applied sciences

MDPI

Article

Visual Enhancement Capsule Network for Aspect-Based Multimodal Sentiment Analysis

Yifei Zhang *, Zhiqing Zhang, Shi Feng and Daling Wang

School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China
* Correspondence: zhangyifei@cse.neu.edu.cn; Tel.: +86-24-8368-7776

Abstract: Multimodal sentiment analysis, which aims to recognize the emotions expressed in multimodal data, has attracted extensive attention in both academia and industry. However, most of the current studies on user-generated reviews classify the overall sentiments of reviews and hardly consider the aspects of user expression. In addition, user-generated reviews on social media are usually dominated by short texts expressing opinions, sometimes attached with images to complement or enhance the emotion. Based on this observation, we propose a visual enhancement capsule network (VECapsNet) based on multimodal fusion for the task of aspect-based sentiment analysis. Firstly, an adaptive mask memory capsule network is designed to extract the local clustering information from opinion text. Then, an aspect-guided visual attention mechanism is constructed to obtain the image information related to the aspect phrases. Finally, a multimodal fusion module based on interactive learning is presented for multimodal sentiment classification, which takes the aspect phrases as the query vectors to continuously capture the multimodal features correlated to the affective entities in multi-round iterative learning. Otherwise, due to the limited number of multimodal aspect-based sentiment review datasets at present, we build a large-scale multimodal aspect-based sentiment dataset of Chinese restaurant reviews, called MTCom. The extensive experiments both on the single-modal and multimodal datasets demonstrate that our model can better capture the local aspect-based sentiment features and is more applicable for general multimodal user reviews than existing methods. The experimental results verify the effectiveness of our proposed VECapsNet.

Keywords: multimodal sentiment analysis; aspect-based sentiment analysis; capsule network; attention mechanism; multimodal fusion

check for updates

Citation: Zhang, Y.; Zhang, Z.; Feng, S.; Wang, D. Visual Enhancement Capsule Network for Aspect-Based Multimodal Sentiment Analysis. *Appl. Sci.* **2022**, *12*, 12146. https://doi.org/10.3390/app122312146

Academic Editor: Vincent A. Cicirello

Received: 7 November 2022
Accepted: 22 November 2022
Published: 28 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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

With the popularity of mobile internet and smartphones, more and more users are used to expressing opinions, reviewing products, or sharing experiences on social networks or e-commerce platforms. Analyzing the emotions embedded in the user-generated data has not only attracted extensive attention from the academic community [1,2], but also brought a wide range of commercial prospects, such as service supervision, game experience, satisfaction survey, product recommendation, and so on. Although much progress has been made in the sentiment analysis on user-generated data, most current research focuses on recognizing the sentimental polarity using the single modal or multiple modal features [3,4] while ignoring the rich and complementary emotional information between the multiple modalities. In addition, the existing studies have paid little attention to the task of aspect-based sentiment analysis. The task, however, offers significant guidance for practical applications. For example, separate ratings of a phone's screen, appearance, battery, price, and so on are better suited for personalized recommendations than simply assessing the phone's overall performance. The reviews expressing opinions in social networks mainly include texts, images, and a small number of short videos, and a short video can also be regarded as a collection of continuous images. Therefore, in this paper, we will study aspect-based multimodal sentiment analysis on the review data including texts and images.

Aspect-based multimodal sentiment analysis is a fine-grained task in multimodal sentiment analysis. Although aspect-based sentiment analysis on text [5–7] and multimodal sentiment analysis [8,9] have become the hot research topics in the multimedia mining community, aspect-based multimodal sentiment analysis on user-generated reviews has received little attention. Xu et al. [10] introduced the task of aspect-based multimodal sentiment analysis for the first time, and proposed a multi-interactive memory network (MIMN) to iteratively learn the text and image information for sentiment classification. Zhou et al. [11] proposed an adversarial training representation model for extracting the unified features of texts and images. In terms of multimodal feature fusion, the post-fusion method is adopted in most studies, which first extracts the features of each modality separately, and then concatenates them as the input of the sentiment classifier. This method uses the modal data independently to infer the sentimental polarity, thus ignoring the semantic correlation and the emotional interaction between different modalities. However, in practical applications, the data of a certain modality is often missing, or the aspect sentiment cannot be obtained from the specified one. According to the research of Truong and Lauw [12], text is often the main carrier to convey information, especially emotional information, and images only serve as supplementary instructions rather than as an independent information source. In this regard, we design a visual enhancement network structure to address the above shortcomings. The model enhances the textual representations using the aspect-guided image features, and interactively learns the textual and visual emotional features to improve the performance of aspect-based sentiment classification.

In this paper, we propose a novel visual enhancement capsule network (VECapsNet) for aspect-based multimodal sentiment analysis. The model consists of three components, i.e., an adaptive mask memory capsule network (MemCapsNet) for extracting the textual features associated with the affective entities, an aspect-guided capsule network (AgCapsNet) for capturing the image emotional features guided by the aspect phrases, and a multimodal fusion module for interactively learning the textual and visual capsule features and exploring the cross-modal interaction and intermodal emotional enhancement. Since no more publicly available datasets exist for multimodal emotion analysis, we built a large-scale text-image aspect sentiment dataset, named MTCom, by crawling the website of Meituan to obtain the review data of Chinese restaurants and reintegrating and annotating the obtained data using six aspect labels. The main contributions of this paper are summarized as follows:

- We present MemCapsNet and AgCapsNet structures, which use an adaptive mask memory attention and aspect-guided attention to extract the aspect emotion features of text and image on the pose matrix of the capsule, respectively.
- We design a visual enhancement network to explore the cross-modal interaction between text and image and the emotional enhancement of image to text.
- We propose a novel VECapsNet model that uses the capsule features of text and images for aspect-based multimodal sentiment analysis. The empirical results show that our model performs satisfactorily on both the Multi-ZOL dataset and our MTCom dataset.

The subsequent sections of the paper are organized as follows: Section 2 summarizes the related work, including the studies on single-modal sentiment analysis, multimodal sentiment analysis, aspect-based sentiment analysis, and multimodal fusion. Section 3 proposes a VECapsNet model for aspect-based multimodal sentiment analysis. In Section 4, we perform the experiments and analyze the experimental results. Finally, Section 5 concludes the whole paper and puts forward future work.

2. Related Work

In this section, we briefly review previous studies on single-modal sentiment analysis, multimodal sentiment analysis, aspect-based sentiment analysis, and multimodal feature fusion.

2.1. Single-Modal Sentiment Analysis

Single-modal sentiment analysis is mainly divided into text sentiment analysis and image sentiment analysis. Text sentiment analysis is to extract and recognize the implicit sentiment or emotion in text, which has been widely studied in the field of natural language processing in recent years.With the advent of deep learning, deep network models have been extensively used in text sentiment analysis for automatically learning the feature embedding representations of text. Dong et al. proposed an AdaRNN model based on a recursive neural network (RNN) to transmit the emotional information of text through a syntactic dependency tree [13]. Kim introduced a convolutional neural network (CNN) for the first time to extract the fine-grained sentiments of text [14]. Chen et al. built a novel sentiment classification model using the sentiment supplementary information of negation and inversion in text [15]. Chen et al. used a directional graph convolutional network (GCN) for joint aspect extraction and sentiment analysis [16]. Furthermore, the long short-term memory network (LSTM) has been widely applied to the tasks of text sentiment classification for capturing the context dependency of affective entities [17–20].

At the same time, great progress has been made in image sentiment analysis based on deep neural networks. Xu et al. analyzed image sentiment using a CNN for the first time [21]. Song et al. introduced a multilevel visual attention mechanism into a CNN sentiment classification [22]. Wu et al. used a weakly supervised interaction discovery network to capture the cross-spacial abstract sentiment relations [23]. Liang et al. employed deep metric networks for image sentiment analysis based on cross-domain semi-supervised learning and via heterogeneous semantics, respectively, in [24] and [25]. Text sentiment analysis and image sentiment analysis have been found to have good effects using different methods. However, the above methods are only for single-modal data, whereas most user-generated data is multimodal.

2.2. Aspect Based Sentiment Analysis

Aspect-based sentiment analysis (ABSA), as one of the more challenging subtasks in sentiment analysis, has been widely studied. Its two fundamental tasks are aspect terms extraction (ATE) and aspect sentiment classification (ASC) [26,27]. The early text-based studies mainly use bidirectional LSTM (BiLSTM) and conditional random field (CRF) to extract the opinion information through the hidden vectors [5,28,29]. Although the ability of CNN used for learning the local features of text has been verified in many studies [14,30,31], it might obtain the local features expressing different sentiments simultaneously and build the false correspondences between opinion words and emotional entities. To this end, Chen et al. [32] and Du et al. [33] applied capsule networks on ABSA, which distinguished overlapping features by clustering feature capsules. The existing studies based on capsule networks mainly weight the activation values of capsules by calculating the gated attention between different contexts and affective entities. This kind of method may lead to the following problems: (1) The input of the gated attention is from the hidden vectors and the weighted objects are the capsules on the next layer, so hidden vector correlation does not mean capsule correlation. (2) The gated attention requires that the length of the weighted capsule sequence be the same as that of the hidden sequence, which will lead to the degradation of the network into a fully connected network, thus limiting the representation ability of the high-level capsules. (3) The gated attention directly weights the activation value of a capsule without considering its credibility, and it is obviously unreasonable to assign a higher attention to a capsule with low credibility. To solve the above problems, we propose a MemCapsNet model for extracting the aspect emotional features of text, which uses a mask memory attention mechanism on the pose matrixes of capsules. Extensive experiments on the public datasets verified the effectiveness of the network.

2.3. Multimodal Sentiment Analysis

Multimodal sentiment polarity analysis and emotion prediction are two fundamental subtasks of multimodal sentimental analysis. Most of the existing end-to-end methods fol-

low the process of feature encoding–multimodal interaction–emotional fusion–sentimental prediction, in which the emotional interaction between different modalities is the core to making accurate sentiment predictions. You et al. used a consistent regression [34] and a syntactic dependency RNN with an attention mechanism [35] for sentiment classification. Xu et al. proposed several network models for multimodal sentiment analysis using image captions, image-guided attention, or iterative features of images and texts [36–38]. Chen et al. designed the weighted cross-modal attention mechanism for multimodal sentiment analysis, which captured the temporal correlation and the spatial dependence between modalities [39]. Peng et al. proposed a cross-modal complementary network with hierarchical fusion for multimodal sentiment classification [40]. In addion, more deep fusion models were used for modeling intra- and inter-modal semantical interaction [41–43]. In many multimodal tasks, unified multimodal representations were learned using attention-based adversarial networks [44] and a correlational multimodal variational autoencoder [45].

Although the current studies have achieved promising results, the aspect-based multimodal sentiment analysis still has some limitations on review data. First of all, compared with text, the emotions expressed by images are more abstract and uncertain, which may be sparse, may have nothing to do with aspects, or may reflect conflicting emotions in some aspects. Second, the learning of deep models relies heavily on large-scale training data, and most multimodal sentiment datasets are labeled by only sentimental polarity [36–38], which are not suitable for fine-grained aspect-based sentiment prediction. Therefore, in this paper, we use affective entities to guide visual feature learning in reviews, rather than simply utilizing image features for training. At the same time, we build a large-scale MTCom dataset from social networks for aspect-based multimodal sentiment analysis.

2.4. Multimodal Fusion

Multimodal fusion plays a crucial role in effectively fusing image and text features for improving the performance of emotional classification results. At present, multimodal fusion in end-to-end network models mainly adopts sharing network layers or cross-modal attention mechanisms to fuse the feature embeddings from multimodal data. Compact bilinear representations were obtained through a novel kernelized analysis of bilinear pooling for fine-grained recognition in [46]. The tensor fusion networks were employed to capture interactive features between different modalities for multimodal sentiment analysis in [42,47,48]. Huang et al. presented an attention-based modal gated network to classify sentiments, which utilized modality-gated LSTM to adaptively choose modal features with strong sentiments [49,50]. Cross-modal attention mechanisms were widely employed in multimodal sentiment analysis to capture the interactive and supplementary features between multiple modalities in many studies [5,33,39]. In this paper, we employ interactive learning to deeply fuse the modal features through a multi-round bidirectional visual enhancement architecture.

3. Methodology

In this section, we propose a visual enhancement capsule network (VECapsNet) for multimodal aspect-based sentiment analysis. The architecture of VECapsNet is shown in Figure 1, which contains three components: an adaptive mask memory capsule network (MemCapsNet) on the text modality, an aspect-guided capsule network (AgCapsNet) on the imaging modality, and a multimodal emotion fusion module base on the interactive learning. In this section, we first formulate the problem and then elaborate the framework of the proposed model, and finally design the algorithm flow for the task of multimodal aspect-based sentiment analysis.

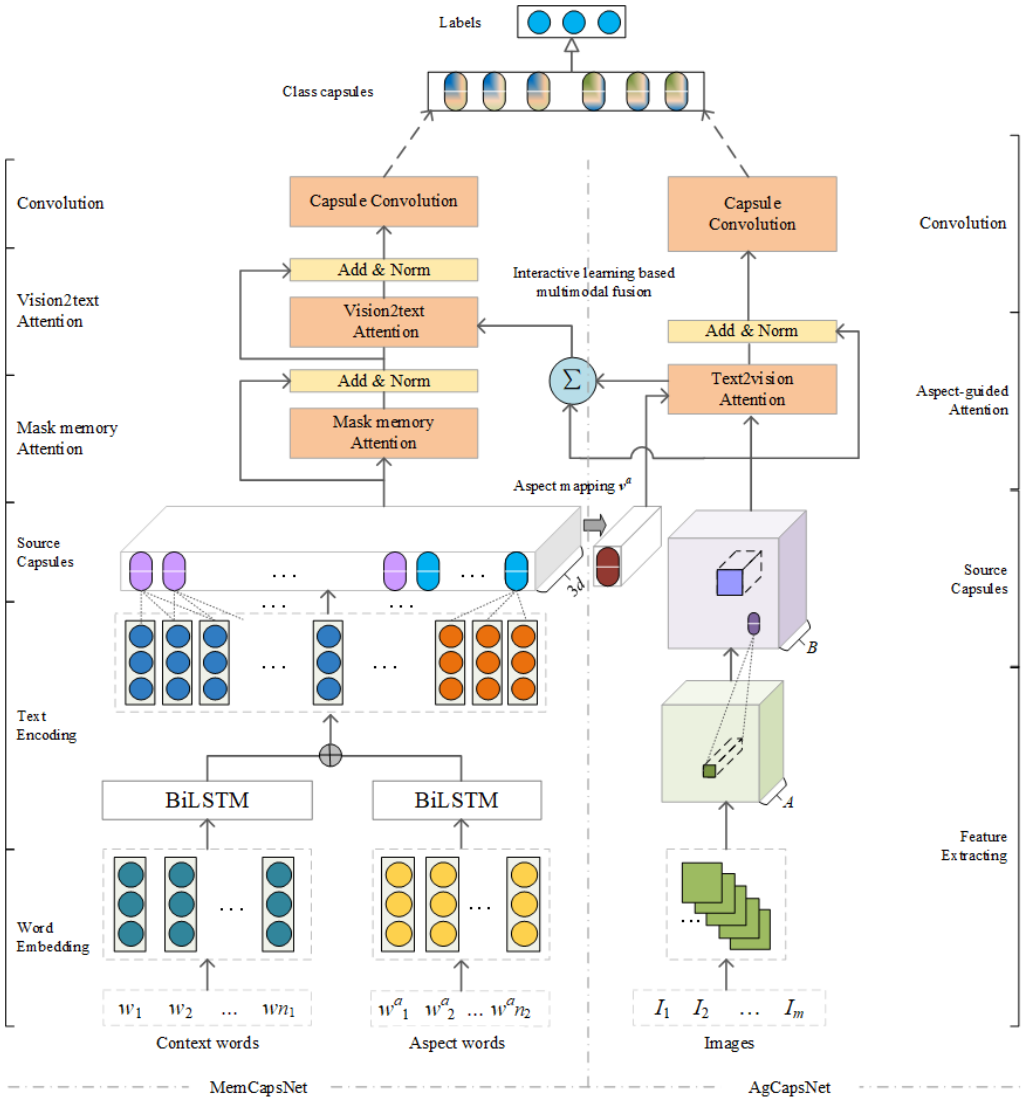

Figure 1. The framework of the VECapsNet model for multimodal aspect-based sentiment analysis. The model is composed of three parts: MemCapsNet based on text features, AgCapsNet based on image features, and the multimodal fusion module based on interactive learning.

3.1. Problem Formalization

The multimodal aspect-based sentiment analysis problem is defined as follows. Suppose that T and I represent a text sample space and an image sample space, respectively, where the ith document is $T^i = \{w_1, w_2, \ldots, w_n\}$, and its corresponding image set is $I^i = \{I_1, I_2, \ldots, I_m\}$, where n and m represent the number of the words and images, respectively. The document T^t and its corresponding image sequence I^t constitute an instance, and the instance set is expressed as $Ins = \{< T^1, I^1 >, < T^2, I^2 >, \ldots, < T^u, I^u >\}$, where u is the size of the text–image set pairs in the training set. Given an aspect set $A = \{a_1, a_2, \ldots, a_l\}$ and the aspect phrase sequence of a_t is represented as $A_t = \{w^a_1, w^a_2, \ldots, w^a_s\}$, where l is the number of aspects and s is the number of aspect phrases of a_t.

Each instance is associated with a sentiment polarity label L_s on each aspect a_t, where the sentiment label $L_s \in \{Positive, Neutral, Negative\}$. The goal of multimodal aspect-based sentiment analysis is to learn a mapping $f : T \times I \times A \rightarrow L$ from the quadruple sequence of multimodal training dataset $\{< T^i, I^i, a_j, L_s > | 1 \leq i \leq u, 1 \leq j \leq l\}$, where $T \times I \times A$ is the cartesian space consisted of three sets, T and I and A, and L is the label set composed of L_s, that is $L = \{Positive, Neutral, Negative\}$. Table 1 provides an overview of the basic notations used in the paper.

Table 1. Description of the basic notations of model.

Notation	Discription
$w_t(w_t^a)$	The tth word (aspect word)
I_t	The tth image
$T^i(I^i)$	Document (image set) included in the ith review, and $w_t \in T^i$, $I_t \in I^i$
$< T^i, I^i >$	The ith instance composed of T^i and I^i
a_t	The tth aspect
$A(A_t)$	Aspect set (aspect phrase sequence of a_t) and $a_t \in A$, $w_t^a \in A_t$
$v_t(v_t^a)$	Word vector of $w_t(w_t^a)$
L_s	A sentimental polarity label
L	A set of the labels composed of L_s
$v_t(v_t^a)$	Word vector of $w_t(w_t^a)$
$\overrightarrow{h_t}$	Feature vector of $w_t(w_t^a)$ gained by forward LSTM
$\overleftarrow{h_t}$	Feature vector of $w_t(w_t^a)$ gained by backward LSTM
h_t	Representation vector gained by Bi-LSTM
$h_t^c(h_t^a)$	Representation vector of $w_t(w_t^a)$ gained by Bi-LSTM
H	Representation matrix obtained by Bi-LSTM
k_i	Size of the ith n-gram convolutional kernel
m_j	Feature mapping matrix of the jth image
$p^{pri}(p_j^{pri})$	Capsule pose matrix of a text (the jth image) from source capsule layer
$a^{pri}(a_j^{pri})$	Capsule active value matrix of a text (the jth image) from source capsule layer
p^{att}	Attention matrix from adaptive mask attention layer
v^a	Aspect mapping capsule vector
β_j	Aspect-guided attention matrix of the jth image
$p^{txt}(p_j^{img})$	Capsule feature representation of a text (the jth image) from capsule convolution layer
$a^{txt}(a_j^{img})$	Active value capsule matrix of a text (the jth image) from capsule convolution layer
$p^{txt\_class}(p^{img\_class})$	Class capsule matrix of a text (its image set) produced by fully connected routing layer
$a^{img\_class}(a^{txt\_class})$	Class active capsule matrix of a text (its image set) produced by fully connected routing layer

3.2. Visual Enhancement Capsule Network Based on Multimodal Fusion

In order to explore the complementation and reinforcement of text and images in emotional representations, we propose a visual enhancement capsule network model (VECapsNet) for multimodal aspect-based sentiment analysis. As shown in Figure 1, the whole framework of the model is composed of three parts: MemCapsNet based on text features, AgCapsNet based on image features, and multimodal emotional fusion based on interactive learning. These three components interact with each other for the task of multimodal sentiment analysis. The details of VECapsNet will be described below.

3.2.1. MemCapsNet

The Adaptive Mask Memory Capsule Network (MemCapsNet) aims to extract text features related to affective entities, as shown in the left of Figure 1, which contains word-embedding layer, encoding layer, N-gram source capsule layer, adaptive mask attention layer, and convolution capsule layer. The following will describe further details about this architecture.

Word-embedding layer. All the words in the vocabulary are vectorized through a sharing word embedding $W_{emb} \in \mathbb{R}^{d \times |V|}$, where $|V|$ and d denote the vocabulary size and the dimension of the word vector, respectively. The embedding matrix can be initialized with random initializations [51] or pretrained models [52]. Given a candidate document $T^i = \{w_1, w_2, \ldots, w_{n_1}\}$ and its aspect phrase sequence $A_t = \{w_1^a, w_2^a, \ldots, w_{n_2}^a\}$, we can use matrix W_{emb} to obtain word vector $v_t(v_t^a) \in \mathbb{R}^d$ for each word $w_t(w_t^a)$. In order to simplify representations, we uniformly denote the word vector matrix as $E = \{v_1, v_2, \ldots, v_n\}$.

Encoding layer. The purpose of this layer is to integrate global semantic information into the embedding representation of each word of the given document T^i and its aspect

phrase sequence A_t^i in aspect a_t. The global semantic information of each word vector is captured with bidirectional LSTM (Bi-LSTM).

Give an input word embedding matrix $E = \{v_1, v_2, \ldots, v_n\}$, where v_t is the embedding vector of the tth word. The output can be computed using Equations (1)–(6).

$$i_t = \sigma(W_i[h_{t-1}, e_t] + b_i) \tag{1}$$

$$o_t = \sigma(W_o[h_{t-1}, e_t] + b_o) \tag{2}$$

$$f_t = \sigma(W_f[h_{t-1}, e_t] + b_f) \tag{3}$$

$$\widetilde{c_t} = tanh(W_c[h_{t-1}, e_t] + b_c) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

where σ is the sigmoid function, $\theta_{bi} = \{W_i, W_o, W_f, W_c, b_i, b_o, b_f, b_c\}$ is the set of training parameters of the Bi-LSTM; i, o, and f are the input gate, output gate, and forget gate, respectively; h and c are the hidden vector and the cell vector, respectively; and $\odot$ expresses the dot product operation of two vectors.

A Bi-LSTM model runs a forward and backward LSTM on a sequence starting from the left and right ends, respectively. The hidden states of v_t generated by these two LSTMs are concatenated to represent the tth word vector and its context. Hence, the vector representation of the tth word can be expressed as follows:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(h_{t-1}, v_t, \theta), \overleftarrow{h_t} = \overleftarrow{LSTM}(h_{t-1}, v_t, \theta) \tag{7}$$

$$h_t = \overrightarrow{h_t} \| \overleftarrow{h_t}, t \in \{1, 2, ..., n\} \tag{8}$$

where $\overrightarrow{h_t}, \overleftarrow{h_t} \in \mathbb{R}^k, h_t \in \mathbb{R}^{2k}$, k is the number of hidden units in LSTM, and $\|$ represents the concatenation of two vectors.

In the model, two different Bi-LSTMs are used on the document and the aspect phrase sequence, respectively, to prevent them from mixing their features through the hidden states. Thus, the information of the document and the aspect phrases can be sufficiently distinguished in the upper layers of the network, and it is more conducive to extracting semantic information correlated to affective entities from the document. So from the encoding layer, we can obtain the hidden states ($h_i^c, i \in \{1, 2, ..., n_1\}$) of the document $E^c = \{e_1^c, e_2^c, \ldots, e_{n_1}^c\}$ and the hidden states ($h_j^a, j \in \{1, 2, ..., n_2\}$) of the aspect phrases $E^a = \{e_1^a, e_2^a, \ldots, e_{n_2}^a\}$, respectively generated by two Bi-LSTMs according to the above process, and then concatenate their text representation vectors as $H = \{h_1^c, h_2^c, \ldots, h_{n_1}^c, h_1^a, h_2^a, \ldots, h_{n_2}^a\}$.

N-gram source capsule layer. In this layer, we extract N-gram source capsule features of the hidden vector H as the input for the capsule network by one-dimensional convolution operation using different sizes of convolution kernels, where each source capsule contains a 4×4 pose matrix and an activation value.

Set the set of one-dimensional convolution kernel sizes $k = \{k_i | k_i \in \mathbf{N}\}$, where $\mathbf{N}$ is a natural number set. In the model, we use $k = \{3, 5, 7\}$ to extract the text local features of 3-gram, 5-gram, and 7-gram, respectively. Suppose $H_{s:t} (1 \leq s \leq t \leq n)$ express the column vectors in H from the sth to tth column, the one-dimensional convolution operation on the jth vector is

$$c_j^{k_i} = ReLu(W^{k_i} \cdot H_{-\frac{k_i-1}{2}+j:\frac{k_i-1}{2}+j} + b^{k_i}) \tag{9}$$

where $W^{k_i} \in \mathbb{R}^{k_i \times n} (n = n_1 + n_2)$ is the weight coefficient of the convolution kernel with a size of k_i, $b^{k_i} \in \mathbb{R}$ is its corresponding bias, and Relu is a rectified linear unit. The k_i-gram output feature map of H is $c^{k_i} = [c_0^{k_i}, c_1^{k_i}, \ldots, c_{n-1}^{k_i}]$.

The implicit vector needs to be padded on its two sides before the one-dimensional convolution operation so that the length of the generated vector c^{k_i} is equal to that of the implicit vector H. The pose matrix sequence P^{pri} and the active value vector a^{pri} of source

capsules can be computed after multiple repeated one-dimensional convolution operations through Equations (10) and (11).

$$P^{pri} = Multiple(\{c^3, c^5, c^7\}, 16 * d) \tag{10}$$

$$a^{pri} = Multiple(\{c^3, c^5, c^7\}, d) \tag{11}$$

where $P^{pri} \in \mathbb{R}^{(3*d)\times(4\times4)\times n}$, $a^{pri} \in \mathbb{R}^{(3*d)\times n}$, $Multiple(a, b)$ denotes to repeat b times on the operation of determining a, and d is the base number of repeating executions with different sizes of the convolution kernel. Computing P^{pri} needs to be repeated $16 * d$ times because the size of a pose matrix is 4×4.

Each column vector of P^{pri} corresponds to one word and its different N-gram local features nearby, and each pose matrix expresses one of the local features of the word. So for each word, there are total $3 * d$ pose matrixes for representing its different local features, and $3 * d$ active values in each column express the active probabilities of the corresponding pose matrix, respectively.

Adaptive mask memory attention layer. Attention is a mechanism for flexibly selecting the reference of context information, which can facilitate global learning [53]. We propose an adaptive mask memory attention for capturing the global emotional information related to the local features over a long-distance text. In this layer, the mask mechanism is first used to clear invalid attention at the padding positions in the text. Then, the attention is adaptively scaled so that less attention is paid to the contexts that are not relevant to the aspect phrases. Last, the adaptive mask attention is used to weight the sum of the memory matrix, that is, the source capsule matrix sequence, for obtaining the capsule features containing global context semantics.

Let $M = P^{pri}$ express the memory matrix and $Q \in \mathbb{R}^{(3*d)\times(4\times4)\times L_q}$ denote the query capsule matrix, where L_q is the width of a query, and $3 * d$ is the height of capsules, which is the number of types of capsules. Since the different types of capsules mean different kinds of local features, we calculate the attention of different types of capsules separately. Memory attention is computed with Equations (12) and (13).

$$g_{jil} = tanh(W_j^{att} \cdot [m_{ji}, q_{jl}] + b_j^{att}) \tag{12}$$

$$att_{jil} = \frac{exp(g_{jil})}{\sum_1^n exp(g_{jil})} \tag{13}$$

where j denotes the jth capsule type; m_{ji} and q_{jl} are the capsules of the jth type at the ith sequence in matrix M and the lth sequence in matrix Q, respectively; $W_j^{att} \in \mathbb{R}^{2*16}$ is the attention weight of the jth type of capsules; $b_j^{att} \in \mathbb{R}$ is the bias of the jth type of capsules; g_{jil} is the similarity score between capsule m_{ji} and q_{jl}; att_{jil} is the result after normalization of g_{jil}.

- **Mask mechanism.** In word embedding, an input sequence into the deep model is generally required to have a uniform length, so the inputted word vectors will be padded. As shown in Figure 2, [pad] is the padding label in a word sequence. The attention is calculated over the global vector, and some invalid attention scores will be obtained in the [pad] positions of the second row. The proportion of invalid attention will increase with the increase of the padding length, which is obviously unreasonable for short texts. Therefore, we will mask these attention scores in padding positions according to the actual length of text, as shown in the third row of Figure 2.

Word	. . .	太	难用	了	。	[pad]	[pad]	. . .	[pad]	照相	功能	[pad]	[pad]	. . .	[pad]
No mask	. . .	0.6	0.8	0.3	0.2	0.2	0.1	. . .	0.2	0.8	0.7	0.3	0.1	. . .	0.1
Mask	. . .	0.6	0.8	0.3	0.2	0	0	. . .	0	0.8	0.7	0	0	. . .	0

Figure 2. Masking mechanism of attention. The first row shows a word sequence with [pad] labels. The attention scores of the according words are shown in the second row, and the results after masking [pad] scores are in the third row.

- **Adaptively scaling.** Since e expect to obtain the contextual features more relevant to the aspect phrases through attention, we do not pay too much attention to irrelevant words. To concentrate on these words that are important for identifying sentiment, we apply an adaptive scaling mechanism on the contextual attention according to the correlation with the aspect phrases, as shown in Figure 3. Specifically, the attention of the jth type of capsules at the lth sequence of matrix Q is scaled as follows:

$$att^c_{jl} = \frac{max(att^a_{jl})}{max(att^c_{jl})} r^{mask} a^c_{jl} \tag{14}$$

where att^c_{jl} denotes the contextual attention of the jth type of capsules at the lth sequence of matrix Q, i.e., $att^c_{jl} = \{att_{jil} | i \in [1, n_1]\}$; i is the ith word of the context; att^a_{jl} denotes the aspect attention of the jth type of capsules at the lth sequence in matrix Q, i.e., $att^a_{jl} = \{att_{jtl} | t \in [1, n_2]\}$; t is the tth word of the aspect phrases; n_1 and n_2 are the numbers of textual words and aspect words, respectively; r^{mask} is mask coefficient.

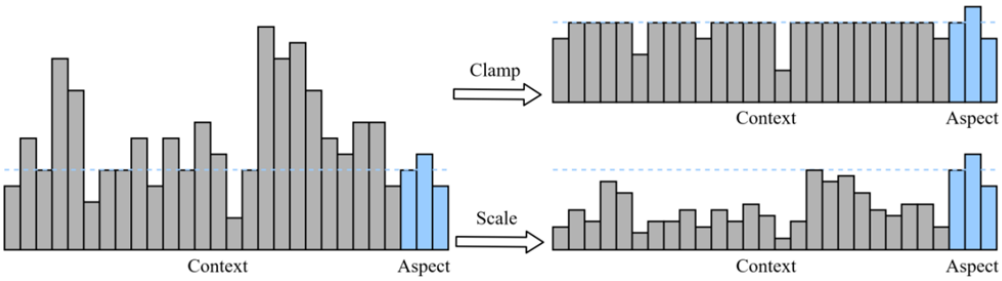

Figure 3. Adaptive scale mechanism of attention. For paying more attention to the context than the aspect phrases, an adaptive scale mechanism on the contextual attention is used according to their correlations with aspect phrases, rather than a clamp mechanism to tailor the contextual attention according to the aspect phrases.

After masked and adaptively scaled processing, the attention vector att_{jl} of the jth type of capsules at the lth sequence of matrix Q can be produced as $att_{jl} = [att^c_{jl}, att^a_{jl}]$. With the vector att_{jl} and the memory matrix M as inputs, we can obtain the aggregated capsule r_{jl} of the jth type of query capsules at the lth sequence by weighted sum operation as follows:

$$r_{jl} = \sum_{i=1}^{n_1+n_2} att_{jil} m_{ji} \tag{15}$$

Then, the jth-type global capsule r_j is produced by calculating the jth type of aggregated capsules at the different locations as $r_j = \{r_{j_1}, r_{j_2}, \ldots, r_{j_{L_q}}\}$. Finally, we can obtain the output of the adaptive mask memory attention layer by computing all types of global capsule matrixes as $p^{att} = \{r_1, r_2, \ldots, r_{(3*d)}\}$.

3.2.2. AgCapsNet

We use AgCapsNet to obtain the local capsule features relevant to the affective entities from an image by an aspect-guided attention, which mainly corresponds to the right of

Figure 1, including the feature extracting layer, the image source capsule layer, and the aspect-guided attention layer.

Feature extracting layer. In this layer, we extract the basic features of an image by the different receptive fields with a size of 5×5, and obtain a feature mapping with the depth of A (channel number) after 2-stride convolution and ReLU activation. That is, for the jth image I_j in set I^i, its feature matrix, $m_j \in \mathbb{R}^{d \times d \times A}$, is obtained through the feature extracting layer, where $d \times d$ is the size of the feature map in each channel.

Image source capsule layer. Similar to the N-gram source capsule layer, in this layer we will transform the basic features of an image into capsule features. That is, each position of the feature matrix is expressed as a capsule, including a 4×4 pose matrix and an active value. Here, a 1×1 convolutional kernel is used for linear transforming on the different channels in a feature matrix to obtain the pose matrixes of image source capsules. Then, the corresponding activation value of each pose matrix is obtained by the sigmoid activation function, as shown in Figure 1. After the feature matrix m_j is encapsulated, the pose matrix of image capsules and its corresponding activation value matrix are denoted as $P_j^{pri} \in \mathbb{R}^{(d \times d) \times B \times (4 \times 4)}$ and $a_j^{pri} \in \mathbb{R}^{(d \times d) \times B}$, respectively, and B is the depth of the source capsule layer.

Aspect-guided attention layer. After the feature representation of an image in the source capsule layer, we use a Text2vision attention to aggregate the aspect-guided information into the image capsule features, and then obtain image local information related to the affective entities for aspect-based sentiment classification. The Text2vision attention mechanism will be elaborated in the next section on multimodal fusion based on interactive learning.

3.2.3. Multimodal Fusion Based on Interactive Learning

The interactive learning of the memory network is mainly to capture the auxiliary information between the text and the image for improving the performance of multimodal sentiment analysis. To explore the complementarity information of text and images, we propose a multimodal multi-round fusion module based on interactive learning using the capsule features from MemCapsNet and AgCapsNet for aspect-based sentiment classification. A new visual enhanced feature fusion network is designed for interactive learning of the different modalities. The network iteratively queries the text and image features by using the next multi-hop storage system to explore the relationship between text and images. The upper part of Figure 1 illustrates the architecture of the module, including three parts of processes: an aspect-guided image learning based on a Text2Vision attention mechanism, an image-guided text learning based on a Vision2Text attention mechanism, and a capsule convolution.

Aspect-guided image learning. In order to learn the image features related to the affective entities, an aspect-guided Text2Vision attention mechanism is designed for allowing the aspect features to help the model find the key feature mapping of an image, that is, using the aspect capsules to conduct the gating selection on the image capsules of different types and at different locations.

Let $p_j^l \in \mathbb{R}^{B \times (4 \times 4)}$ be the pose matrix vector of the jth image I_j at the lth position and $l \in [1, d \times d]$. v^a is an aspect mapping vector from the hidden vector of aspect phrases in MemCapsNet. As shown in Figure 1, an aspect mapping is obtained by performing a temporal maximum pooling on the aspect phrase indices of the text source capsule sequence P^{pri}:

$$v^a = pooling(p_x^{pri} | x \in [n_1 + 1, n]) \tag{16}$$

where p_x^{pri} is the pose vector of $3 * d$ capsules of the xth aspect word. In order to gain the aspect-guided attention, we first compute the projects of the pose vector p_j^l and the aspect vector v^a onto the Text2Vision attention space as k_l and q^a, respectively:

$$k_l = tanh(w_k p_j^l + b_k) \tag{17}$$

$$q^a = tanh(w_q v^a + b_q) \tag{18}$$

where $w_k \in \mathbb{R}^{d_{att} \times B \times (4 \times 4)}$ and $w_q \in \mathbb{R}^{d_{att} \times B}$ are the coefficient matrixes of projecting, b_k, $b_q \in \mathbb{R}^{d_{att}}$ are their corresponding biases, and d_{att} is the size of the attention space. Then, the aspect-guided correlation score g_l and the attention score β_l are computed as follows:

$$g_l = V \cdot (k_l \odot q^a + k_l) \tag{19}$$

$$\beta_l = \frac{exp(g_l)}{\sum_l exp(g_l)} \tag{20}$$

where $V \in \mathbb{R}^{d_{att}}$ is a randomly initialized contextual vector that can be trained, and we can obtain the attention matrix of images $\beta_j = \{\beta_{jl} | 1 \leq l \leq d * d\} (1 \leq j \leq m)$. Lastly, the most relevant image with the given affective entity is found as below:

$$J = \arg\max_l (\|\beta_l\|_2) \tag{21}$$

where $\| \cdot \|_2$ is the 2-norm function of the matrix. We can complete the gating selection of the image capsule using attention matrix β_l to weight the capsule features of I_J, and obtain the capsule feature map p_J related to the affective entity.

Subsequently, we conduct the multilayer attention computation on p_J to get the final attention matrix β_J and the weighted capsule feature mapping $\beta_J \odot p_J$, which will be inputted into the add and norm layer for the further sentiment computation. At the same time, with β_J and p_J as the inputs for $\sum$ operation, we can obtain the global feature matrix:

$$v^{img} = \frac{\sum_k \beta_{Jk} \odot p_{Jk}}{\sum_k \beta_{Jk}} \tag{22}$$

Then, v^{img} can be inputted into the Vision2Text attention layer for image-guided text feature learning.

Image-guided text learning. After Text2Vision attention learning, we get the aspect-guided image features relevant to the affective entity. Then, to further utilize the mutual information between text and image for sentiment analysis, we continue to learn the image-guided text features based on the Image2text attention mechanism, which enhances the sentiment of text using the auxiliary information from aspect-guided image features.

Similar to Text2vision attention calculation, we first compute the projects of the text capsule matrix P^{att} from MemCapsNet and global image vector v^{img} onto the attention spaces k_l and q^a:

$$k_l = tanh(w_k p_j^{att} + b_k) \tag{23}$$

$$q^a = tanh(w_q v^{img} + b_q) \tag{24}$$

where $w_k \in \mathbb{R}^{d_{att} \times (3*d) \times (4 \times 4)}$ and $w_q \in \mathbb{R}^{d_{att} \times (3*d)}$ are the weight matrixes of projecting. Then, we compute the correlation score g_l and the attention score γ_l after normalization as follows:

$$g_l = K \cdot (k_l \odot q^{img} + k_l) \tag{25}$$

$$\gamma_l = \frac{exp(g_l)}{\sum_l exp(g_l)} \tag{26}$$

where, similar to V, $K \in \mathbb{R}^{d_{att}}$ is a randomly initialized contextual vector that can be trained. Finally, we use the attention matrix $\gamma = \{\gamma_l | 1 \leq l \leq L_q\}$ to weight the capsule features of P^{att} on the width dimension and complete the image-guided text feature aggregation.

Capsule convolution. The EM routing algorithm is used to perform a convolution operation on the capsule matrixes, and then the cluster centers are extracted to find the

sentiment information of text and images that is more related to the aspect phrases. Finally, the results of sentiment classification are output through a fully connected routing layer. It should be noted that the text capsule matrix and the image capsule matrix are convolved using one- and two-dimension convolutional kernels, respectively.

Firstly, a transformation matrix $\mathbf{M}_{ij}$ is applied to the pose matrix p_i of a low-level capsule $i \in \Omega_L$, where Ω_L is the Lth convolutional layer, and the vote matrix $\mathbf{V}_{ij} = p_i \mathbf{M}_{ij}$ is obtained which can be routed to the high-level capsule $j \in \Omega_{L+1}$. Then, EM routing is conducted as follows: (1) M stage. We compute the mean μ_j^h and the variance $(\sigma_j^h)^2$ of the mixture Gaussian distribution of capsule j on the dimension h using Equations (27) and (28).

$$\mu_j^h = \frac{\sum_i R_{ij} V_{ij}^h}{\sum_i R_{ij}} \tag{27}$$

$$(\sigma_j^h)^2 = \frac{\sum_i R_{ij} (V_{ij}^h - \mu_j^h)^2}{\sum_i R_{ij}} \tag{28}$$

where R_{ij} is a routing assigned coefficient, and V_{ij}^h is the value of V_{ij} on h dimension. Then the active value act_j of capsule j is computed according to the shortest description distance as follows:

$$act_j = sigmoid(\lambda(\beta_a - \sum_h cost_j^h)) \tag{29}$$

$$cost_j^h = (\beta_u + log(\sigma_j^h)) \sum_i R_{ij} \tag{30}$$

where β_u and β_a are trainable parameters, representing the description distance costs of mixed Gaussian model for each data point and for its mean and variance when its capsule is activated, respectively, and λ is the inversion coefficient, denoting the sensitive degree of the active value to the description distance cost. (2) E stage. We adjust the R_{ij} of the data point to determine the lower bound of the log-likelihood of the mixed Gaussian distribution under the current parameters using Equation (31).

$$R_{ij} = \frac{act_j p_j}{\sum_j act_j p_j} \tag{31}$$

where the pose matrix p_j is obtained by the joint probability of the mixed Gaussian distribution:

$$p_j = \frac{act_j p_j}{\sqrt{\prod_h 2\pi(\sigma_j^h)^2}} exp(-\sum_h \frac{(V_{ij}^h - \mu_j^h)^2}{2(\sigma_j^h)^2}) \tag{32}$$

Thus, the sequences of pose matrix and active value of all capsules, $p = \{p_j | j \in \Omega_{L+1}\}$ and $act = \{act_j | j \in \Omega_{L+1}\}$, can be obtained through iterating M stage and E stage.

In the last layer of capsule convolution, we conduct a fully connected routing on the feature capsules, which is to construct a line transformation from the feature capsules to three sentiment capsule classes corresponding to the label set $L = \{Positive, Neutral, Negative\}$, and the active values of three class capsules are acted as the logit values used for sentiment classification.

3.3. Training and Predicting

During training, MemCapsNet and AgCapsNet are separately used for encoding text and images. In the other words, the capsule representations for the text and images are obtained by two models, respectively. Multimodal fusion based on interactive learning is applied to select and enhance the aspect-based sentimental information of capsule features using the mutual information between text and images. Finally, the probabilities of aspect-based sentiment labels are outputted from the capsule convolution layer. Our training

objective is to maximize the interval between the active value act_t of the objective label and act_i of the other labels, and the loss function is defined as

$$Loss = \sum_{i \neq t} Loss_i \tag{33}$$

$$Loss_i = (\max(0, m - (act_t - act_i)))^2 \tag{34}$$

where m is an interval coefficient. m is set as 0.2 at the beginning, and linearly increased to 0.9 with the training. We start training with a smaller interval value to avoid the problem of dead capsules due to early excessive punishment. Since the two modalities of text and image correspond to two losses $Loss^{txt}$ and $Loss^{img}$, respectively, in order to make the text capsules with more aspect-oriented information play a leading role in the training, we use the weighted interval loss as the joint objective function of the network, i.e.,

$$Loss = Loss^{txt} + w^{img} Loss^{img} \tag{35}$$

where weighted coefficient $w^{img} \in [0, 0.5]$. $Loss^{txt}$ and $Loss^{img}$ are two modal losses obtained from Equations (33) and (34), respectively.

Algorithm 1 is the operational flow of VECapsNet.

Algorithm 1 Multimodal aspect-based sentiment analysis based on a VECapsNet.

Input: Multimodal triplet test dataset: $\{< T, I, A_t >\}$
Output: The sentiment label set L_t at aspect a_t
1. Extract the context word vector e^c, the aspect word vector e^a and the image feature mapping set M
2. Obtain the text hidden vectors h^c&h^a from Bi-LSTMs, and concatenate them as sequence H
3. Get the pose matrix p^{pri} and the active matrix a^{pri} of H from source capsule layer
4. Obtain the attention p^{att} from adaptive mask attention layer
5. Obtain the aspect mapping v^a using Equation (16).
6. Get the pose matrix p_m^{pri} and the active matrix a_m^{pri} of each image m in M, and obtain its attention matrix β_m using p_m^{pri} and v^a from Text2vision module according to Equations (16)–(20)
7. Compute the most relevant image index J from all β_m using Equations (21)
8. Compute the weighted image pose matrix p_J^{img} using β_J and p_J^{pri}
9. Repeat
10. Compute the Text2vision attention β_J using p_J^{img} and v^a
11. Obtain the global image vector v^{img} using Equation (22)
12. Compute the weighted matrix p_J^{img} using β_J and p_J^{img}
13. Compute the Vision2text attention p^{txt} using v^{img} and p^{att} according to Equations (23)–(26)
14. Obtain new p_J^{img} and a_J^{img} by image convolution layer
15. Obtain new p^{txt} and a^{txt} by text convolution layer
16. Update Loss using Equation (35)
17. Until the accuracy of the validation dataset no longer increases over ten epochs or reaches default N times
18. Obtain the class capsules p^{img_class} and p^{txt_class}, and the class active a^{img_class} and a^{txt_class} from last fully connected routing layer
19. Retrieve the predicted label set L_t from the aggregated distribution of a^{img_class} and a^{txt_class}.

4. Experiment

4.1. Experimental Data and Preprocessing

We conduct the multimodal aspect-based sentiment analysis task using text-image review data from social media, and adopt six review datasets for these experiments: Lap14,

Rest14, Rest15, Rest16, Multi-ZOL, and MTCom, where Lap14, Rest14, Rest15, and Rest16 are single-text datasets from the International Conference on Semantic Evaluation (SemEval), published from 2014 to 2016. They correspond to the product review datasets in English for one laptop review dataset (Lap14) and three restaurant review datasets (Rest14, Rest15, and Rest16), respectively. Table 2 shows the details of the above four datasets.

Table 2. Statistics of the datasets on SemEval.

Dataset	Positive		Neutral		Negative	
	Train	Test	Train	Test	Train	Test
Lap14	987	341	460	169	866	218
Rest14	2164	728	633	196	805	196
Rest15	955	34	272	340	28	195
Rest16	1297	63	466	474	29	127

Multi-ZOL [10] is a multimodal aspect-level sentiment dataset, including a total of 5228 text-image reviews for mobile phones from ZOL.com in China. Each review includes textual content (an average of 315 words long) and an image set (an average of 4.5 images) in six aspect categories. A total of 28,469 aspect–review pairs are obtained by pairwise combining each aspect category and each review. Each aspect-based sentiment has an emotional score from 1 to 10, which can be used as a label in the aspect-based sentiment classification.

In order to further analyze the aspect-based sentiment of multimodal data, we built a multimodal review dataset by crawling the reviews of Chinese restaurants from the Meituan website. The mate dataset consists of 791,852 text-image reviews. Each review contains textual content, a user's self-report tag, an image set including at least three images, and a star rating from 1 to 5. Since the original data is unevenly distributed over 191 tags and 5 ratings, we classified original reviews into six aspects according to users' self-report tags. The six aspects are "general comment", "location", "environment", "prices", "tastes", and "service". For each aspect, each review has a star rating from 1 to 5, which is regarded as a label for sentiment classification. After randomly discarding the data from more than 10,000 reviews with a single star rating, we finally obtain 42,543 pieces of multimodal reviews for the task of the aspect-based sentiment analysis. Table 3 shows the statistics of stars in MTCom.

Table 3. Statistics of the stars in MTCom.

Star Rating	Review Number	Percentage
5	10,000	23.50%
4	10,000	23.50%
3	10,000	23.50%
2	7428	17.45%
1	5115	12.02%

4.2. Experimental Setup

In the experiment, The parameter settings and evaluation methods of the model are as follows:

Network settings. First, in MemCapsNet, the GloVe model is used in the word embedding layer, and the unit numbers of the input vector layer and the hidden layer are both set to 300 in BiLSTMs. Second, in AgCapsNet, ResNet with RGB three channels is used as the feature mapping network in the feature extracting layer, where the size of the convolution kernel is 5×5 and the stride is 2. Last, for capsule convolution networks, the number of iterations of EM routing in capsule convolution of text and image is uniformly set as 3. The number of convolution layers in both networks is 2. The one-dimensional convolution kernel of text capsules in two layers are 5 and 3, respectively, and the strides are 3 and 2 respectively. Meanwhile, the two-dimensional convolution kernel of image capsules is 3×3 and the strides in two layers are 2 and 1, respectively.

Model setup and metrics. In the multimodal experiments, the maximum number of samples used at the same time in training is 8. If only MemCapsNet is used for text sentiment analysis, the maximum number of samples used simultaneously is 64. In our all neural network-based experiments, the Adam optimizer with a learning rate of 0.0005 is used to minimize the interval loss of the model, where the L2 regularization coefficient was 0.0000002. A dropout strategy with a ratio of 0.3 is used in the add and norm layers. The metrics used in our experiment are accuracy and F_1-score (F_1), and the definition of F_1-score is expressed below:

$$F_1 = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{36}$$

where $Prec$ and Rec are precision and recall, which are used to calculate F_1. The micro-F_1 method is employed to evaluate the results of aspect-based fine-grained sentiment analysis in our experiments.

4.3. Baselines

We compare our model with the following baseline models. To highlight the advantages of multimodal feature fusion, we separately evaluate MemCapsNet and AgCapsNet for single-modal aspect-based sentiment analysis. ATAE-LSTM, MemNet, and IACapsNet are used as single-text baseline models on the SemEval datasets for aspect-based sentiment analysis. Since there were no publicly available methods and datasets for aspect-based image sentiment analysis on review data, we compare the results of AgCapsNet with the reproduced results on VGG-16 [54] and Capsule Network (CapsNet) [55] only using the image data in Multi-ZOL and MTCom. Finally, the comparison experiments are conducted for our VECapsNet and the strong baseline method MIMN on the Multi-ZOL dataset and our MTCom dataset.

ATAE-LSTM [6]: An aspect-based text sentiment analysis model, which uses the attention mechanism between hidden states and aspect words to obtain aspect-related keyword representations for text sentiment analysis.

MemNet [26]: A memory model using multiple attention mechanisms on a memory matrix stacked by inputted word vectors.

IACapsNet [33]: A capsule network model using a cross-attention mechanism for textual aspect-based sentiment analysis.

MIMN [10]: A multimodal aspect-based sentiment analysis model proposed by Xu et al., which uses two interactive memory networks with four types of attention to supervise the textual and visual information with the given aspect. They presented the task of aspect-based multimodal sentiment analysis for the first time, and provided a multimodal aspect-level sentiment dataset, Multi-ZOL.

4.4. Experimental Results and Analysis

4.4.1. Aspect-Based Text Sentiment Analysis

We first conduct the comparative experiment with MEMCapNet and three baseline methods on four textual evaluation datasets, Lap14, Rest14, Rest15, and Rest16. In the experiment using MEMCapNet for single-text aspect-based sentiment classification, the Vision2Text attention module in Figure 1 is removed. The experimental results are shown in Table 4, where we can observe that the MEMCapNet proposed in our method performs better than the other models in the terms of accuracy and F1-score. Obviously, we can find the reason from the model structures of these models. Compared with ATAE-LSTM, MemCapsNet has a convolutional neural network and a memory module, which can better extract local features of the text. Although MemNet uses the memory network, it only models the global context with long-distance dependence, which cannot find the structural information of local phrases and clauses. IACapsNet uses a capsule network for extracting the local features of text, but it does not get a better effect due to the lack the auxiliary information on the global context.

Then we further conduct the comparative experiment only using the text features on two multimodal datasets, and rows 1-3 of Table 5 show the results of MemCapNet and the reproduced results of the baseline models. It can be observed that although the results on all models are degraded, the MemCapNet proposed in our method still outperforms other baseline models. Experiments on two kinds of datasets both verify the effectiveness of MemCapNet on single-text aspect-based sentiment analysis. Compared with the results in Table 4, the performance in Table 5 is lower due to the degraded quality of online crawled review datasets relative to the well-defined evaluation datasets, and the lack of visual support in the same text-image review for sentiment labels.

Table 4. The metrics of accuracy and F1-score on four text datasets.

Dataset	Lap14		Rest14		Rest15		Rest16	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
ATAE-LST	68.70	-	77.20	-	-	-	-	-
MemNet	70.64	65.17	79.61	69.64	77.31	58.28	85.44	69.55
IACapsNet	76.80	73.29	81.79	73.40	80.41	61.67	88.71	64.48
MemCapsNet (our)	**77.56**	**74.14**	**83.56**	**74.78**	**80.41**	**61.67**	**89.16**	**71.62**

Table 5. The metrics of accuracy and F1-score on the multimodal datasets.

Dataset	Multi-ZOL		MTCom	
	Accuracy	F1	Accuracy	F1
MemNet-T	59.51	58.73	56.08	53.03
IACapsNet-T	58.69	57.47	53.48	50.77
MemCapsNet (our)	**59.47**	**58.91**	**56.32**	**54.71**
VGG-16-I	44.05	43.67	33.46	31.86
CapsNet-I	45.10	44.29	31.41	30.13
AgCapsNet (our)	**46.59**	**44.36**	**41.37**	**40.16**
MIMN	**61.59**	**60.51**	54.35	52.94
VECapsNet (our)	61.23	59.63	**57.87**	**56.81**

The suffix "*-T" indicates the model only used the text features in the multimodal datasets; the suffix "*-I" indicates the use of only image features in the multimodal datasets.

4.4.2. Aspect-Based Image Sentiment Analysis

We also evaluate the performance of AgCapsNet on aspect-based image sentiment analysis by the comparative experiments with two baseline models of VGG-16 and CapsNet. Due to the lack of a single aspect-based image emotional dataset, only the annotated image data in the Multi-ZOL and MTCom datasets are used in the experiments. As shown in Figure 1, the input of aspect vector v^a from the source capsule layer in MemCapsNet and the Text2vision attention module are included in the AgCapsNet for aspect-based image sentiment classification.

The experimental results are shown in rows 4–6 of Table 5. Obviously, the AgCapsNet proposed by us is better than the baseline models. No aspect information is integrated into the VGG-16 and CapsNet, proving that adding the relevant information of the affective entity to the aspect-based sentiment analysis task with a single-image modality can improve the accuracy of emotional recognition, as well as the effectiveness of the capsule features of the image in this task is verified.

4.4.3. Aspect-Based Multimodal Sentiment Analysis

We conduct the experiment on the multimodal datasets of Multi-ZOL and MTCom, and compare our model with the baseline MIMN model. The results of the two models are listed in the last two rows in Table 5. It can be observed that the experimental results of VECapsNet proposed in our method on the Multi-ZOL dataset are close to and a bit lower than MIMN, but the result on the MTCom dataset is significantly better than MIMN. Since the same model behaves differently on two datasets, we will analyze the experimental results from two aspects of the composition of the data set and the limitations of the model.

Analysis of datasets. Multi-ZOL is a multimodal sentiment annotated dataset provided by Xu et al. [10] for evaluating the MIMN model, which contains six aspects: "price-performance ratio", "performance configuration", "battery life", "appearance and feeling", "photographing effect", and "screen". Except "price-performance ratio", the corresponding affective entities with the aspects can be intuitively found in the images in reviews. For example, "performance configuration" corresponds to the run points' scores with different colors or the line charts and other statistical charts; "battery life" corresponds to the battery consumption diagrams; "appearance and feeling" corresponds to the photos of mobile phones; "photographing effect" corresponds to the close-range or long-range photos; "screen" corresponds to the lighted mobile phone screens, etc. In the MTCom dataset, for the annotated aspects of "general comment", "location", "environment", "prices", "tastes", and "service", it is difficult to find their corresponding affective entities in the review images, except for "environment" and "tastes". Therefore, the reason for the performance differences of the model in the two datasets is likely to be the non-correspondence between the aspects and the affective entities of images in most of reviews.

At the same time, we also further explore the compositions of the text content and images in two datasets: Multi-ZOL and MTCom. Figure 4 sorts the average text lengths and image numbers of samples in two datasets, respectively, and the statistics of their 91st to 99th quantiles are listed in Table 6. Taking the statistics in the typical 95th quantile as an example, we can see that the average text length of MultiZOL is 803, while that of MTCom is only 146, and the average number of images in Multi-ZOL is 14, while that of MTCom is only 7. This is because, for expensive electronic products such as mobile phones, people tend to use longer text to describe all aspects of them in detail, as well as supplementing their reviews with corresponding images for almost every aspect. However, for the daily consumption of food and beverage, people usually do not spend a lot of time giving long reviews unless the dining experience is very good or very bad. So the text lengths and image numbers of reviews in the MTCom dataset are shorter and less than those in the Multi-ZOL dataset, and more representative of the majority of review data.

Analysis of limitations. The analysis of the datasets also verifies our design idea from another aspect, that is, in the task of multimodal sentiment analysis, text information is dominant, and an image is auxiliary rather than a complete ideographic unit. Furthermore, VECapsNet proposed in our method uses capsule networks to model the local sentimental features, so it has more advantages for processing short text than MIMN, which uses bidirectional LSTMs for capturing long-distance dependencies of context. Although the effectiveness of VECapsNet can be verified by the overall experimental results, for long text content and a small number of image comments with almost no explicit emotion, the performance of VECapsNet is shown to be degraded, which may be due to the fact that text-guided image feature learning may not be able to well mine the induced emotion in images. We also point out that for the direction of our further research, we still need to strengthen the emotion analysis for long text compared with the uncertainty of image emotion. We may consider decomposing the long text into short texts according to syntactic structures or integrating the structural information of long text into the capsule features in MMCapsNet for sentiment prediction, so as to capture more sufficient semantic features from the global context.

Table 6. The quantile statistics of context length and image number of samples on two multimodal datasets.

Quantile	Multi-ZOL		MTCom	
	Textual Length	Image Number	Textual Length	Image Number
91	537	9	105	6
93	648	9	122	7
95	803	14	146	7
97	1195	21	185	8
99	1922	34	285	9

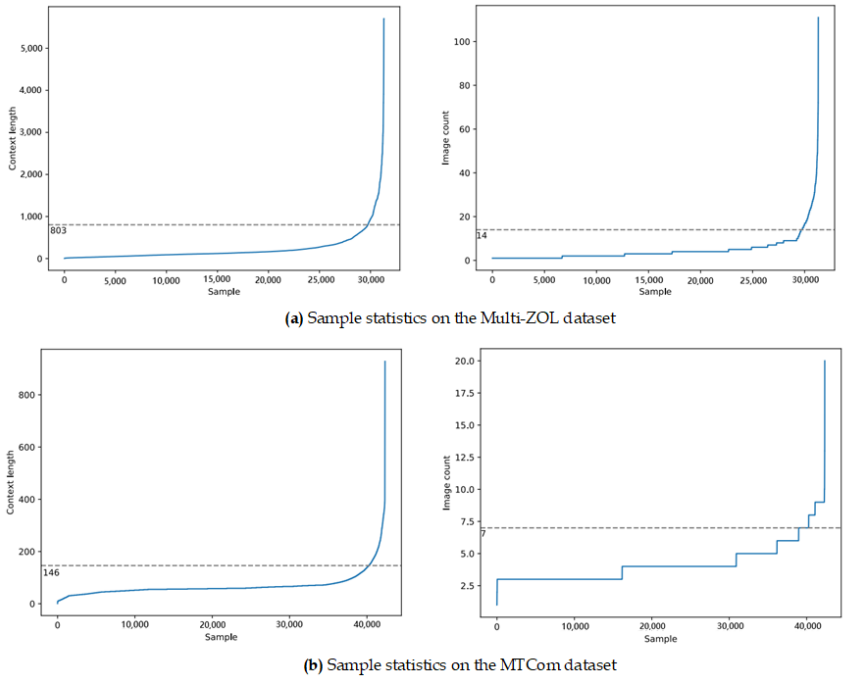

Figure 4. Average context length and image number on two multimodal datasets. (**a**,**b**) show the statistics of samples on the Multi-ZOL dataset and the MTCom dataset, respectively.

5. Conclusions

Aspect-based multimodal sentiment analysis is an important task in emotion analysis. In this paper, we proposed a visual enhancement capsule network (VECapsNet) based on interactive learning and built an aspect-based multimodal sentiment dataset (MTCom) by crawling reviews from the Meituan website. First, an adaptive mask memory capsule network is proposed for text feature learning. The model combines the memory networks and the capsule networks to integrate context semantic information into local word embedding and obtain capsule features related to aspect phrases by an adaptive mask attention mechanism. Second, we propose a novel aspect-guided capsule network for learning image sentiment features, which captures the local image information related to aspect phrases using a Text2vision attention mechanism to guide the sentimental learning on image feature capsules. Last, the multi-round fusion module is constructed by enhancing the text sentimental representations using the aspect-guided image features for aspect-based sentiment classification. The results of extensive experiments on the publicly available single-text evaluation datasets, the multimodal Multi-ZOL dataset, and on our MTCom dataset show that our method proposed in the paper performs satisfactorily on the different multimodal aspect-based sentiment analysis tasks. Compared with the existing models, VECapsNet can better capture the local semantic features correlated to the aspect phrases in text and images, which is more suitable for general text-based multimodal user-generated reviews. In the future, we will further study how to more deeply integrate long-text semantics and image interactions into the capsule features and apply our method to different multimodal sentiment analysis tasks.

Author Contributions: Conceptualization, Methodology, Original draft preparation, Y.Z.; Data processing, Experiments and analysis, Z.Z.; Supervision, Funding acquisition, Review and editing, D.W. and S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62172086, 61872074).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors are grateful to Xu, N. and Mao, W. for providing the datasets used in the experiments.

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

References

1. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
2. Yue, L.; Chen, W.; Li, X.; Zuo, W.; Yin, M. A survey of sentiment analysis in social media. *Knowl. Inf. Syst.* **2019**, *60*, 617–663. [CrossRef]
3. Abdi, A.; Shamsuddin, S.M.; Hasan, S.; Piran, J. Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Inf. Process. Manag.* **2019**, *56*, 1245–1259. [CrossRef]
4. Rao, T.; Li, X.; Zhang, H.; Xu, M. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing* **2019**, *333*, 429–439. [CrossRef]
5. Li, L.; Liu, Y.; Zhou, A. Hierarchical Attention Based Position-Aware Network for Aspect-Level Sentiment Analysis. In Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL), Brussels, Belgium, 31 October–1 November 2018; pp. 181–189.
6. Wang, Y.; Huang, M.; Zhao, L.; Zhu, X. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 606–615.
7. Li, P.; Chang, W.; Zhou, S.; Xiao, Y.; Wei, C.; Zhao, R. A conflict opinion recognition method based on graph neural network in Aspect-based Sentiment Analysis. In Proceedings of the 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 22–24 July 2022; pp. 1–6.
8. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
9. Kaur, R.; Kautish, S. Multimodal sentiment analysis: A survey and comparison. *Int. J. Serv. Sci. Manag. Eng. Technol. (IJSSMET)* **2019**, *10*, 38–58. [CrossRef]
10. Xu, N.; Mao, W.; Chen, G. Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 371–378.
11. Zhou, J.; Zhao, J.; Huang, J.X.; Hu, Q.V.; He, L. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing* **2021**, *455*, 47–58. [CrossRef]
12. Truong, Q.; Lauw, H. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 305–312.
13. Dong, L.; Wei, F.; Tan, C.; Tang, D.Y.; Zhou, M.; Xu, K. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Baltimore, MD, USA, 22–27 June 2014; Volume 2, pp. 49–54.
14. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv: 1408.5882.
15. Chen, X.; Rao, Y.; Xie, H.; Wang, F.L.; Zhao, Y.; Yin, J. Sentiment classification using negative and intensive sentiment supplement information. *Data Sci. Eng.* **2019**, *4*, 109–118. [CrossRef]
16. Chen, G.; Tian, Y.; Song, Y. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In Proceedings of the 28th International Conference on Computational Linguistics (COLING), Online, 8–13 December 2020; pp. 272–279.
17. Tang, D.Y.; Qin, B.; Feng, X.C.; Liu, T. Effective LSTMs for Target-Dependent Sentiment Classification. In Proceedings of the 26th International Conference on Computational Linguistics (COLING), Osaka, Japan, 11–16 December 2016; pp. 3298–3307.
18. Feng, S.; Wang, Y.; Liu, L.R.; Wang, D.; Yu, G. Attention based hierarchical LSTM network for context-aware microblog sentiment classification. *World Wide Web* **2019**, *22*, 59–81. [CrossRef]
19. Huang, M.; Cao, Y.; Dong, C. Modeling rich contexts for sentiment classification with LSTM. *arXiv* **2016**, arXiv:1605.01478.
20. Zhao, Z.; Lu, H.; Cai, D. He, X.; Zhuang, Y. Microblog Sentiment Classification via Recurrent Random Walk Network Learning. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 19–25 August 2017; pp. 3532–3538.
21. Xu, C.; Cetintas, S.; Lee, K.; Li, L. Visual sentiment prediction with deep convolutional neural networks. *arXiv* **2016**, arXiv:1411.5731.
22. Song, K.; Yao, T.; Ling, Q.; Mei, T. Boosting image sentiment analysis with visual attention. *Neurocomputing* **2018**, *312*, 218–228. [CrossRef]
23. Wu, L.; Zhang, H.; Shi, G.; Deng, S. Weakly Supervised Interaction Discovery Network for Image Sentiment Analysis. In *Asian Conference on Pattern Recognition*; Springer: Cham, Switzerland, 2022; Volume 13188, pp. 501–512.

24. Liang, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. Cross-Domain Semi-Supervised Deep Metric Learning for Image Sentiment Analysis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 4150–4154.
25. Liang, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. Deep Metric Network Via Heterogeneous Semantics for Image Sentiment Analysis. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1039–1043.
26. Tang, D.; Qin, B.; Liu, T. Aspect level sentiment classification with deep memory network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 214–224.
27. Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; Zhou, G. Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), Virtual, Dominican Republic, 7–11 November 2021; pp. 4395–4405.
28. Wang, B.; Lu, W. Learning Latent Opinions for Aspect-level Sentiment Classification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5537–5544.
29. Xu, L.; Bing, L.; Lu, W.; Huang, F. Aspect Sentiment Classification with Aspect-Specific Opinion Spans. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Virtual, Online, 16–20 November 2020; pp. 3561–3567.
30. Li, X.; Bing, L.; Lam, W.; Shi, B. Transformation Networks for Target-Oriented Sentiment Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, VIC, Australia, 15–20 July 2018; pp. 946–956.
31. Johnson, R.; Zhang, T. Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 919–927.
32. Chen, Z.; Qian, T. Transfer Capsule Network for Aspect Level Sentiment Classification. In Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 547–556.
33. Du, C.; Sun, H.; Wang, J. Qi, Q.; Liao, J.; Xu, T.; Liu, M. Capsule Network with Interactive Attention for Aspect-Level Sentiment Classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 3–7 November 2019; pp. 5488–5497.
34. You, Q.; Luo, J.; Jin, H.; Yang, J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 13–22.
35. You, Q.; Cao, L.; Jin, H.; Luo, J. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1008–1017.
36. Xu, N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 152–154.
37. 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.
38. 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.
39. Chen, Q.; Huang, G.; Wang, Y. The Weighted Cross-Modal Attention Mechanism with Sentiment Prediction Auxiliary Task for Multimodal Sentiment Analysis. *IEEE/ACM Trans. Audio Speech, Lang. Process.* **2022**, *30*, 2689–2695. [CrossRef]
40. Peng, C.; Zhang, C.; Xue, X.; Gao, J.; Liang, H.; Niu, Z. Cross-Modal Complementary Network with Hierarchical Fusion for Multimodal Sentiment Classification. *Tsinghua Sci. Technol.* **2022**, *27*, 664–679. [CrossRef]
41. Ji, R.; Chen, F.; Cao, L.; Gao, Y. Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning. *IEEE Trans. Multimed.* **2018**, *21*, 1062–1075. [CrossRef]
42. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
43. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [CrossRef]
44. Huang, F.; Zhang, X.; Li, Z. Learning joint multimodal representation with adversarial attention networks. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 1874–1882.
45. Huang, F.; Zhang, X.; Xu, J.; Zhao, Z.; Li, Z. Multimodal learning of social image representation by exploiting social relations. *IEEE Trans. Cybern.* **2021**, *51*, 1506–1518. [CrossRef]
46. Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact bilinear pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 317–326.
47. Wang, Z.; Xu, G.; Zhou, X.; Kim, J.Y.; Zhu, H.; Deng, L. Deep Tensor Evidence Fusion Network for Sentiment Classification. *IEEE Trans. Comput. Soc. Syst.* **2022**, 1–9. [CrossRef]
48. Xue, H.; Yan, X.; Jiang, S. Lai, H. Multi-Tensor Fusion Network with Hybrid Attention for Multimodal Sentiment Analysis. In Proceedings of the 2020 International Conference on Machine Learning and Cybernetics (ICMLC), Adelaide, Australia, 2 December 2020; pp. 169–174.

49. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowl.-Based Syst.* **2019**, *167*, 26–37. [CrossRef]
50. Huang, F.R.; Wei, K.M.; Weng, J.; Li, Z.J. Attention based modality-gated networks for image-text sentiment analysis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2020**, *16*, 79. [CrossRef]
51. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
52. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality, In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
53. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR)—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
54. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
55. Hinton, G.E.; Sabour, S.; Frosst, S. Matrix capsules with EM routing. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.