, G_n\}$ of the target input:

$$G_i = tanh(W_g[S_i : F_i] + b_g), \tag{8}$$

where $G_i \in \mathbb{R}^{d_s}$ and $W_g \in \mathbb{R}^{d_s \times (d_s + d_a)}$. Thus, the overall process of IIF can be summarized as follows:

$$G = IIF(S, A). \tag{9}$$

Therefore, the textual-specific visual representation V_g and the visual-specific textual representation X_g are obtained as follows:

$$V_g = IIF(V, X), \tag{10}$$

$$X_g = IIF(X, V). \tag{11}$$

3.4 Specific Information Extraction Layer

After interactively fusing two modality information, we need to extract the most informative representation and control the proportion contributing to the final sentiment classification. As shown in the top part of Fig. 1, we introduce the SIE layer for this task and the details of the SIE layer is depicted in Fig. 2(b).

The SIE layer is based on convolutional layers and gated units. Given a padded input vector $Q = \{q_1, q_2, \ldots, q_k\} \in \mathbb{R}^{d_q \times k}$, we pass it through the SIE layer to get the final representation. First, n_k one dimensional convolutional kernel pairs are applied to capture the active local features. Each kernel corresponds a feature detector which extracts a specific pattern of active local features [11]. However, there are differences within the kernel pairs for their different nonlinearity activation function. The first kernel of kernel pairs is adopted to transform the information and obtain informative representation. While the second kernel of kernel pairs is a gate which controls the proportion of the result of the first kernel flowing to the final representation. Specifically, a convolution kernel pair of W_a and W_b maps r columns in the receptive field to a single feature a and b with *tanh* and *sigmoid* activation function, respectively. e is the result of multiplication of a and b, which stands for the representation after extraction and adjustment. As the filter slide across the whole sentence, a sequence of new feature $\mathbf{e} = \{e_1, e_2, \ldots, e_{k-r+1}\}$ is obtained by:

$$a_i = tanh(q_{i:i+r-1} * W_a + b_a), \tag{12}$$

$$b_i = sigmoid(q_{i:i+r-1} * W_b + b_b), \tag{13}$$

$$e_i = a_i \times b_i, \tag{14}$$

where $W_a, W_b \in \mathbb{R}^{d_q \times r}$ are weights of the convolution kernel pair, and $b_a, b_b \in \mathbb{R}$ are bias of the convolution kernel pair. "$*$" denotes the convolution operation. As there are n_k kernel pairs, the output features can form a matrix $E \in \mathbb{R}^{(k-r+1) \times n_k}$. Finally, we apply a max-pooling layer to obtain the most informative features for each convolution kernel pair, which results in a fixed-size vector z whose size is equal to the number of filter pairs n_k as follows:

$$z = [max(\mathbf{e}_1), \ldots, max(\mathbf{e}_{n_k})]^T. \tag{15}$$

The above process can be summarized as follows:

$$z = SIE(Q). \tag{16}$$

We treat V_g and X_g as the input of SIE to obtain the final visual and textual representation, respectively. The process is formulated as follows:

$$V_z = SIE(V_g), \tag{17}$$

$$X_z = SIE(X_g). \tag{18}$$

Table 1. Hyper-parameters of our model.

Hyper-parameter	Value
IIF shared space size d_h	100
SIE convolution kernel pair size n_k	50
SIE convolution kernel size r	3
Dropout rate	0.3

3.5 Output Layer

After obtaining the final feature representation vectors for image and text, we concatenate them as the input of a fully connected layer for classification:

$$p = Softmax(W_p[V_z : X_z] + b_p), \tag{19}$$

where $W_p \in \mathbb{R}^{class \times 2n_k}$ and $b_p \in \mathbb{R}^{class}$ are learnable parameters.

4 Experiments and Results

4.1 Datasets and Settings

Datasets. We use MVSA-Single and MVSA-Multiple [15] two datasets. The former contains 5129 text-image pairs from Twitter and is labeled by a single annotator. The later has 19600 text-image pairs labeled by three annotators. For fair comparison, we process the original two MVSA datasets on the same way used in [22,23]. We randomly split the datasets into training set, validation set and test set by using the split ratio 8:1:1.
Tokenization. On the one hand, to tokenize the sentences for Glove-based embedding method, we apply the same rule as [16], except we separate the tag '@' and '#' with the words after. On the other hand, we use the WordPiece tokenization introduced in [6] for BERT-based embedding method.
Word Embeddings. To initialize words as vectors, FENet-Glove uses the 300-dimensional pretrained Glove embeddings, and FENet-BERT applies 768-dimensional pretrained BERT embeddings which contains 110M parameters.
Pretrained CNNs. We use the pretrained ResNet152 [8] from Pytorch.
Optimization. The training objective is cross-entropy, and Adam optimizer [12] is adopted to compute and update all the training parameters. Learning rate is set to 1e−3 and 2e−5 for Glove-based and BERT-based embedding, respectively.
Hyper-parameters. We list the hyper-parameters during our training process in Table 1. All hyper-parameters are tuned on the validation set, and the hyper-parameters collection producing the highest accuracy score is used for testing.

4.2 Compared Methods

We compare with the following baseline methods on MVSA datasets.
SentiBank & SentiStrength [2] extracts 1200 adjective-noun pairs as the middle-level features of image and calculates the sentiment scores based on English grammar and spelling style of texts.
CNN-Multi [3] learns textual features and visual features by applying two individual CNN, and uses another CNN to exploiting the internal relation between text and image for sentiment classification.
DNN-LR [27] trains a CNN for text and employs a deep convolutional neural network for image, and uses average strategy to aggregate probabilistic results which is the output of logistics regression.
MultiSentiNet [22] extracts deep semantic features of images and introduces a visual feature attention LSTM model to absorb the text words with these visual semantic features.
CoMN [23] proposes a memory network to iteratively model the interactions between visual contents and textual words for sentiment prediction.

Besides, this paper also presents two ablations of FENet to evaluate the contribution of our components.
FENet w/o IIF removes the IIF component from the original model, and the text embedding and image embedding are fed into the SIE layer directly.
FENet w/o SIE replaces the SIE component with a max-pooling layer to get the final representation vector for sentiment classification.

4.3 Results and Analysis

Table 2 shows the performance comparison results of FENet with other baseline methods. As shown in Table 2, we have the following observations.

(1) **SentiBank & SentiStrength** is the worst since it only uses traditional statistical features to present image and text multimodality information, which can not make full of the high-level characteristic of multimodal data. Both **CNN-Multi** and **DNN-LR** are better than **SentiBank & SentiStrength** and achieve close performances by applying CNN architecture to learn two modality representation. **MultiSentiNet** and **CoMN** get outstanding results as they take the interrelations of image and context into consideration. **CoMN** is slightly better than **MultiSentiNet** because **MultiSentiNet** only considers the single-direction influence of image to text and ignores the mutual reinforcing and complementary characteristics between visual and textual information. However, **CoMN** employs the coarse-grained attention mechanism which may cause information loss, and directly uses redundant textual and visual representations for final sentiment classification. In contrast, **FENet** applies an information-fusion layer based on fine-grained attention mechanisms, and leverages visual and textual information for sentiment prediction by adopting an information extraction layer. Thus, **FENet** variants perform better than **CoMN** and achieves a new state-of-the-art performance.

Table 2. Experimental results of different models on two MVSA datasets. For fair comparison, ablated FENet is based on Glove embedding. CoMN(6) indicates that CoMN with 6 memory hops. The results of baseline methods are retrieved from published papers and the best two performances are marked in bold. The marker ♮ refers to p-value < 0.01 when comparing with MultiSentiNet, while the marker ♯ refers to p-value < 0.01 when comparing with CoMN(6).

	Model	MVSA-Single		MVSA-Multiple	
		ACC	F1	ACC	F1
Baselines	SentiBank & SentiStrength	0.5205	0.5008	0.6562	0.5536
	CNN-Multi	0.6120	0.5837	0.6630	0.6419
	DNN-LR	0.6142	0.6103	0.6786	0.6633
	MultiSentiNet	0.6984	0.6963	0.6886	0.6811
	CoMN(6)	0.7051	0.7001	0.6892	0.6883
Ablated FENet	FENet w/o IIF	0.6920	0.6882	0.6837	0.6795
	FENet w/o SIE	0.7120	0.7102	0.6989	0.6964
FENet variants	FENet-Glove	**0.7254**$^{\natural\sharp}$	**0.7232**$^{\natural\sharp}$	**0.7057**$^{\natural}$	**0.7038**$^{\natural\sharp}$
	FENet-BERT	**0.7421**$^{\natural\sharp}$	**0.7406**$^{\natural\sharp}$	**0.7146**$^{\natural\sharp}$	**0.7121**$^{\natural\sharp}$

(2) The results of both two ablations of **FENet** in accuracy and F1 are inferior to those of **FENet** variants. On the one hand, after removing the interactive information extraction layer, **FENet** cannot capture the interrelations between image and text, which are significant for sentiment analysis. Specifically, the performance of **FENet w/o IIF** degrades more than **FENet w/o SIE** by 2.0% of accuracy in MVSA-Single and 1.5% of accuracy in MVSA-Multiple. It verifies that the visual-specific textual representation and the textual-specific visual representation bring useful information for sentiment classification. On the other hand, **FENet w/o SIE** removes the SIE layer from FENet and only contains the IIF layer, which achieves better performances than **CoMN**. It is suggested that fine-grained attention can capture more specific information than coarse-grained attention. Furthermore, the SIE component also plays a key role in our model. **FENet-Glove** outperforms **FENet w/o SIE** in two datasets by 1.3% and 0.7% of accuracy respectively, which demonstrates that the SIE layer can exert significant effects after integrated with the IIF layer.

(3) **FENet-BERT** remarkably improves the performance of **FENet-Glove**, which reflects the powerful embedding capability of BERT.

5 Case Study

Figure 3 shows a example of visual and textual attention visualization. We use the first feature map of image and the first token of sentence as attention query,

respectively. With the help of interactive fine-grained attention mechanisms, the model can successfully focus on appropriate regions based on the associated sentences and pay more attention to the relevant tokens. For example, Fig. 3(a) depicts a traffic accident, and the corresponding text describes the casualties. As shown in Fig. 3(b), our model pay more attention to the head and seat of broken car according to the sentence context. Also, based on the accident image, the important words such as "serious" and "injury" have greater attention weight in Fig. 3(c). Thus, our model correctly catches the important parts of text and image, and predicts the sentiment of this sample as negative.

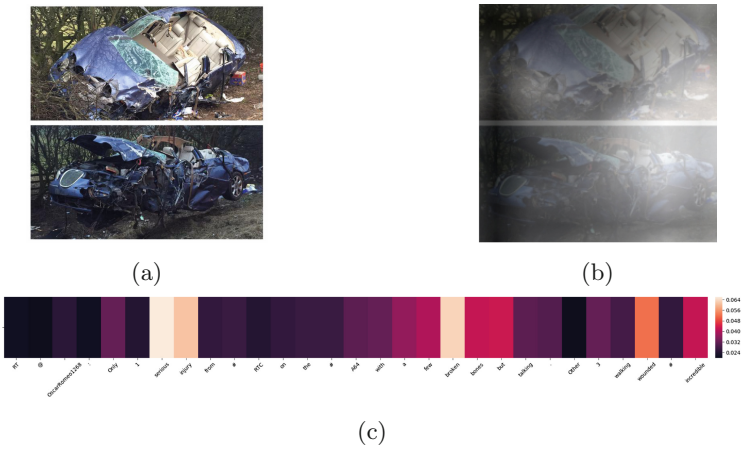


Fig. 3. An example of visual and textual attention. (a) An example image. The corresponding text of the example image is: "RT @OscarRomeo1268: Only 1 serious injury from #RTC on the #A64 with a few broken bones but talking. Other 3 walking wounded #incredible." (b) Visual attention visualization. (c) Textual attention visualization.

6 Conclusion

This paper proposes FENet for sentiment analysis in multimodal social media. Compared with the previous works, we employ a fine-grained attention mechanism to effectively extract the relationship and influence between text and image. Besides, we explore a new approach based on gated convolution mechanisms to extract and leverage visual and textual information for sentiment prediction. The experimental results on two datasets demonstrate that our proposed model outperforms the existing state-of-the-art methods.

Acknowledgement. This work is supported by the National Key R&D Program of China (2018AAA0101203), and the National Natural Science Foundation of China (61673403, U1611262).

References

1. Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., Hussain, A.: Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cogn. Comput. **7**(4), 487–499 (2015). https://doi.org/10.1007/s12559-014-9316-6
2. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: In ACM MM. Citeseer (2013)
3. Cai, G., Xia, B.: Convolutional neural networks for multimedia sentiment analysis. In: Li, J., Ji, H., Zhao, D., Feng, Y. (eds.) NLPCC-2015. LNCS (LNAI), vol. 9362, pp. 159–167. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25207-0_14
4. Cambria, E., Poria, S., Bajpai, R., Schuller, B.: SenticNet 4: a semantic resource for sentiment analysis based on conceptual primitives. In: COLING (2016)
5. Cambria, E., Poria, S., Hazarika, D., Kwok, K.: SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In: AAAI (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
7. Fan, F., Feng, Y., Zhao, D.: Multi-grained attention network for aspect-level sentiment classification. In: EMNLP, pp. 3433–3442 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
10. Johnson, R., Zhang, T.: Semi-supervised convolutional neural networks for text categorization via region embedding. In: NIPS, pp. 919–927 (2015)
11. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: ACL, vol. 1, pp. 655–665 (2014)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
13. Li, Z., Wei, Y., Zhang, Y., Yang, Q.: Hierarchical attention transfer network for cross-domain sentiment classification. In: AAAI (2018)
14. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
15. Niu, T., Zhu, S., Pang, L., El Saddik, A.: Sentiment analysis on multi-view social data. In: Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., Liu, X. (eds.) MMM 2016. LNCS, vol. 9517, pp. 15–27. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27674-8_2
16. Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text with word clusters. In: NAACL-HLT, pp. 380–390 (2013)
17. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
18. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
19. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR (2015)
20. Truong, Q.T., Lauw, H.W.: VistaNet: visual aspect attention network for multimodal sentiment analysis. In: AAAI (2019)
21. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: EMNLP (2005)

22. Xu, N., Mao, W.: MultiSentiNet: a deep semantic network for multimodal sentiment analysis. In: CIKM, pp. 2399–2402. ACM (2017)
23. Xu, N., Mao, W., Chen, G.: A co-memory network for multimodal sentiment analysis. In: SIGIR, pp. 929–932. ACM (2018)
24. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. In: AAAI (2019)
25. Xue, W., Li, T.: Aspect based sentiment analysis with gated convolutional networks. In: ACL, pp. 2514–2523 (2018)
26. You, Q., Jin, H., Luo, J.: Visual sentiment analysis by attending on local image regions. In: AAAI (2017)
27. Yu, Y., Lin, H., Meng, J., Zhao, Z.: Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms **9**(2), 41 (2016)
28. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: EMNLP, pp. 1103–1114 (2017)