



## Article

# Multimodal Interaction and Fused Graph Convolution Network for Sentiment Classification of Online Reviews

Dehong Zeng <sup>1</sup>, Xiaosong Chen <sup>2</sup>, Zhengxin Song <sup>2</sup>, Yun Xue <sup>1</sup> and Qianhua Cai <sup>1,\*</sup><sup>1</sup> School of Electronics and Information Engineering, South China Normal University, Foshan 528225, China; 2021022423@m.scnu.edu.cn (D.Z.); xueyun@m.scnu.edu.cn (Y.X.)<sup>2</sup> School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China; 20202332033@m.scnu.edu.cn (X.C.); 2020022258@m.scnu.edu.cn (Z.S.)

\* Correspondence: caiqianhua@m.scnu.edu.cn

**Abstract:** An increasing number of people tend to convey their opinions in different modalities. For the purpose of opinion mining, sentiment classification based on multimodal data becomes a major focus. In this work, we propose a novel **Multimodal Interactive and Fusion Graph Convolutional Network** to deal with both texts and images on the task of document-level multimodal sentiment analysis. The image caption is introduced as an auxiliary, which is aligned with the image to enhance the semantics delivery. Then, a graph is constructed with the sentences and images generated as nodes. In line with the graph learning, the long-distance dependencies can be captured while the visual noise can be filtered. Specifically, a cross-modal graph convolutional network is built for multimodal information fusion. Extensive experiments are conducted on a multimodal dataset from Yelp. Experimental results reveal that our model obtains a satisfying working performance in DLMSA tasks.

**Keywords:** document-level multimodal sentiment classification; graph convolutional networks

**MSC:** 18C50



**Citation:** Zeng, D.; Chen, X.; Song, Z.; Xue, Y.; Cai, Q. Multimodal Interaction and Fused Graph Convolution Network for Sentiment Classification of Online Reviews. *Mathematics* **2023**, *11*, 2335. <https://doi.org/10.3390/math11102335>

Academic Editor: Ivan Lorencin

Received: 6 April 2023

Revised: 14 May 2023

Accepted: 15 May 2023

Published: 17 May 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Sentiment analysis at the document level aims to identify the opinion on a main topic expressed by a whole document. Instead of understanding the sentiment at the sentence or aspect level, DLSA tends to extract the overall sentiment of the whole document. Driven by the commercial demands, document-level sentiment analysis (DLSA), on the basis of deep learning algorithms, is currently widely employed to deal with the online product reviews [1]. That is, the general sentiment toward a product or service based on an overwhelming abundance of textual data is captured directly and classified as either positive, neutral or negative [2]. As such, DLSA is capable of delivering opinions in a way that clearly facilitates the product recommendation and sales prediction [3].

Typically, the task of DLSA mainly focuses on dealing with the textual information. In line with the flourish of deep neural networks, researchers exploit a variety of methods to extract textual features and capture the context information from the document. Zhou et al. utilize a convolutional neural network (CNN) to extract a sequence of higher-level phrase representations and feed them into a long short-term memory recurrent neural network (LSTM) to obtain the sentence representation [4]. Yang et al. propose a hierarchical attention network, which aims to extract the features at both the word and sentence level to construct the document representation [5]. The DLSA models based on graph neural networks are also developed [6]. In this model, graphs for each input text are built with significant local features extracted and the memory consumption reduced.

More recently, the widespread use of smartphones has given rise to more opportunities to express opinions via different modalities (i.e., textual, acoustic and visual modalities).

On social media, the text and the image are generally taken to mutually reinforce and complement each other; see Figure 1. For this reason, there is an ongoing trend to devise document-level multimodal sentiment analysis (DLMSA) methods that tackle multimodal information. In practice, the major challenge of DLMSA models lies in aligning and fusing textual and visual information using data of distinguishing format and structure. On the task of multimodal sentiment analysis, Zadeh et al. work on computing the outer product between modalities to characterize the multimodal relevance [7]. However, this scheme greatly increases the feature vector dimension, which results in the difficulty and complexity of model training. Furthermore, recent publications report the multimodal fusion at the feature level. Truong et al. consider visual information as a source of alignment at the sentence level and assign more attention to image-related sentences [8]. In addition, Du et al. use image features to emphasize the text segment by the attention mechanism and take a gating unit to retain valuable visual information [9].



📷 3 photos

The Chinese club special soup is really good. I got the beef shank braised in soy. I didn't care for it much until I added the hot sauces. Then it was better. I 'll come back and try something else. The menu sounds cool.



**Figure 1.** An example of multimodal review.

In spite of the achievements in DLMSA, three principal limitations can be observed:

- (1) Previous work generally takes RNN and its variants to encode text, which is challenging to capture long-range contextual dependencies among sentences, especially for a large number of texts.
- (2) In most DLMSA models, images are used as a complement to texts. The alignment of textual and visual information is still limited.
- (3) The fusion of multimodal information based on the attention mechanism or gating mechanism not just fails to remove the irrelevant visual information but also introduce more noise during weighted summation.

In this work, we propose a DLMSA approach based on a cross-modal graph convolutional network to address the issues mentioned above. As such, the document is encoded to capture semantic information at both the word and sentence level. On the other hand, the global features of each image are extracted based on which the visual and textual information is aligned to enhance the visual representations. To thoroughly integrate the textual and visual features as well as capture the long-distance dependencies among sentences, an intramodal fully connected graph is constructed to establish the relation between nodes. Then, the intermodal graph is developed by setting the relation edges of text nodes to the most relevant image nodes. The multimodal information is aggregated via the encoding of a cross-modal graph and sent to sentiment classification.

The contributions of our work are summarized as follows:

- A cross-modal graph convolutional network is proposed to capture the long-range contextual dependencies of text and filter the visual noise irrelevant to the text based on which the multimodal information can be sufficiently integrated.
- The description of images is introduced into the proposed model. In this way, the alignment of images with their corresponding description is conducted through a multi-head attention mechanism.

- Experiments are carried out on datasets from Yelp.com. Experimental results verify the effectiveness of our model comparing with the state of the art.

The rest of this paper is organized as follows: Section 2 presents some related work. Section 3 describes our proposed method in detail. In Section 4, experiments and result analysis are performed. Section 5 summarizes the concluding remarks of our work.

## 2. Related Work

### 2.1. Document-Level Sentiment Analysis

Sentiment analysis is a major focus in the field of natural language processing that has gained an increasing amount of attention. Sentiment analysis determines sentiment polarity or predicts sentiment scores from a given text. With the advancement in social media, massive user-generated texts are accessible, which has further promoted the research in sentiment analysis [10].

In general, a document consists of multiple sentences. While once restricted to processing methods, development in DLSA greatly progresses with advances in deep learning algorithms. On the basis of deep neural networks, a variety of DLSA approaches are reported [11,12]. Chen et al. train a convolutional neural network (CNN), which is applied to sentence-level sentiment analysis via pre-trained word vectors, achieving a satisfying working performance [11]. Lai et al. propose an integrated model by combining the superiorities of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [13]. That is, the context information is captured via RNNs, while the document representation and the local feature of sentence are derived via CNNs. On the other hand, RNN-based methods, integrated with attention mechanisms, also have their distinctiveness in DLSA [14,15]. Specifically, the hierarchical-structure networks are the most pronounced to process on both word and sentence levels. Yang et al. establish a hierarchical attention network that, respectively, applies attention mechanisms to word and sentence levels, which fuses more valuable information into each document [5]. Huang et al. develop a hierarchical multi-attention network to accurately assign the attentive weights on distinguishing levels [16]. Huang et al. establish a hierarchical hybrid neural network with multi-head attention to extract the global and local features of each document [17]. Due to the distinguishing contribution of each sentence to the sentiment polarity, Choi et al. propose a gating-mechanism-based method to identify the sentence importance in a document [18]. So far, there is an ongoing trend to model the document based on its hierarchical structure and thus precisely extract the document feature [19,20].

### 2.2. Document-Level Multimodal Sentiment Analysis

In the multimodal sentiment analysis domain, deep-learning based methods play a pivotal role. Previous work tends to directly fuse unimodal features to construct a multimodal representation for sentiment analysis [21–23]. In [21,22], feature vectors from different modalities are concatenated for multimodal integration. Soujanya et al. extract textual and visual features using CNN, concatenate the multimodal features, and classify the sentiment polarity via a multicore learning classifier. However, such approaches fail to deal with the cross-modal interaction [23]. In [7], a TFN model is proposed to use tensor outer products to dynamically model data across modalities. This approach generally results in oversized models for training.

More recently, studies have addressed multimodal interaction and information fusion, especially by using attention mechanisms [24–27]. Amir et al. develop a multi-level attention network to extract multimodal interaction by assuming the interactions of different information between modalities [24]. Xu et al. propose a visual feature guided attention LSTM model to extract words for sentiment delivery and aggregate the representation of informative words with visual semantic features, objects and scenes [25]. Since textual and visual information reinforce and complement each other, Xu et al. construct a co-memory network to iteratively interact the textual and visual information for multimodal sentiment analysis [26]. Similarly, Zhu et al. apply an image–text interaction network for multimodal

analysis to explore the interaction between text and image regions through a cross-modal attention mechanism [27].

Notwithstanding, all the aforementioned work is carried out based on the one-to-one correspondence between text and images. While in practice, for most multimodal samples such as blog posts and e-commerce reviews, no conformity between text and image information is set in advance. For example, a single document can contain multiple images. As we know, the DLMSA is a more text-oriented task, and the image features are auxiliary for better analysis [8,28]. Instead of directly feeding images into sentiment classifiers, visual information is typically considered as a source on sentence-level alignment. Truong et al. exploit pre-trained VGG networks to obtain image features and then align the visual information as attention to each sentence, based on which more focus is assigned to image-related sentences [8]. Guo et al. leverage a set of distance-based coefficients for image and text alignment and learn sentiment representations of documents for online news sentiment classification [29]. Aiming to obtain the sentiment-related information, Du et al. propose a method based on a gated attention mechanism [9]. In this method, a pre-trained CNN is taken to extract fine-grained features of images, and then, the gated attention network is employed to fuse the image and text representations, based on which a better sentiment analysis result is achieved.

### 3. Methodology

#### 3.1. Task Definition

The DLMSA task is defined as follows: consider that a document  $d$  contains  $L$  sentences and each sentence  $s_i (i \in [1, L])$  contains words  $w_{i,t}$  with  $t \in [1, T]$ . Meanwhile,  $N$  images  $a_j = \{a_1, a_2, \dots, a_N\}, j \in [1, N]$  are attached to the document  $d$ . Notably, the sentiment of each document is labeled as  $y \in \{1, 2, \dots, C\}$  where  $C = 5$  in our work. The main purpose of our method is to predict the sentiment labels of the document based on the textual and visual information.

#### 3.2. Model Architecture

The architecture of the proposed **Multimodal Interaction and Fused Graph Convolutional Network** (MIFGCN) is presented in Figure 2. Generally, our model consists of four main modules: a text encoder, an image encoder, a cross-modal graph convolutional network module and a sentiment classifier. Specifically, three major steps are performed. To start with, a text encoder and an image encoder are employed to extract the textual features and visual features from the input, respectively. Then, the cross-modal graph convolutional network module is utilized to interact and fuse the textual and visual information, together with capturing the long-range dependencies between sentences. Lastly, the document representation is derived with the integration of visual features based on which the sentiment label is predicted.

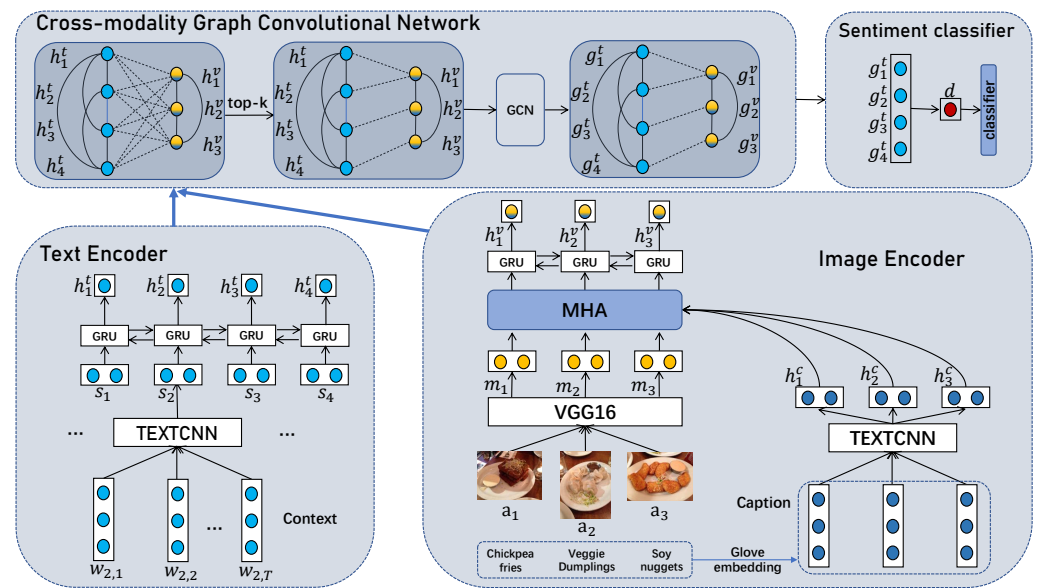
##### 3.2.1. Text Encoder

With respect to the user-generated content such as e-commerce reviews, the textual information is generally presented as a document. The simple feature extracting from the entire document fails to precisely capture the contextual information. Basically, a document consists of multiple sentences, while a sentence further consists of multiple words. Inspired by the work of [5], a hierarchical structure that deals with the semantics on both the word level and sentence level is proposed for document comprehending.

Each word is mapped into a low-dimensional vector by looking up in a pre-trained word-embedding matrix  $W_e \in \mathbb{R}^{|V| \times d_e}$  where  $|V|$  is the lexicon size and  $d_e$  is the dimension of word vector [30], i.e.,

$$e_{i,t} = w_{i,t} W_e \in \mathbb{R}^{d_e} \quad (1)$$

where  $w_{i,t} \in \mathbb{R}^{|V|}$  is a one-hot vector denoting the  $t$ -th word of the  $i$ -th sentence in the given document, with  $t \in [1, T]$  and  $i \in [1, L]$ .



**Figure 2.** Overall architecture of the proposed Multimodal Interaction and Fused Graph Convolution Network.

Then, the word vectors are taken for further processing. Following the work of Hazarika et al. [31], the convolutional neural network (CNN) is employed to extract the semantic information at the word level. Three distinguishing convolutional kernels of sizes 1, 2 and 3, with 100 feature maps, are used to distill the n-gram features of each sentence. The output is thus sent to the pooling procedure. Existing approaches typically adopt maximum pooling to tackle the semantic features. However, considering the quantity of words in every single sentence, a Top-K pooling operation is developed, which preserves the two largest values in each feature map. Subsequently, an ELU (Exponential Linear Unit) activation function and a 100-dimensional linear layer are performed for contextual information enhancing. Then, sentence vector  $x_i$  is thereby generated with the word-level features extracted and aggregated via TEXTCNN:

$$E_i = \{e_{i,1}, e_{i,2}, \dots, e_{i,T}\} \in \mathbb{R}^{T \times d_e} \quad (2)$$

$$x_i = \text{TEXTCNN}(E_i), x_i \in \mathbb{R}^{d_h} \quad (3)$$

$$X = \{x_1, x_2, \dots, x_L\} \in \mathbb{R}^{L \times d_h} \quad (4)$$

Furthermore, each sentence  $s_i$  in the document is encoded via Bi-GRU. For the sentence vector  $x_i$ , the forward hidden states vector  $\vec{h}_i$  and the backward hidden states vector  $\overleftarrow{h}_i$  are concatenated as the output hidden states  $h_i^t = [\vec{h}_i, \overleftarrow{h}_i]$ . Specifically, the information of not only  $s_i$ , but also its neighboring sentences are collected; see Equation (5):

$$H^t = \{h_1^t, h_2^t, \dots, h_L^t\} = \text{Bi-GRU}(X) \in \mathbb{R}^{L \times 2d_h} \quad (5)$$

The output  $H^t = \{h_1^t, h_2^t, \dots, h_L^t\}$  is applied to construct a fully connected graph and interact with the visual information.

### 3.2.2. Image Encoder

In most cases, a document involves more than one image. For DLMSA tasks, the image features are auxiliary for information comprehending and sentiment delivery. According to Figure 1, the images of restaurant reviews relating to food and service tend to convey a positive sentiment. In our model, the VGG-16 network [32] is employed for visual feature



extraction due to its pre-training on the large-scale dataset ImageNet [33]. Concretely, the image  $a_j$  is fed into a pre-trained VGG-16 network for encoding. The outcome of the last fully connected layer is taken as the encoding output, which is denoted as  $m_j$ :

$$m_j = \text{VGG}(a_j), m_j \in \mathbb{R}^{4096} \quad (6)$$

For the alignment of text and image, the image caption is introduced to describe the content of the image. For example, an image caption of “veggie dumplings” refers to the dumplings placed on the plate. Hence, benefiting from the textual feature extracting process, the hidden vector representation of the image caption  $H_c = \{h_1^c, h_2^c, \dots, h_N^c\} \in \mathbb{R}^{N \times 2d_h}$  can be derived, whose number equals to the image number. The image vector  $M = \{m_1, m_2, \dots, m_N\} \in \mathbb{R}^{N \times 4096}$  is sent to a fully connected layer, which we have:

$$H_m = M * W_m + b_m \quad (7)$$

where  $H_m \in \mathbb{R}^{N \times 2d_h}$  and  $W_m \in \mathbb{R}^{4096 \times 2d_h}$ .

Then, instead of the cross-modal concatenation, the multi-head attention mechanism is carried out for alignment:

$$\text{head}_i = \text{softmax}\left(\left(H_c W^Q\right) \times \left(H_m W^K\right)^T\right) \times \left(H_m W^V\right) \quad (8)$$

$$d_{\text{head}} = \frac{2d_h}{n_{\text{head}}} \quad (9)$$

$$H_{att}^v = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_{n_{\text{head}}}) \quad (10)$$

where  $W^Q$ ,  $W^K$  and  $W^V \in \mathbb{R}^{2d_h \times d_{\text{head}}}$  are linear-layer weight matrices;  $d_{\text{head}}$  is the dimension of the attention layer and  $n_{\text{head}}$  is a hyperparameter indicating the attention head number.

The image and the text are aligned to obtain the semantic information of images. Notably, there exist internal connections of images attached to the same document, which benefits the comprehending of textual content. In light of essential connections among images, a Bi-GRU is taken to obtain the enhanced image representation based on the aligned image vector  $H_{att}^v$ , which is:

$$H^v = \{h_1^v, h_2^v, \dots, h_N^v\} = \text{Bi-GRU}(H_{att}^v) \in \mathbb{R}^{N \times 2d_h} \quad (11)$$

### 3.2.3. Cross-Modal Graph Convolutional Network Module

In order to capture long-range dependencies between sentences and fuse the multi-modal information, we propose a cross-modal graph convolutional network on the task of DLMSA. A fully connected graph is first established using hidden state vectors of text. Then, the attention mechanism is combined with the top-k approach for text and image alignment, which aims to construct a cross-modal graph. The cross-modal graph is fed into a multilayer graph convolutional network for textual and visual information fusion.

#### Graph construction

To represent a document with  $L$  sentences and  $N$  images, we use a graph where nodes correspond to both sentences and images. The edges in the graph depict long-range dependencies between sentences, relationships between images, and interactions between modalities.

**Node construction:** Each sentence  $h_i^t$  and each image  $h_j^v$  are characterized by nodes within the graph. As mentioned above, the document  $d$  contains  $L$  sentences and  $N$  images. Thus, the number of nodes for the given document is  $L + N$ .

**Edge construction:** Different sentences have certain connections with each other to convey sentiment information. To precisely model their relationship, the nodes of two

sentences are established with an edge, based on which the long-range dependencies can be captured. Specifically, we employ the attention mechanism to obtain the semantic relation between sentences. Then, the attention weight is derived and used as an edge weight for the node connection, which is given by:

$$A^t = \text{softmax}\left(\left(H^t W_t^Q\right) \times \left(H^t W_t^K\right)^T\right) \quad (12)$$

where  $W_t^Q$  and  $W_t^K \in \mathbb{R}^{2d_h \times d_{att}}$  are linear-layer weight matrices.

The greater the edge weight between nodes that is computed, the greater the importance determined between them.

In such a manner, the relation of image nodes can also be built:

$$A^v = \text{softmax}\left(\left(H^v W_v^Q\right) \times \left(H^v W_v^K\right)^T\right) \quad (13)$$

where  $W_v^Q$  and  $W_v^K \in \mathbb{R}^{2d_h \times d_{att}}$  are linear-layer weight matrices.

Assume now that each sentence involves one most relevant image; then, the visual information from other images can be filtered.

For nodes of different modalities, we tend to filter the noise of irrelevant images during the edge construction procedure. In this way, for the multimodal nodes, the attention mechanism is performed to compute the relevance of each sentence toward each image. Moreover, to remove the interference from unrelated images, we apply the top-k approach to maintain the most relevant image of each sentence as its neighboring node. Accordingly, the edge between the sentence and the image is established, which is used for image and text alignment. The attention weight as the edge weight of connected nodes is computed as:

$$A^{tv} = \text{topk}\left(\text{softmax}\left(\left(H^t W_{tv}^Q\right) \times \left(H^v W_{tv}^K\right)^T\right)\right) \quad (14)$$

$$A^{vt} = (A^{tv})^T \quad (15)$$

where both  $W_{tv}^Q$  and  $W_{tv}^K \in \mathbb{R}^{2d_h \times d_{att}}$  are linear-layer weight matrices.

Apparently, there are three categories of edges in the graph: (1) unimodal edges connecting text nodes, (2) unimodal edges connecting image nodes, and (3) cross-modal edges connecting text and image nodes. In order to distinguish the three relationships denoted by the edges, the edge-weighting is performed as:

$$A(i, j) = \begin{cases} \mu \times A^t & \text{if } i < L, j < L \\ \gamma \times A^v & \text{if } i \geq L, j \geq L \\ A^{tv} & \text{if } i < L, j \geq L \\ A^{vt} & \text{if } i \geq L, j < L \end{cases} \quad (16)$$

where  $\mu$  and  $\gamma$  are hyperparameters and  $A$  is the adjacency matrix between the nodes.

### Graph learning

During the multimodal feature extraction, to stack more graph convolutional layers to distill information from higher-order neighbors while mitigating over-smoothing, following the work of Chen et al. [34], the deep graph convolutional network GCNII is employed for information encoding, which further fuses multimodal information. Specifically, graph convolution on the  $(l + 1)$ -th layer is defined as:

$$G^{(l+1)} = \sigma\left(\left((1 - \alpha)\tilde{P}G^{(l)} + \alpha G^{(0)}\right)\left(\left(1 - \beta^{(l)}\right)I + \beta^{(l)}W^{(l)}\right)\right) \quad (17)$$

together with

$$G^{(0)} = [H^t, H^v] \quad (18)$$

and

$$\tilde{P} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} = (D + I)^{-1/2} (A + I) (D + I)^{-1/2} \quad (19)$$

where  $\tilde{P}$  denotes the normalized graph Laplacian matrix;  $D$  is the pairwise degree matrix;  $I$  is the unit matrix; and  $\sigma$  refers to the activation function. In addition,  $\alpha$  is the hyperparameter to control the residual connectivity, based on which the final representation of each node concerns the input features, together with alleviating the issue of over-smoothing caused by deep layers. The parameter  $\beta$  is set to ensure that the decay of weight matrix  $W^{(l)}$  adaptively adjusts with the increasing of graph convolution layers. For  $\beta = \log(\frac{\eta}{l} + 1)$ ,  $\eta$  is the hyperparameter.

### 3.2.4. Sentiment Classifier

As we know, the DLMSA is a more text-oriented task, and the image features are auxiliary for better classification. In this context, we send the textual representation  $G^t = \{g_1^t, g_2^t, \dots, g_L^t\} \in \mathbb{R}^{L \times 3d_h}$  that incorporates visual information into the sentiment classifier. A soft-attention network is employed to map  $G^t$  into the attention space via a nonlinear activation function, based on which to learn the importance  $\omega_i$  of each sentence  $g_i^t$ . The final document representation  $d$  is derived as:

$$p_i = P^T \tanh(W_g g_i^t + b_g) \quad (20)$$

$$\omega_i = \frac{\exp(p_i)}{\sum_i \exp(p_i)} \quad (21)$$

$$d = \sum_i \omega_i g_i^t \quad (22)$$

where  $P$  is a randomly initialized global attention vector,  $W_g$  is the weight of the linear layer and  $b_g$  is the bias term.

Lastly, a Softmax function is taken as a classifier to predict the sentiment label of the document:

$$\hat{y} = \text{softmax}(W_o d + b_o) \quad (23)$$

where  $\hat{y} \in \mathbb{R}^C$ .

### 3.2.5. Model Training

The model training is carried out by using cross-entropy and  $L_2$  regularization as the loss function, which is given as:

$$\mathcal{L} = - \sum_{x=1}^{\mathcal{N}} \sum_{z=1}^C y_x^z \log \hat{y}_x^z + \lambda \|\theta\|^2 \quad (24)$$

where  $\mathcal{N}$  denotes the number of training samples, and  $C$  denotes the number of sentiment labels.  $\hat{y}_x^z \in \mathbb{R}^C$  is the probability distribution of the predicted sentiment of the  $x$ -th data and  $y_x^z$  refers to its real sentiment label. Notably, the  $L_2$  regularization is carried out to prevent overfitting. The parameter  $\lambda$  is the regularization factor, while  $\theta$  stands for the set of all trainable parameters in the model.

## 4. Experiment

In this section, experiments are conducted to evaluate the working performance of MIFGCN in DLMSA tasks. The results of our model are analyzed in comparison with the baselines. Then, an ablation study is carried out to investigate the effectiveness of different components. In addition, the significance of hyperparameters is studied. We investigated the performance of the MIFGCN under different hyperparameter conditions. In addition, we analyzed the cross-modal graph convolutional network module in feature visualization.



#### 4.1. Dataset

We carry out our experiment on the restaurant reviews of a dataset from Yelp [8], which involves 44,000 reviews and 244,000 images across five cities in the US, including Boston (BO), Los Angeles (LA), Chicago (CH), New York (NY), and San Francisco (SF). Each sample contains one document and three or more images. Specifically, every single review is rated from 1 to 5 to represent each consumer's opinion. The whole dataset is further divided into a training set, a validation set and a test set. Details of the dataset are given in Table 1.

**Table 1.** Statistics of the Yelp dataset.

Dataset	Train	Valid	Test				
			BO	CH	LA	NY	SF
#Docs	35,435	2215	315	325	3730	1715	570
Avg. #Words	225	226	211	208	223	219	244
Max. #Words	1134	1145	1099	1095	1103	1080	1116
Min. #Words	10	12	14	15	12	14	10
Avg. #Images	5.54	5.35	5.25	5.60	5.43	5.52	5.69
Max. #Images	147	38	42	97	128	222	74
Min. #Images	3	3	3	3	3	3	3

#### 4.2. Experimental Settings

The initialization of all word embeddings is conducted using 300-dimensional vectors pre-trained by Glove [35]. The hidden layer of the model is set to 100. The head number of the multi-head attention network is 8. In addition, the layer number of the cross-modal graph convolutional network is 4, while the number of  $k$  is 1 in the top- $k$  module. The weights  $\mu$ ,  $\gamma$  for the different types of edges in the cross-modal graph convolutional network are designated as 0.8 and 0.1, respectively. All parameter matrices in the model are initialized by Xavier uniform distribution. Moreover, the Adam optimizer [36] is adopted during model training with the learning rate of 0.0003 and the batch size of 32. The  $L_2$  regularization factor  $\lambda$  is 0.00003. A dropout rate of 0.4 is taken to alleviate the overfitting problem.

In the dataset, each sample contains at least three images. We draw images randomly from each sample, and the number of images is fixed at three. This is because when there are more than three images, a large proportion of the documents will be excluded, resulting in a sharp reduction in the amount of data, which is about 40% in the dataset. We therefore ensured that all data, regardless of category, had the same number of images to eliminate bias in the data.

#### 4.3. Baselines

Comprehensively, we take six baseline methods to demonstrate the working performance of the proposed model:

- **Bi-GRU [37]:** A classical model that is based on bi-directional gated units for extracting word-level textual features and generating high-quality textual representations. Considering the multiple images, a pooling operation is used to aggregate visual features, which are further concatenated to textual features for sentiment classification. The average pooling and maximum pooling of the images are performed by Bi-GRU-a and Bi-GRU-m, respectively.
- **HAN [5]:** A hierarchical attention network that separately extracts word-level and sentence-level features and then generates document representation by aggregating sentence features. With respect to visual features, we also use the variants HAN-a and HAN-m to conduct average pooling and maximum pooling, respectively.
- **TFN [7]:** A tensor fusion approach that calculates the correlation of intermodal features and fuses the multimodal information. With respect to visual features, we also use

the variants TFN-a and TFN-m to conduct average pooling and maximum pooling, respectively.

- **VistaNet [8]**: A HAN-based approach that computes the sentence representation attention using visual features as QUERY to highlight sentence importance. The textual and visual fusion is carried out via weighted summation.
- **LD-MAN [29]**: A HAN-based approach that models the textual layout as visual locations, aiming to align images with corresponding text. The multimodal representation is learned via distance-based coefficients and using a multimodal attention module.
- **GAFN [9]**: A gated attention network that fuses visual and textual information to generate vector representations for sentiment classification.
- **HGLNET [38]**: A hierarchical global–local feature fusion network that fuses global features of textual and visual modalities as well as captures the fine-grained local semantic interactions between two modalities.

#### 4.4. Experimental Results and Analysis

Experimental results of **MIFGCN** and the baselines are presented in Table 2. In this experiment, we adopt the accuracy as the evaluation metric to exactly demonstrate the working performance, where Avg. represents the weighted average accuracy of the results based on the document numbers of five cities referring to Table 1. Among all the methods, our model achieves the best results on two city settings, i.e., LA and SF. The average accuracy reaches 62.29%, which outperforms the baselines.

**Table 2.** Performance comparison to baselines on the Yelp dataset (Accuracy).

Methods	BO	CH	LA	NY	SF	Avg.
TFN-a	46.35	43.69	43.91	43.79	42.81	43.89
TFN-m	48.25	47.08	46.70	46.71	47.54	46.87
Bi-GRU-a	51.23	51.33	48.99	49.55	48.60	49.32
Bi-GRU-m	53.92	53.51	52.09	52.14	51.36	52.20
HAN-a	55.18	54.88	53.11	52.96	51.98	53.16
HAN-m	56.77	57.02	55.06	54.66	53.69	55.01
VistaNet	63.81	65.74	<b>62.01</b>	61.08	60.14	61.88
LD-MAN	61.90	64.00	61.02	61.57	59.47	61.22
GAFN	61.60	66.20	59.00	61.00	60.70	60.10
HGLNET	<b>65.47</b>	<b>69.58</b>	60.78	<b>63.43</b>	60.35	62.07
<b>MIFGCN (Ours)</b>	62.86	67.38	<b>62.01</b>	61.17	<b>64.04</b>	<b>62.29</b>

Bold numbers represent the best results among methods.

The **TFN-a** and **TFN-m** methods obtain the worst results with the application of tensor fusion for cross-modal interaction. The main reason is that DLMSA is a more text-oriented task, and the images are supplementary to the document. It is challenging to capture sentiment information by simply fusing the features from both modalities through computations.

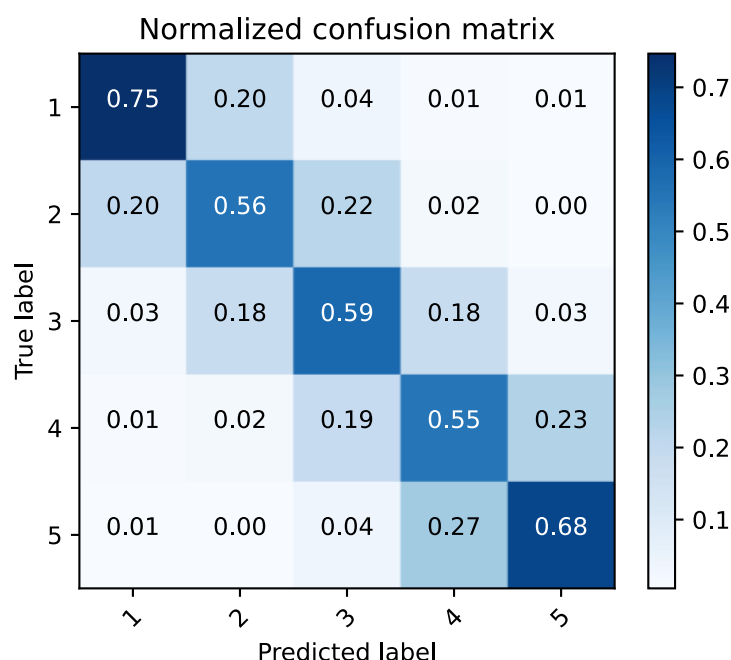
In contrast, **LD-MAN** and **VistaNet** extensively improve the results by modeling visual features as sentence attention to extract important textual information instead of distilling sentiment from images. Notably, **LD-MAN** considers both textual and visual features for sentiment polarity prediction, which leads to an inferior result than **VistaNet**.

The HAN-based models (**HAN-a** and **HAN-m**) show superiority to the Bi-GRU-based models (**Bi-GRU-a** and **Bi-GRU-m**). The HAN-based models are capable of modeling the document hierarchically, in which way both word- and sentence-level semantics are extracted. In such a manner, **LD-MAN** and **VistaNet** are also more competitive than **GAFN**. In addition, one can easily observe that the maximum pooling overperforms the average pooling in DLMSA tasks.

In comparison with the state of the art, our model achieves the best classification accuracy on average. The minimum performance gap is 0.22% against HGLNET. The main

reason is the application of GCN to the DLMSA task. For one thing, **MIFGCN** models the sentences of a document as nodes in a graph within a hierarchical structure, which not just captures the long-distance information between sentences but also enhances the sentiment delivery between sentences. For another, we also transform the images to nodes in the graph and construct a cross-modal graph convolutional network. In this way, the noises caused by irrelevant visual features can be effectively filtered. Experimental results identify the distinctiveness of our model in DLMSA, which sets a solid foundation for modeling text and images into graphs for multimodal information fusion.

The confusion matrix of **MIFGCN** over all the test data is exhibited in Figure 3. The colors on the main diagonal represent the prediction results. The closer the normalized result approaches 1, the higher the accuracy reaches, and the darker the color presents. Clearly, the proposed model obtains a satisfying performance on different labels. Specifically, the accuracies on label 1 and label 5 reach 75% and 68%, respectively, which substantially demonstrates the superiority of our model.



**Figure 3.** The confusion matrix on total test data of Yelp dataset (five cities).

#### 4.5. Ablation Study

In order to determine the importance of the different components in the proposed model, an ablation study is conducted; see Table 3.

**Table 3.** Ablation study results.

Model	BO	CH	LA	NY	SF	Avg.
<b>MIFGCN (Full Model)</b>	62.86	<b>67.38</b>	<b>62.04</b>	<b>61.17</b>	<b>64.04</b>	<b>62.29</b>
<i>w/o C</i>	<b>65.08</b>	67.08	61.39	61.05	58.77	61.53
<i>w/o V &amp; C</i>	57.46	65.54	61.96	60.58	61.05	61.49
<i>w/o G + Attn</i>	64.13	65.54	61.47	60.76	59.12	61.41
<i>w/o G</i>	62.22	66.46	60.70	60.52	61.58	61.08

Bold numbers represent the best results.

*w/o C*: The image caption for visual feature alignment is removed. The 4096-dimensional image features extracted from the pre-trained VGG network are sent to the cross-modal graph convolutional network.

*w/o  $\mathcal{V}$  &  $\mathcal{C}$* : Image processing modules are ablated. Only the textual features from the text encoder are fed into the graph convolutional network for sentiment classification.

*w/o  $\mathcal{G}$  + Attn*: The cross-modal graph convolutional network module is replaced with a multi-head attention network.

*w/o  $\mathcal{G}$* : The cross-modal graph convolutional network module is ablated from the basic model.

The results show that the most important module for our method is the cross-modal graph convolutional network module. The accuracy decreases by 1.94% with the removal of the cross-modal graph convolutional network module, indicating the significance of cross-modal interaction in DLMSA. Similarly, the replacement of the cross-modal graph convolutional network module with the multi-head attention network also results in an accuracy drop of 1.41%. Moreover, the image caption also makes a contribution to enhance the semantic representations with the alignment of visual features. Comparing the results of *w/o  $\mathcal{V}$  &  $\mathcal{C}$*  and *w/o  $\mathcal{G}$* , the graph convolutional network structure is retained in *w/o  $\mathcal{V}$  &  $\mathcal{C}$* , which shows a slightly better performance. Even if only modeling the text as a graph can the long-range dependencies between sentences be precisely captured. As a result, the accuracy of sentiment classification is thus improved, which further demonstrates the effectiveness of graphs in document modeling.

#### 4.6. Impact of GCN Layers

The effect of different GCN layers on the model performance is also investigated; see Table 4. We observe that the model performs the worst with one-layer GCN. A possible explanation is that little sentiment information is aggregated with the GCN of a small layer number. With the increasing of layer number, the working performance of our model also presents an increasing trend. The optimal number of the GCN layer is 4 for Yelp. In contrast, the continually increasing layer numbers also cause the overfitting of textual and visual features, which leads to the accuracy decline.

**Table 4.** Comparison of performance with different numbers of layers.

layers	Avg.
1	61.26
2	61.31
<b>4</b>	<b>62.29</b>
8	61.18
16	61.16
32	60.72

Bold numbers represent the best results.

#### 4.7. Impact of Top-k in Cross-Modal Graph Convolutional Network

While constructing the cross-modal graph, we take the top-k method to select the most relevant images and thus generate edges between image nodes. As such, the value of  $k$  exactly affects the performance of the proposed model. As presented in Table 5, our model obtains the lowest average accuracy with the  $k$  value of 3. Since a large proportion of documents involves three images in the dataset, the parameter  $k = 3$  indicates that all visual information is incorporated. In such a manner, the noise is introduced to confuse the sentiment prediction. By contrast, our model reaches the best performance when  $k = 1$ . Following this result, in most DLMSA samples, only one most-related image contains a large proportion of semantic-related features for sentiment delivery. By modeling images into nodes within a graph, each textual node connects to only the most relevant visual features, preventing the interference of unrelated images and enhancing the cross-modal semantics aggregation.

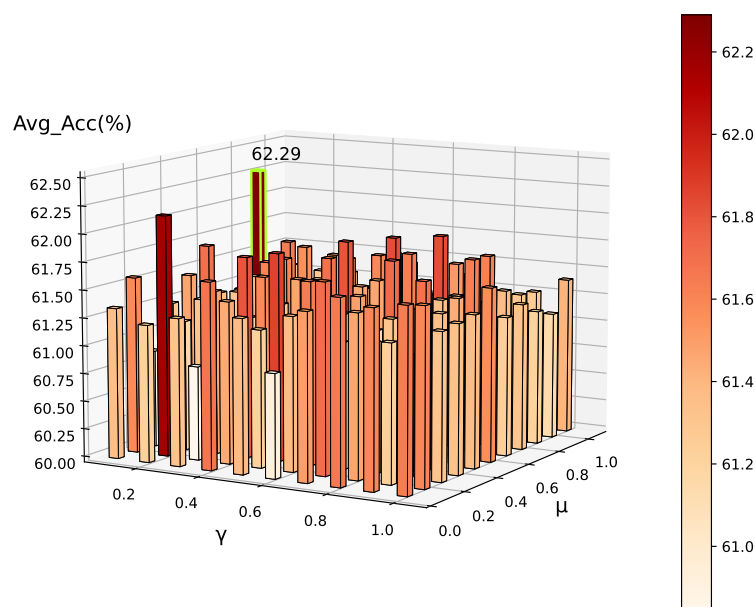
**Table 5.** Comparison of different values of  $k$  in top-k module.

$k$	Avg.
3	61.43
2	61.60
<b>1</b>	<b>62.29</b>

Bold numbers represent the best results.

#### 4.8. Hyperparameter Analysis

There are two hyperparameters in our model, i.e., the edge weights  $\mu$  and  $\gamma$  in Equation (16). The results of distinguishing hyperparameter configuration are given in Figure 4. The optimal values for  $\mu$  and  $\gamma$  are 0.8 and 0.1, respectively. Since the intramodal information has a tight connection for sentiment delivery, the application of GCN precisely captures the sentiment for further classification. Notably, the textual edges obtain higher weights than the visual edges, revealing that the text involves more with the sentiment information in DLMSA tasks.

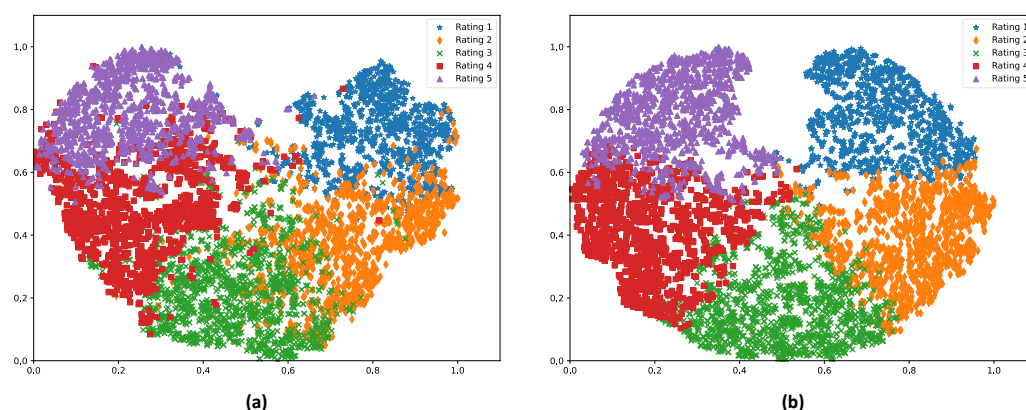
**Figure 4.** Impact of different  $\mu$  and  $\gamma$  on the performance of MIFGCN.

#### 4.9. Visualization

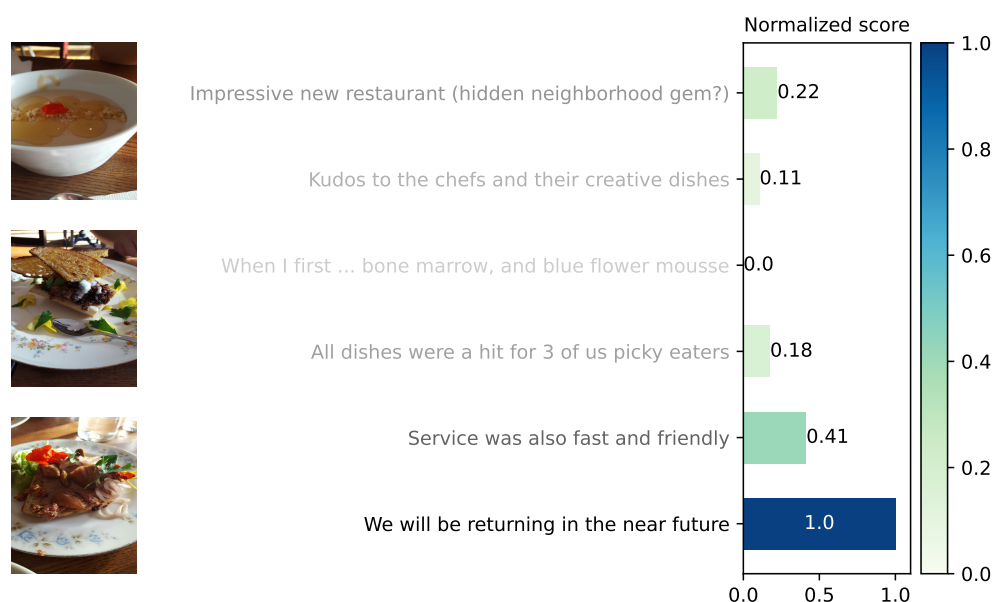
The working performance of our model is further studied by visualizing the effectiveness of the cross-modal graph convolutional network. Both the input and output textual representations from the cross-modal graph convolutional network are extracted and down-scaled into a two-dimensional space using the TSNE algorithm [39]; see Figure 5. Figure 5a illustrates the inputs to the cross-modal graph convolutional network, while Figure 5b presents the output textual representations with visual features integrated. The labels rated from 1 to 5 stand for the five sentiments of the Yelp dataset. In Figure 5a, the boundaries between the different categories are not clear. A large proportion of intersecting parts can mislead the sentiment classification. After the multilayer graph learning process, the five sentiment classes are distinguished with a longer distance, as shown in Figure 5b. Moreover, the representations of the same sentiment are aggregated with the intra-class distance decreases. For sentiment rating 2 and rating 3, the overlapping between classes is significantly reduced; so is that of rating 4 and rating 5. In this way, the cross-modal graph convolutional network module is capable of aggregating multimodal features containing sentiment information, which effectively improves the sentiment classification accuracy.

#### 4.10. Case Study

The effectiveness of our model is further validated by visualizing the attention scores of different sentences in a given document. In Figure 6, a normalized attention score of each sentence from a review rated five is presented. The darker the color is, the higher the attention assigned, and thus the larger the contribution made to the sentiment delivery. With respect to our model, the focus is not just given to the adjacent context (e.g., “Service was also fast and friendly.”) but also to the long-dependency information (e.g., “Impressive new restaurant ...”). Notably, the sentence “We will be returning in the near future” receives the highest attention score, despite the absence of intuitive sentiment words. As a consequence, our model is capable of exploiting and integrating contextual information into sentiment convey, even for long-distance dependency.



**Figure 5.** Visualization of text features projected to two-dimensional space. (a) Textual feature inputs to a cross-modal GCN. (b) Textual feature fused with visual feature output of a cross-modal GCN.



**Figure 6.** Case study. An online review rated 5 from Yelp.com

## 5. Conclusions

In this work, a novel Multimodal Interaction and Fused Graph Convolutional Network (MIFGCN) is proposed on the task of DLMSA. Our model deals with not just the texts and images but also the image captions. To start with, the image caption is aligned to the image to precisely convey the semantic information. Furthermore, the sentences and images are modeled as nodes to construct a graph. A cross-modal graph convolutional network is established to capture the long-distance contextual information and filter the visual noise,



in which way the multi-modal information can be thoroughly interacted and fused. To the best of our knowledge, this is the first model that exploits cross-modal graph convolutional network to integrate the textual and visual features for sentiment information extraction. Comparing with the state-of-the-art methods, our model achieves a competitive result on the dataset from Yelp. The effectiveness of our model and its components is validated in a variety of experiments.

Nevertheless, the proposed model still has limitations in dealing with the more fine-grained information of the image. In future work, more focus will be given to extract fine-grained information from images to facilitate the sentiment classification. Other than that, we tend to explore and exploit the external knowledge in DLMSA, which facilitates the learning of latent semantics from sentences.

**Author Contributions:** Conceptualization, D.Z. and Y.X.; methodology, D.Z.; formal analysis, D.Z. and X.C.; writing—original draft preparation, D.Z. and X.C.; writing—review and editing, D.Z. and X.C.; supervision, Q.C., Y.X. and Z.S.; funding acquisition Q.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011370, the Characteristic Innovation Projects of Guangdong Colleges and Universities (No. 2018KTSCX049), and the Science and Technology Plan Project of Guangzhou under Grant No. 202102080258.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Rhanoui, M.; Mikram, M.; Yousfi, S.; Barzali, S. A CNN-BiLSTM model for document-level sentiment analysis. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 832–847. [[CrossRef](#)]
2. Chambers, A. *Statistical Models for Text Classification and Clustering: Applications and Analysis*; University of California: Irvine, CA, USA, 2013.
3. Jiang, D.; He, J. Text semantic classification of long discourses based on neural networks with improved focal loss. *Comput. Intell. Neurosci.* **2021**, 2021. [[CrossRef](#)] [[PubMed](#)]
4. Zhou, C.; Sun, C.; Liu, Z.; Lau, F. A C-LSTM neural network for text classification. *arXiv* **2015**, arXiv:1511.08630.
5. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
6. Huang, L.; Ma, D.; Li, S.; Zhang, X.; Wang, H. Text level graph neural network for text classification. *arXiv* **2019**, arXiv:1910.02356.
7. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
8. Truong, Q.T.; Lauw, H.W. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 305–312.
9. Du, Y.; Liu, Y.; Peng, Z.; Jin, X. Gated attention fusion network for multimodal sentiment classification. *Knowl.-Based Syst.* **2022**, *240*, 108107. [[CrossRef](#)]
10. Xiong, H.; Yan, Z.; Zhao, H.; Huang, Z.; Xue, Y. Triplet Contrastive Learning for Aspect Level Sentiment Classification. *Mathematics* **2022**, *10*, 4099. [[CrossRef](#)]
11. Chen, Y. Convolutional Neural Network for Sentence Classification. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2015.
12. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
13. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
14. Liu, P.; Qiu, X.; Huang, X. Recurrent neural network for text classification with multi-task learning. *arXiv* **2016**, arXiv:1605.05101.
15. Rao, G.; Huang, W.; Feng, Z.; Cong, Q. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing* **2018**, *308*, 49–57. [[CrossRef](#)]

16. Huang, Y.; Chen, J.; Zheng, S.; Xue, Y.; Hu, X. Hierarchical multi-attention networks for document classification. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1639–1647. [\[CrossRef\]](#)
17. Huang, W.; Chen, J.; Cai, Q.; Liu, X.; Zhang, Y.; Hu, X. Hierarchical Hybrid Neural Networks With Multi-Head Attention for Document Classification. *Int. J. Data Warehous. Min. (IJDWM)* **2022**, *18*, 1–16. [\[CrossRef\]](#)
18. Choi, G.; Oh, S.; Kim, H. Improving document-level sentiment classification using importance of sentences. *Entropy* **2020**, *22*, 1336. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Sinha, K.; Dong, Y.; Cheung, J.C.K.; Ruths, D. A hierarchical neural attention-based text classifier. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 817–823.
20. Liu, F.; Zheng, J.; Zheng, L.; Chen, C. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. *Neurocomputing* **2020**, *371*, 39–50. [\[CrossRef\]](#)
21. Wang, H.; Meghawat, A.; Morency, L.P.; Xing, E.P. Select-additive learning: Improving generalization in multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 949–954.
22. Anastasopoulos, A.; Kumar, S.; Liao, H. Neural language modeling with visual features. *arXiv* **2019**, arXiv:1903.02930.
23. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 439–448.
24. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
25. Xu, N.; Mao, W. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 2399–2402.
26. Xu, N.; Mao, W.; Chen, G. A co-memory network for multimodal sentiment analysis. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 929–932.
27. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Multimodal sentiment analysis with image-text interaction network. *IEEE Trans. Multimed.* **2022**, *1*. [\[CrossRef\]](#)
28. Tian, Y.; Sun, X.; Yu, H.; Li, Y.; Fu, K. Hierarchical self-adaptation network for multimodal named entity recognition in social media. *Neurocomputing* **2021**, *439*, 12–21. [\[CrossRef\]](#)
29. Guo, W.; Zhang, Y.; Cai, X.; Meng, L.; Yang, J.; Yuan, X. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition. *IEEE Trans. Multimed.* **2020**, *23*, 1785–1798. [\[CrossRef\]](#)
30. Bengio, Y.; Ducharme, R.; Vincent, P. A neural probabilistic language model. *Adv. Neural Inf. Process. Syst.* **2000**, *13*, 1137–1155.
31. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
32. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
33. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
34. Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; Li, Y. Simple and deep graph convolutional networks. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; PMLR: Coventry, UK, 2020; pp. 1725–1735.
35. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
36. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
37. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
38. Wu, J.; Zhao, J.; Xu, J. HGLNET: A Generic Hierarchical Global-Local Feature Fusion Network for Multi-Modal Classification. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
39. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.