



# Gated attention fusion network for multimodal sentiment classification

Yongping Du<sup>\*</sup>, Yang Liu, Zhi Peng, Xingnan Jin

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

## ARTICLE INFO

### Article history:

Received 29 April 2021

Received in revised form 27 December 2021

Accepted 30 December 2021

Available online 5 January 2022

### Keywords:

Multimodal sentiment classification

Gated attention mechanism

Convolutional neural network

Feature fusion

## ABSTRACT

Sentiment classification can explore the opinions expressed by people and help them make better decisions. With the increasing of multimodal contents on the web, such as text, image, audio and video, how to make full use of them is important in many tasks, including sentiment classification. This paper focuses on the text and image. Previous work cannot capture the fine-grained features of images, and those models bring a lot of noise during feature fusion. In this work, we propose a novel multimodal sentiment classification model based on gated attention mechanism. The image feature is used to emphasize the text segment by the attention mechanism and it allows the model to focus on the text that affects the sentiment polarity. Moreover, the gating mechanism enables the model to retain useful image information while ignoring the noise introduced during the fusion of image and text. The experiment results on Yelp multimodal dataset show that our model outperforms the previous SOTA model. And the ablation experiment results further prove the effectiveness of different strategies in the proposed model.

© 2022 Elsevier B.V. All rights reserved.

## 1. Introduction

Sentiment is an important factor affecting people's decision-making. Sentiment classification aims to automatically determine the sentiment polarity (e.g., positive, negative or neutral) of massive opinions and reviews emerging on the Internet. It is one of the important tasks in the field of natural language processing.

Cambria [1] describes the related technologies of sentiment analysis in detail, and introduces related problems and solutions. Driven by the development of Internet, documents on the Web (e.g., reviews, blog posts, tweets) are increasingly multimodal in nature, including text, image, audio and video, which promotes the research of multimodal sentiment classification. We focus on text and image in this paper.

As the textual sentiment classification has made great progress, more research works focus on the text–image multimodal sentiment classification task and there are two challenging problems. Firstly, how to extract valuable information from the images. Secondly, how to fuse image features with textual features for better representation. Previous works usually use convolutional neural networks to extract features of the entire image, and then merge them with textual features by concatenation or attention mechanism. However, these methods do not make good use of the fine-grained features in the images and the

complementary relationship between images and text. Moreover, these models do not take into account the negative effects of noise images on sentiment classification. This paper proposes a novel text–image multimodal sentiment classification model based on the gated attention mechanism, which resolves the above problems well. It uses a convolutional neural network pre-trained on the large scale data to extract the fine-grained features of the entity in the image. More importantly, the gated attention mechanism is used to fuse them with text features to generate the final vector representation for sentiment classification. The gated attention mechanism can capture the text information associated with the images and give them higher weights under the premise of less image noise. Therefore, the model can make more effective use of the text–image information for sentiment classification.

In summary, the key contributions of our work are as follows:

- A convolutional neural network pre-trained on the large scale object detection dataset is used to extract the fine-grained features from images.
- Gated attention mechanism is adopted to fuse textural features and image features to get better representation and reduce the image noise.
- Our model outperforms the SOTA model VistaNet on Yelp dataset by accuracy and F1 score.

## 2. Related work

Traditional sentiment classification task usually relies on textual content. The previous feature engineering is not needed with

<sup>\*</sup> Corresponding author.

E-mail addresses: [ypdu@bjut.edu.cn](mailto:ypdu@bjut.edu.cn) (Y. Du), [yangsince97@outlook.com](mailto:yangsince97@outlook.com) (Y. Liu), [peng\\_xiaozhi@163.com](mailto:peng_xiaozhi@163.com) (Z. Peng), [jinxingnan@outlook.com](mailto:jinxingnan@outlook.com) (X. Jin).

the development of deep learning technology. The pre-trained word vectors are used for character and word representation [2, 3], and deep neural networks are adopted popularly. Kim [4] uses a multi-channel convolutional neural network to extract the n-gram features of the text and form the final sentence representation for classification. Lai et al. [5] combine the recurrent neural network with the advantage of the convolutional neural network and propose an RCNN network for text classification. It also achieves good performance on sentiment classification tasks. Yang et al. [6] believe that the importance of words in the sentence is varied, and the importance of sentences in the document is also different. The hierarchical attention network is used to classify documents and achieves better results, also the model can be applied to sentiment classification task. Basiri et al. [7] build an efficient sentiment classification network by using RNN, CNN and attention in different layers. Akhtar et al. [8] build a classifier by integrating different deep neural networks and it also achieves good results. Recently, pre-training tasks are implemented to allow models to learn knowledge from large-scale corpus and further improving the performance of downstream tasks, including sentiment classification task. Peters et al. [9] combine the word embedding and hidden layer representation of multi-layer LSTM to solve the polysemy problem of a word. Devlin et al. [10] use Transformer [11] to construct a masked language modeling task to extract features in the text. And so the representation of each word is integrated with the features of the context. The model named BERT achieves significant improvement in the downstream tasks. Yang et al. [12] construct a permutation language modeling task for pre-training, which overcomes the problem of inconsistent input of upstream and downstream tasks in the masked language model. The development of pre-training tasks also promotes text sentiment classification task for better performance. Other efforts also have been tried. Valdivia et al. [13] improve the model's performance by characterizing the boundary between positive and negative reviews. Wang et al. [14] propose a new sentiment analysis scheme, a multi-level fine-scaled sentiment sensing with ambivalence handling. It can drill deeper into the text to reveal multi-level fine-scaled sentiments and different types of emotions. Emotion recognition in conversations has also attracted significant attention. Jiao et al. [15] propose hierarchical memory network to save the current and historical information of the conversation and the model outperforms the state-of-the-art approaches with significant margins. Ghosal et al. [16] incorporate commonsense knowledge to enhance emotion recognition and achieve the best results on four benchmark conversational datasets. It alleviates the issues of difficulty in detecting emotion shifts. Li et al. [17] propose a fast, compact and parameter-efficient party-ignorant framework named bidirectional emotional recurrent unit (BiERU) for conversational sentiment analysis. The results show that BiERU outperforms current state-of-the-art models on three standard datasets in most cases.

In recent years, multimodal sentiment classification has attracted more and more attention. Cambria et al. [18] propose a scalable methodology for fusing multiple cognitive and affective recognition modules. Lazaridou et al. [19] merge the features extracted from multimodal data and obtain better results than a single modal. Gu et al. [20] use hierarchical attention strategy to learn multimodal representations, which improves the performance on multimodal sentiment classification task. Pham et al. [21] propose a method to learn robust representations by modal information conversion and achieve better performance. Dumpala et al. [22] propose an autoencoder for the alignment of visual and auditory information to perform multimodal sentiment classification task. Tsai et al. [23] believe that the frequency of different modal information are not same and they

cannot be completely aligned. They propose a Transformer without alignment to extract multimodal information and achieve better performance. Chaturvedi et al. [24] adopt deep learning to extract features from each modality and then project them to a common affective space. And a fuzzy logic classifier is used to predict the degree of a particular emotion in affective space. Stappen et al. [25] explore sub-symbolic representations gained from semantic concepts to gain insights into the emotional and contextual information provided by video transcriptions. Furthermore, they successfully leverage the derived features to automatically classify video segments. The performance of deep learning models is highly related to the quality of training data. Therefore, many researchers apply other techniques for multimodal emotion recognition. Li et al. [26] propose a novel framework based on reinforcement learning for pre-selecting useful images for emotion classification in facial expression recognition, and it can improve classification performance from noise data. Shu and Xu [27] propose a novel method on music emotion recognition through exploring the domain knowledge of music elements, and the experimental results on two benchmark datasets demonstrate the importance of the domain knowledge. Zhang et al. [28] propose a novel multimodal emotion recognition model for conversational videos based on reinforcement learning and domain knowledge. It achieves the state-of-the-art results on weighted average and most of the specific emotion categories.

Text-image multimodal data is also a common multimedia form, especially in product reviews. Early text-image multimodal sentiment classification task mainly uses feature engineering to build models. For example, Borth et al. [29] construct 1200 adjective-noun pairs as the basis for judgment of sentiment polarity, and then detect whether these features exist in the image, finally achieve better results than those models using text features only. With the development of deep learning, a series of multimodal sentiment classification models based on neural networks have been proposed. Yu et al. [30] use pre-trained convolutional neural networks to extract the features of text and image respectively and further combine them for sentiment classification. Xu and Mao [31] believe that the information of scene and object in the image is critical for sentiment classification task, and they use two different convolutional neural networks to extract the scene features and object features respectively, and further combine them with the text features for correct sentiment classification. Later, they consider the impact of different modal information to improve the model [32], so that different modal information can complement each other for better performance. Cai et al. [33] use convolutional neural networks to extract image attribute, and then merge them with image feature and text feature for text-image sarcasm detection task. Truong and Lauw [34] believe that the image should be used to emphasize the key part of the text instead of being input into the classifier as the feature. The model they proposed, VistaNet, uses a pre-trained VGG network [35] to extract the features of the image through the attention mechanism to emphasize some sentences in the review text, and achieves better performance.

### 3. Approach

#### 3.1. Multimodal sentiment classification task

The increasing documents in the web contain more than text and they are multimodal. While "multimodal" could refer to image, audio or video and so on, here we focus on text and image. For the text-image multimodal sentiment classification task, it is assumed to contain labeled data  $D = \{(x^i, img^i, y^i)\}_{i=1}^N$ , where  $x^i$  represents the text of the  $i$ th sample in the dataset,  $img^i$  represents the images corresponding to the text  $x^i$ ,  $y^i$  represents the sentiment label of the sample, and  $N$  represents the

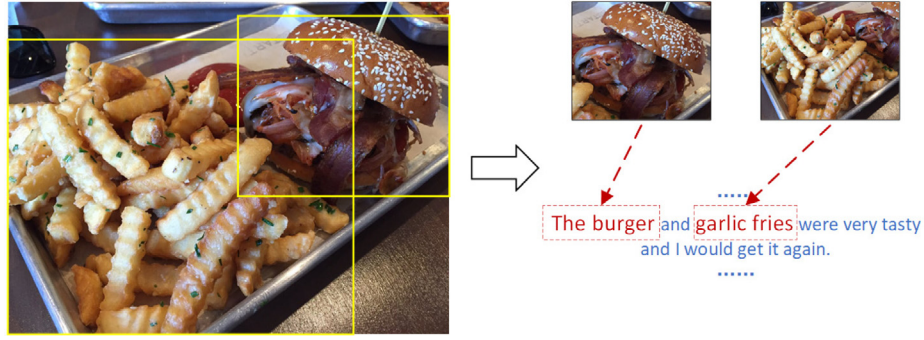


Fig. 1. A sample of text-image review dataset.

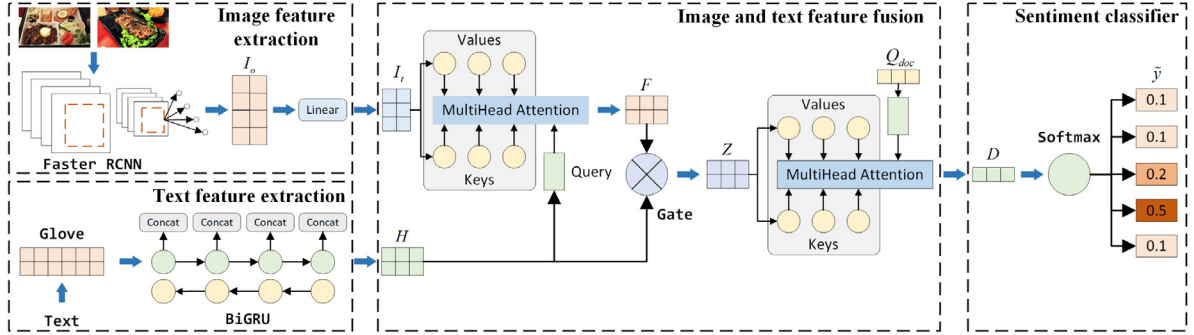


Fig. 2. Structure of Gated Attention Fusion Network.

number of samples. The goal of text-image multimodal sentiment classification task is using the existed labeled data  $D$  to train a classifier  $f$ , so that the classifier can determine the sentiment polarity of the multimodal sample containing text and images, that is,  $f(x^i, img^i) = y^i$ .

The text-image multimodal information usually appears in the reviews of various goods or services. The entities in the image often correspond to the words or phrases in the text to highlight it. As shown in Fig. 1, the “The burger” and “garlic fries” in the review appear in the image. Compared with other words, these entities often play a more important role to judge the sentiment polarity of the review.

### 3.2. Gated attention fusion network

We propose a novel text-image multimodal sentiment classification model named Gated Attention Fusion Network (GAFN). The model uses the attention mechanism to capture the text associated with the image and gives them higher weight. Moreover, the gating mechanism enables the model to retain useful image information while ignoring useless information. The structure of GAFN is shown in Fig. 2 and it contains different modules in the following.

- (1) Text feature extraction module: The input text is encoded to obtain the word embedding, and then the Bidirectional Gated Recurrent Unit (BiGRU) [36] is used to perform feature extraction to get the representation containing contextual semantic features.
- (2) Image feature extraction module: A pre-trained convolutional neural network is used to extract fine-grained feature vectors from the image.
- (3) Image and text feature fusion module: The multi-head attention mechanism is used to extract the text-related feature vectors from the image features, and then these vectors are fused with the textual features by the gating mechanism.

- (4) Sentiment classification module: The fusion feature of text and image is used as the input for sentiment polarity classification.

#### 3.2.1. Text feature extraction module

The distributed representation of words can better express the semantic information of the text. The proposed model uses the pre-trained GloVe [3] word vector to generate a distribution representation for each word in the text. Firstly, generate the matrix  $L \in R^{|V| \times d_e}$ , where  $d_e$  represents the dimension of the word vector and  $|V|$  represents the size of the vocabulary. A unique index is specified for each word in the vocabulary, and the index number corresponds to the  $L$  index number. The required word embedding can be generated for each word in the data set through the mapping relationship. In this way, the given input data  $x = \{w_1, w_2, \dots, w_n\}$  can be mapped to the embedding  $E = \{e_1, e_2, \dots, e_n\} \in R^{n \times d_e}$ .

However, the word embedding cannot resolve the word order and polysemous problem. BiGRU is used to generate the representation  $H = \{h_1, h_2, \dots, h_n\} \in R^{n \times d}$  that incorporates the semantic features of the context. The word embedding  $E$  of the review text is transformed to the new representation by BiGRU shown in Eq. (1).

$$h_i = [\vec{GRU}(e_i); \overleftarrow{GRU}(e_i)] \in R^{n \times d} \quad (1)$$

where  $h_i$  is concatenated by the hidden layer vectors of the forward GRU and the backward GRU.

#### 3.2.2. Image feature extraction module

For the images corresponding to the review text, the convolutional neural network Faster-RCNN [37] pre-trained on the large scale object detection dataset is used to extract the feature vectors of the objects. The attributes of the images, such as color, have an important impact on decide the review's sentiment polarity. But the traditional object detection dataset only contains

object labels. We use the Visual Genome dataset [38,39] to train Faster-RCNN, and it contains both the labels of the objects and the corresponding attribute labels.

After the pretraining on Visual Genome dataset, Faster-RCNN can not only find the objects in the images, but also capture their attributes. The output feature vectors contain the key information of the images. Since the output vectors' dimension of Faster-RCNN and BiGRU are different, the model transforms the output vectors of Faster-RCNN into a unified vector space with text feature vectors, as shown in Eqs. (2)–(3).

$$I_o = \text{Faster-RCNN}(\text{img}_1, \text{img}_2, \dots, \text{img}_p) \in R^{m \times d_i} \quad (2)$$

$$I_t = \tanh(W_t I_o + b_t) \in R^{m \times d} \quad (3)$$

where  $p$  represents the number of images,  $I_o$  represents the feature vectors of the images extracted by Faster-RCNN, a total number of  $m$  with the dimension  $d_i$ ,  $I_t$  represents the feature vectors after the transforming,  $W_t$  and  $b_t$  are the parameters to be trained.

### 3.2.3. Image and text feature fusion module

The obtained image feature vectors are used to enhance the text representation by the multi-head attention mechanism, instead of being treated as the feature of sentiment classifier. In detail, the multi-head attention mechanism is adopted to calculate the similarity between the image representation and the word representation, and further the feature vectors related to the text is selected from the image representation, as shown in Eqs. (4)–(11):

$$Q_j = W_{Q_j} I_x + b_{Q_j} \quad (4)$$

$$K_j = W_{K_j} I_x + b_{K_j} \quad (5)$$

$$V_j = W_{V_j} I_x + b_{V_j} \quad (6)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (7)$$

$$\text{head}_j = \text{Attention}(Q_j, K_j, V_j) \quad (8)$$

$$\begin{aligned} \text{MultiheadAttention}(Q, K, V) \\ = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_{\text{num}})W_o \end{aligned} \quad (9)$$

$$f_i = \text{MultiheadAttention}(h_i, I_t, I_t) \quad (10)$$

$$F = \{f_1, f_2, \dots, f_n\} \in R^{n \times d} \quad (11)$$

where  $W_{Q_j}$ ,  $W_{K_j}$ ,  $W_{V_j}$ ,  $b_{Q_j}$ ,  $b_{K_j}$ ,  $b_{V_j}$  are the parameters to be trained,  $\text{num}$  is the number of heads and we set  $\text{num}$  to 4,  $F$  represents the final image feature vectors related to the text. In addition, most words in the text are not associated with the image and there are also some images that have useless information due to various reasons (such as poor image quality). In order to avoid introducing noise, gating mechanism is used to control the combination of image features and text features, as shown in Eqs. (12)–(15):

$$g_i = \sigma(W_g f_i + U_g h_i + b_g) \quad (12)$$

$$l_i = g_i \circ f_i \quad (13)$$

$$z_i = \tanh(W_z l_i + U_z h_i + b_z) \quad (14)$$

**Table 1**

Statistics in Yelp restaurant review dataset.

City	Number of samples	Average number of sentences	Average word count	Number of images
Boston	2080	13.4	222.3	10743
Chicago	2165	13.5	219.0	12360
LosAngeles	24860	14.4	227.2	137920
NewYork	11425	13.4	217.5	61474
San Francisco	3775	14.8	237.3	22072

$$Z = \{z_1, z_2, \dots, z_n\} \in R^{n \times d} \quad (15)$$

where  $\sigma$  represents sigmoid function,  $W_g$ ,  $W_z$ ,  $U_g$ ,  $U_z$ ,  $b_g$ ,  $b_z$  represent the parameters to be trained and  $Z$  represents the vectors after image–text feature fusion. After training the model, it can capture the image features associated with the text while ignoring the useless image input. Finally, the multi-head attention mechanism is used to fuse all word representations into the document vector representation for classification, as shown in Eq. (16):

$$D = \text{MultiheadAttention}(Q_{\text{doc}}, Z, Z) \in R^{1 \times d} \quad (16)$$

where  $Q_{\text{doc}} \in R^{1 \times d}$  represents the parameters to be trained.

### 3.2.4. Sentiment classification module

The document representation obtained by the feature fusion module is used as the input of the softmax classifier, and the final probability distribution of each sentiment polarity is obtained, as shown in Eq. (17). The cross entropy is used as the loss function in Eq. (18).

$$\tilde{y} = \text{softmax}(W_d D + b_{\text{doc}}) \quad (17)$$

$$\text{loss} = - \sum_{k=1}^K y \log \tilde{y}(k) \quad (18)$$

where  $K$  represents the number of sentiment polarity categories,  $y$  represents the ground truth label and  $\tilde{y}(k)$  represents the probability that the sentiment polarity is predicted as the  $k$ th category.

## 4. Experiment and analysis

### 4.1. Dataset and experimental settings

Experiments are conducted on Yelp restaurant review dataset [34] which contains the text–image pairs of restaurant reviews in five cities of the United States, and it is collected from the Yelp restaurant review website. The review text is relatively long, containing multiple sentences. There are three or more images for each review mostly. The rating of the reviews in the dataset is used as the sentiment polarity label. The higher of the rating, the more satisfied the user is. The number of samples in the five categories is exactly the same. The statistical information of the dataset is shown in Table 1. There are a total number of 44305 samples and they are divided into training set, development set and test set according to the ratio of 8:1:1.

The word embedding is initialized with a 100-dimension pre-trained Glove [3] vectors. The hidden layer size of the model is set to 100, and unknown words are initialized randomly with a uniform distribution  $U[-0.01, 0.01]$ . After Faster-RCNN is pre-trained, it extracts 16–32 feature vectors from each picture as the objects' feature representation. During the training process, the model is optimized using Adam optimizer [40], which defaults to an initial learning rate of 1e-3. The batch size is set to 64. The early stopping strategy is used for regularization.



**Table 2**  
Comparison results on Yelp restaurant review dataset (Accuracy).

Model	Boston	Chicago	Los Angeles	New York	San Francisco	Total
TextCNN	0.556	0.554	0.544	0.542	0.530	0.543
TextCNN_VGG16	0.543	0.548	0.540	0.536	0.535	0.539
BiGRU	0.597	0.582	0.567	0.580	0.565	0.573
BiGRU_VGG16	0.549	0.560	0.565	0.583	0.528	0.565
HAN	<b>0.616</b>	0.585	0.576	0.571	0.530	0.573
VistaNet	0.584	0.637	<b>0.590</b>	0.591	0.551	0.589
<b>GAFN(Ours)</b>	<b>0.616</b>	<b>0.662</b>	<b>0.590</b>	<b>0.610</b>	<b>0.607</b>	<b>0.601</b>

**Table 3**  
Comparison results on Yelp restaurant review dataset (F1).

Model	Boston	Chicago	Los Angeles	New York	San Francisco	Total
TextCNN	0.548	0.548	0.538	0.533	0.525	0.537
TextCNN_VGG16	0.526	0.534	0.527	0.523	0.523	0.526
BiGRU	0.597	0.576	0.568	0.581	0.562	0.573
BiGRU_VGG16	0.551	0.562	0.568	0.584	0.531	0.568
HAN	0.607	0.574	0.571	0.545	0.525	0.568
VistaNet	0.570	0.628	0.585	0.583	0.545	0.582
<b>GAFN(Ours)</b>	<b>0.614</b>	<b>0.661</b>	<b>0.592</b>	<b>0.612</b>	<b>0.610</b>	<b>0.603</b>

#### 4.2. Comparison experiment

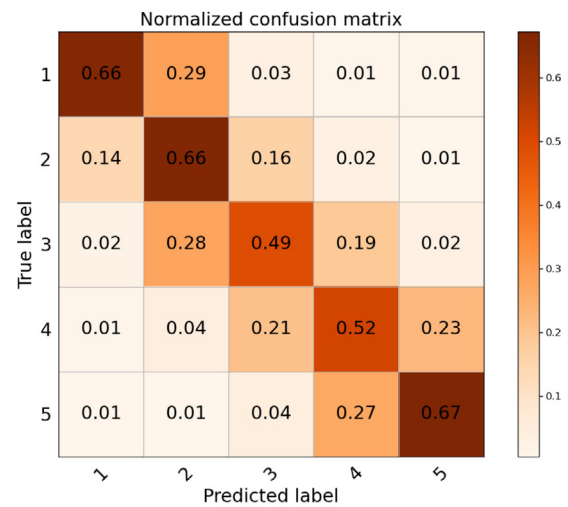
We choose the following classic methods on sentiment classification task for comparison to verify the effectiveness of the model proposed in this paper.

- TextCNN: Kim [4] uses convolutional neural network to extract text features, which can capture important information in the text to guide sentiment polarity prediction. And further TextCNN\_VGG16 uses the VGG16 [35] network to extract the images' features representation and concatenate it with the text representation for classification.
- BiGRU: The use of gating mechanism to solve the long-distance dependence problem of sequence modeling which can generate higher quality text representation. BiGRU\_VGG16 also uses the VGG16 network to extract the images' features representation, and concatenate it with the text representation for classification.
- HAN: A hierarchical attention network proposed by Yang et al. [6]. It considers the importance of different words in sentence, and the importance of different sentences in the document, and finally generates a document-level representation of the text.
- VistaNet: Truong and Lauw [34] propose a multimodal sentiment classification network based on HAN, which uses visual feature to weight sentence representation.
- GAFN(Ours): The model proposed by this paper fuses image and text information by a gated attention mechanism, which can not only make full use of multimodal information but also alleviate the impact of the noise image.

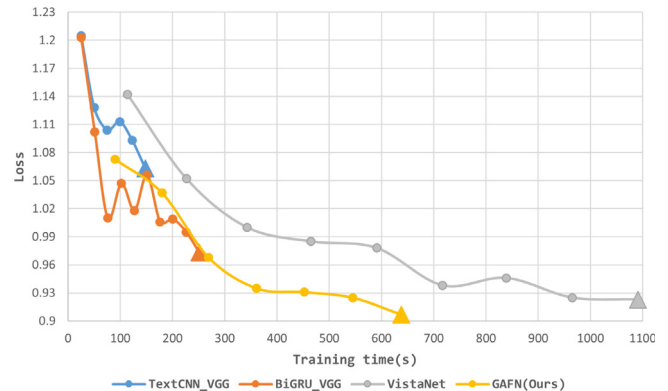
The sentiment classification performance of the above baseline methods and our proposed model are compared on the dataset. It is evaluated by accuracy and F1 values. The experimental results are shown in Tables 2 and 3.

It can be seen that the accuracy of the proposed model GAFN performs best on different datasets. It also gets the best results on all datasets from different cities by F1 value. Especially on the San Francisco data, it achieves nearly 12% improvement than the result of the previous SOTA model VistaNet. The confusion matrix of GAFN in Total data is shown in Fig. 3 and it shows the model's prediction performance on different ratings.

Compared with the VistaNet model, our model does not use a hierarchical attention structure and so it is more efficient and has



**Fig. 3.** The confusion matrix on Yelp restaurant review dataset(Total).



**Fig. 4.** Comparison of model training time and Loss value change.

the faster training speed. The loss value with training time change on the development set is shown in Fig. 4. The triangle nodes in the figure indicate that the model has reached the optimal state under the early stopping strategy. It can be seen that the single-epoch training time of Text\_CNN\_VGG16 and BiGRU\_VGG16 is




**Table 4**  
Comparison results in ablation experiments (Accuracy).

Model	Boston	Chicago	Los Angeles	New York	San Francisco	Total
<b>GAFN(Full Model)</b>	<b>0.616</b>	<b>0.662</b>	<b>0.590</b>	<b>0.610</b>	<b>0.607</b>	<b>0.601</b>
-Gate	0.603	0.585	0.594	0.594	0.570	0.592
-Attn	0.587	0.575	0.576	0.578	0.563	0.576
-Gate-Attn-RCNN	0.597	0.582	0.567	0.580	0.565	0.573
-BiGRU+ELECTRA	0.537	0.560	0.553	0.562	0.546	0.554

**Table 5**  
Comparison results in ablation experiments (F1).

Model	Boston	Chicago	Los Angeles	New York	San Francisco	Total
<b>GAFN(Full Model)</b>	<b>0.614</b>	<b>0.661</b>	<b>0.592</b>	<b>0.612</b>	<b>0.610</b>	<b>0.603</b>
-Gate	0.594	0.570	0.588	0.586	0.561	0.585
-Attn	0.580	0.567	0.574	0.574	0.562	0.573
-Gate-Attn-RCNN	0.597	0.576	0.567	0.581	0.562	0.573
-BiGRU+ELECTRA	0.544	0.567	0.559	0.568	0.551	0.560

**Table 6**  
Several image-text samples in the dataset.

Image	Text	Gold Label	Predicted Label
	...The <b>steak</b> was good but considering ...	<b>2</b>	<b>2</b>
	...This place has really good chicken <b>wings</b> . ...	<b>4</b>	<b>4</b>
	...We also had the <b>tiramisu</b> which the portion was...	<b>4</b>	<b>4</b>

short. And they reach the optimal state quickly, but their final loss value is still high. The proposed GAFN model has achieved better results in terms of single-epoch training time, number of convergence epochs, and Loss value than VistaNet which uses a hierarchical attention structure.

#### 4.3. Ablation experiment

In order to verify the effectiveness of different strategies in the proposed model, we conduct ablation experiments and the results are shown in [Tables 4 and 5](#).

- GAFN(Full Model): The complete gated attention mechanism model proposed in this paper.
- -Gate: The features extracted by Faster-RCNN are integrated with the text feature vectors by the multi-head attention mechanism. But no gating mechanism is used.
- -Attn: The gating mechanism is used to control the combination of the features extracted by Faster-RCNN and the text feature vector. But no attention mechanism is used.
- -Gate-Attn-RCNN: The model using only text information and no image information.
- -BiGRU+ELECTRA: Replace the GAFN's text feature extraction module by ELECTRA-small [41].

The full model achieves the best results in the ablation experiment. Using image features to emphasize the text segment by the attention mechanism allows the model to focus on the text that affects the sentiment polarity. More importantly, the gating mechanism can alleviate the problem of noise introduced during the fusion of image and text, so that the model can extract valuable image information.

#### 4.4. Image text association sample

Several samples in the test data set are shown in [Table 6](#). The object in the yellow box in the figure is identified by Faster-RCNN, which has the highest attention score computed by multi-head attention and it is associated with the entity in the review text.

For the first example in [Table 6](#), the image-text attention score given by our model shows that the “steak” in the image has the highest correlation with the word “steak” in the review text. For the second sample in [Table 6](#), the word “wings” in the text has the highest correlation with the “chicken wings” in the image. And the last sample in [Table 6](#), the relationship between “tiramisu” in the text and the pastry in the picture is also captured by the model. These samples show that the image attention mechanism adopted in this paper can effectively capture the object in image associated with the text.

nice environment , good service . loved fried jumbo shrimp ! tried steak tartare and tuna steak for the first time , should have ordered just medium for the tuna though , but still , everything was delicious .

Fig. 5. A sample of text attention visualization.

#### 4.5. Text attention visualization

In addition, the visualized result of the attention for a given review text is shown in Fig. 5. It can be seen that the model pays great attention to the words highly related to the final positive prediction, like ‘good’ and ‘delicious’.

## 5. Conclusion

This paper proposes a novel text–image multimodal sentiment classification model based on the gated attention mechanism. The pre-trained convolutional neural network is used to extract the fine-grained feature vectors of the objects in the image. And the gated attention mechanism is adopted to fuse the image and text representation, which can capture the text segments associated with the entities in the image, so as to give indication for predicting the sentiment polarity. The comparative experiments prove that our model achieves better results than other related models on the Yelp multimodal datasets. The ablation experiments further verify the effectiveness of different strategies in the model.

In the future work, we will try to capture more valuable information in the image for building the relationship with the text at multiple levels, and further improve the performance of sentiment classification.

## CRediT authorship contribution statement

**Yongping Du:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Yang Liu:** Methodology, Data curation, Software, Writing – Original draft preparation. **Zhi Peng:** Visualization, Resources. **Xingnan Jin:** Data Curation, Software.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work is supported by Beijing Natural Science Foundation, China under grant NO. 4212013 and National Key R&D Program of China under grant NO. 2019YFC1906002.

## References

- [1] E. Cambria, *Affective computing and sentiment analysis*, IEEE Intell. Syst. 31 (2) (2016) 102–107.
- [2] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st International Conference On Learning Representations, ICLR, Scottsdale, Arizona, USA, Workshop Track Proceedings, 2013.
- [3] J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in: Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
- [4] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings Of The Conference On Empirical Methods In Natural Language Processing, EMNLP, Doha, Qatar, 2014, pp. 1746–1751.
- [5] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Proceedings Of The Twenty-Ninth AAAI Conference On Artificial Intelligence, 2015, pp. 2267–2273.
- [6] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings Of The Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, San Diego, California, 2016, pp. 1480–1489.
- [7] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis, Future Gener. Comput. Syst. 115 (2021) 279–294.
- [8] M.S. Akhtar, A. Ekbal, E. Cambria, How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble, IEEE Comput. Intell. Mag. 15 (1) (2020) 64–75.
- [9] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings Of The Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, 2018, pp. 2227–2237.
- [10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings Of The Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers), Minneapolis, Minnesota, 2019, pp. 4171–4186.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
- [12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, 2019.
- [13] A. Valdivia, M.V. Luzón, E. Cambria, F. Herrera, Consensus vote models for detecting and filtering neutrality in sentiment analysis, Inf. Fusion 44 (2018) 126–135.
- [14] Z. Wang, S. Ho, E. Cambria, Multi-level fine-scaled sentiment sensing with ambivalence handling, Int. J. Uncertain. Fuzziness Knowl. Based Syst. 28 (4) (2020) 683–697.
- [15] W. Jiao, M. Lyu, I. King, Real-time emotion recognition via attention gated hierarchical memory network, Proc. Assoc. Adv. Artif. Intell. Conf. Artif. Intell. 34 (05) (2020) 8002–8009.
- [16] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, COSMIC: Commonsense knowledge for emotion identification in conversations, in: Findings Of The Association For Computational Linguistics: EMNLP, 2020, pp. 2470–2481.
- [17] W. Li, W. Shao, S. Ji, E. Cambria, BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82.
- [18] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics, in: IEEE Symposium On Computational Intelligence For Human-Like Intelligence, 2013, pp. 108–117.
- [19] A. Lazaridou, N.T. Pham, M. Baroni, Combining language and vision with a multimodal skip-gram model, in: Proceedings Of The Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Denver, Colorado, 2015, pp. 153–163.
- [20] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, I. Marsic, Multimodal affective analysis using hierarchical attention strategy with word-level alignment, in: Proceedings Of The 56th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 2225–2235.
- [21] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings Of The Association For The Advance Of Artificial Intelligence Conference On Artificial Intelligence, vol. 33, no. 01, 2019, pp. 6892–6899.
- [22] S.H. Dumpala, I. Sheikh, R. Chakraborty, S.K. Kopparapu, Audio-visual fusion for sentiment classification using cross-modal autoencoder, in: Proc. Neural Inf. Process. Syst., NIPS, 2019, pp. 1–4.
- [23] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings Of The 57th Annual Meeting Of The Association For Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569.
- [24] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.
- [25] L. Stappen, A. Baird, E. Cambria, B.W. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95.

- [26] H. Li, H. Xu, Deep reinforcement learning for robust emotional classification in facial expression recognition, *Knowl.-Based Syst.* 204 (2020) 106172.
- [27] Y. Shu, G. Xu, Emotion recognition from music enhanced by domain knowledge, in: *The Pacific Rim International Conference On Artificial Intelligence 2019: Trends In Artificial Intelligence*, 2019, pp. 121–134.
- [28] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, *IEEE Trans. Circuits Syst. Video Technol.* (2021) 1.
- [29] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: *Proceedings Of The 21st ACM International Conference On Multimedia*, New York, NY, USA, 2013, pp. 223–232.
- [30] Y. Yu, H. Lin, J. Meng, Z. Zhao, Visual and textual sentiment analysis of a microblog using deep convolutional neural networks, *Algorithms* 9 (2) (2016).
- [31] N. Xu, W. Mao, MultiSentiNet: A deep semantic network for multimodal sentiment analysis, in: *Proceedings Of The ACM On Conference On Information And Knowledge Management*, New York, NY, USA, 2017, pp. 2399–2402.
- [32] N. Xu, W. Mao, G. Chen, A co-memory network for multimodal sentiment analysis, in: *The 41st International ACM SIGIR Conference On Research & Development In Information Retrieval*, New York, NY, USA, 2018, pp. 929–932.
- [33] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in Twitter with hierarchical fusion model, in: *Proceedings Of The 57th Annual Meeting Of The Association For Computational Linguistics*, Florence, Italy, 2019, pp. 2506–2515.
- [34] Q.-T. Truong, H.W. Lauw, VistaNet: Visual aspect attention network for multimodal sentiment analysis, *Proceedings Of The Association For The Advance Of Artificial Intelligence Conference On Artificial Intelligence* (2019) 305–312.
- [35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *3rd International Conference On Learning Representations, ICLR, Conference Track Proceedings*, 2015.
- [36] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: *Proceedings Of The Conference On Empirical Methods In Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734.
- [37] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [38] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. Shamma, M. Bernstein, F.-F. Li, Visual genome: Connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vis.* 123 (2017) 32–73.
- [39] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: *IEEE/CVF Conference On Computer Vision And Pattern Recognition*, 2018, pp. 6077–6086.
- [40] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), *3rd International Conference On Learning Representations, ICLR, Conference Track Proceedings*, 2015.
- [41] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: *8rd International Conference On Learning Representations, ICLR*, 2020.