Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis

Nan Xu, Wenji Mao, Guandan Chen
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences, Beijing, China
{xunan2015, wenji.mao, chenguandan2014}@ia.ac.cn

Abstract

As a fundamental task of sentiment analysis, aspect-level sentiment analysis aims to identify the sentiment polarity of a specific aspect in the context. Previous work on aspect-level sentiment analysis is text-based. With the prevalence of multimodal user-generated content (e.g. text and image) on the Internet, multimodal sentiment analysis has attracted increasing research attention in recent years. In the context of aspect-level sentiment analysis, multimodal data are often more important than text-only data, and have various correlations including impacts that aspect brings to text and image as well as the interactions associated with text and image. However, there has not been any related work carried out so far at the intersection of aspect-level and multimodal sentiment analysis. To fill this gap, we are among the first to put forward the new task, aspect based multimodal sentiment analysis, and propose a novel Multi-Interactive Memory Network (MIMN) model for this task. Our model includes two interactive memory networks to supervise the textual and visual information with the given aspect, and learns not only the interactive influences between cross-modality data but also the self influences in single-modality data. We provide a new publicly available multimodal aspect-level sentiment dataset to evaluate our model, and the experimental results demonstrate the effectiveness of our proposed model for this new task.

Introduction

Aspect-level sentiment analysis is a fundamental task in the field of sentiment analysis, which has a number of practical applications in domains such as business, public management and social security. Existing work on aspect-level sentiment analysis is based on text modality (Wang et al. 2016; Tang, Qin, and Liu 2016; Ma et al. 2017; Chen et al. 2017). With the prevalence of multimodal user-generated content (e.g. text and image) on the Internet, multimodal sentiment analysis has attracted increasing research attention in recent years (Borth et al. 2013; Yu et al. 2016; Xu, Mao, and Chen 2018). In the context of aspect-level sentiment analysis, multimodal data are often more important than text-only data. For example, customers will browse product reviews before they buy a product, and those multimodal reviews are more likely to attract customers' attention. Based on a recent study on the reviews of over 1000 kinds of cellphone in ZOL.com, approximately 40% of the reviews contain both text and image, and the numbers of Reply and Like in these multimodal reviews are 3 times and 4 times greater respectively, than those in text-only reviews. This highlights the importance of analyzing aspect-level sentiment on multimodal data.

For the task of aspect-level sentiment analysis, image information is often as indicative as text information. On the one hand, in multimodal data, both text and image are highly associated with aspect sentiment. For example, when reviewing the aspect 'photographing effect' of a cellphone, customers may write down positive words and add high-quality photos into their reviews to show their satisfaction, or negative words and error image samples (e.g. red/ purple noise in low light photos photographed by HTC M7) to express their dissatisfaction. Furthermore, different aspects might be related to different parts on each modality data. In other words, customers may write down different words or attach different images for different aspects. On the other hand, text and image information can mutually reinforce and complement each other to enhance the analysis of specific aspect sentiment. In summary, various correlations exist in multimodal data for aspect-level sentiment analysis. However, there has not been any related work carried out so far at the intersection of aspect-level and multimodal sentiment analysis.

In this paper, we fill the research gap and propose the new task of aspect based multimodal sentiment analysis by firstly introducing the image modality data to traditional text based aspect-level sentiment analysis. Specifically, we focus on aspect based multimodal sentiment classification, with the aim of determining the sentiment polarity that user opinions conveyed in multimodal data on a specific aspect. To capture the impacts that aspect brings to text and image, as well as the multiple interactions associated with text and image, we propose a novel Multi-Interactive Memory Network (MIMN) for this task. Our model includes two interactive memory networks to supervise the textual and visual information with the given aspect, and then learns not only the interactive influences between cross-modality data but also the self influences in single-modality data. In addition, to advance and further test aspect based multimodal sentiment analysis research, we build a new multimodal dataset

crawled from ZOL.com[1], the leading IT information and business web portal in China. We conduct experiment on this dataset, and the results show that our proposed model outperforms the baseline methods, including the representative textual aspect-level sentiment analysis methods and the variant of the state-of-the-art multimodal sentiment analysis method.

The contributions of our model are as follows.

- We are among the first to propose the new task, aspect based multimodal sentiment analysis, which fills the gap between aspect-level sentiment analysis and multimodal sentiment analysis.
- We propose a novel multi-interactive memory network to capture the multiple correlations in multimodal data for aspect-level sentiment analysis, including impacts that aspect brings to text and image, and interactions in and between text and image.
- We provide a new publicly available multimodal aspect-level sentiment dataset to evaluate our model. The experimental results on our constructed dataset show the effectiveness of our proposed model.

Related Work

The new task of aspect based multimodal sentiment analysis stemmed from two lines of research, namely aspect-level sentiment analysis and multimodal sentiment analysis. We briefly introduce these two research areas.

Aspect-level Sentiment Analysis

Aspect-level sentiment analysis aims to identify the sentiment polarity of a textual sentence on a given aspect. Its research methods can be divided into two groups: traditional feature selection based methods and neural network based methods.

Traditional feature selection based methods mainly focus on designing a series of feature templates or introducing external resources like parser and sentiment lexicons to train a sentiment classifier (Jiang et al. 2011; Mohammad, Kiritchenko, and Zhu 2013; Kiritchenko et al. 2014; Wagner et al. 2014). These methods are all labor-intensive and require painstaking feature engineering.

Recently, neural networks have been demonstrated powerful performance in learning feature representation and achieved significant improvement than previous feature selection based methods in many text-based analysis tasks such as classification (Kim 2014), machine translation (Sutskever, Vinyals, and Le 2014), and QA (Dong et al. 2015). For aspect-level sentiment analysis task, Dong et al. (2014) firstly introduce the recursive neural network to this field, which adaptively propagates the sentiments of words to target depending on the context and syntactic relationships between them. However, the performance of their method depends on the syntax parsing which is likely ineffective or error in practice (Vo and Zhang 2015). To solve this problem, Vo and Zhang (2015) extract a rich set of automatic features using distributed word representations and neural pooling functions. Zhang, Zhang, and Vo (2016) further address the limitation of pooling functions and use two gated neural networks to capture tweet-level syntactic and semantic information and model the interactions between the left and right context of a given target. Tang et al. (2016) introduce the recurrent neural network and propose a target-dependent LSTM to model the context information, separating sentence into left and right context.

To further model the correlation between the context and aspect, attention mechanism is introduced into aspect-level sentiment analysis task. Tang, Qin, and Liu (2016) develop a memory network to focus on those informative context words which have a relationship with the given aspect by multiple attention mechanism. Wang et al. (2016) combine each context word with aspect embedding at input layer, and also propose the aspect guided attention mechanism to strengthen the reasonability of hidden representation. In fact, the aspect also havs sequence, to better model the sequence information of both context and aspect, Ma et al. (2017) use two LSTM models to respectively learn the representations of context and aspect and propose the interactive attention mechanism to interactively learn attentions between them. To better aggregate feature representations, Chen et al. (2017) use a memory model to learn multiple attentions and non-linearly combine these attentions by a recurrent neural network to strengthen the expressive power. To further deal with multi-aspect sentences and the syntactically complex sentence structures, Liu et al. (2018) propose the sentence-level content attention mechanism to capture the important information about given aspects from a global perspective, and context attention mechanism to simultaneously take the order of context words and their correlations into account.

Multimodal Sentiment Analysis

With the prevalence of multiform user-generated-content (e.g. text, image, speech or video), sentiment analysis has gone beyond traditional text-based analysis. Multimodal sentiment analysis is an emerging research area that integrates textual and non-textual information into user sentiment analysis.

The text-image pair is the most common form of multimodal data. Early work adopts feature-based methods. For example, Borth et al. (2013) generate visual features by extracting 1200 adjective-noun pairs from image, and textual features by calculating the sentiment scores of text based on English grammar and spelling style. With the development of deep learning technologies, some neural network based models have been proposed for multimodal sentiment analysis, achieving significant progress. Yu et al. (2016) pre-train text CNN and image CNN to extract feature representations from text and image respectively and combine these multimodal features to train a logistics regression model. To fully capture the visual semantic information, Xu and Mao (2017) extract scene and object features from image and absorb text words with these visual semantic features to model the influence that image brings to text. In fact, text and image can mutually reinforce and complement each other in sentiment analysis. Thus, Xu, Mao, and Chen (2018) propose

[1]http://www.zol.com.cn/

a co-memory attentional mechanism to interactively model the interaction between text and image. Their model considers the influence of one modality to another (i.e. text to image and image to text) and achieves better performance than other related methods.

There is also other multimodal sentiment analysis work dealing with textual, visual, and audio modalities, in which visual (such as facial expression) and audio (such as pitch) and utterance are utilized for emotion recognition and sentiment analysis (Poria, Cambria, and Gelbukh 2015; Poria et al. 2016; Wang et al. 2016; Zadeh et al. 2017).

To bridge the gap between aspect-level sentiment analysis and multimodal sentiment analysis, we propose a novel multi-interactive memory network for aspect based multimodal sentiment analysis in this paper. Our model fully captures the impacts that aspect brings to text and image, as well as the multiple interactions in and between text and image. With non-linear combination of multi-interactive memory attentions by a recurrent neural network, our model can learn a global memory abstraction for aspect based multimodal sentiment analysis.

Proposed Model

The overall architecture of our MIMN model is shown in Figure 1. Given a sample, suppose the multimodal inputs include a textual content $T = \{W_1, W_2, ..., W_L\}$ and an image set $I = \{I_1, I_2, ..., I_K\}$, the goal of our model is to predict the sentiment label with a given aspect phrase $A = \{A_1, A_2, ..., A_N\}$, where L is the length of textual context, K is the number of images, and N is the length of aspect phrase.

Feature Extraction

Aspect feature embedding. An aspect is a multi-word expression in most case. MemNet (Tang et al., 2016b) maps aspect into a single vector using average pooling, which ignores context information and cannot synthesize phrase-like feature in original aspect words. To overcome this, we use a Bidirectional LSTM model, similar to the work by (Chen et al. 2017; Ma et al. 2017). Given an aspect phrase $A = \{A_1, A_2, ..., A_N\}$, the D_{text}-dimensional embedding vector a_j of each word A_j is initialized by the word representation method (Li et al. 2018). At each time step, the LSTM unit learns the hidden representation $v_j \in \mathbb{R}^{2*D_h}$ of each aspect word embedding a_j.

$$a_j = Embed(A_j), j \in [1, N] \quad (1)$$

$$\overrightarrow{v_j} = \overrightarrow{LSTM}(a_j), j \in [1, N] \quad (2)$$

$$\overleftarrow{v_j} = \overleftarrow{LSTM}(a_j), j \in [N, 1] \quad (3)$$

$$v_j = [\overrightarrow{v_j}, \overleftarrow{v_j}], \quad j \in [1, N] \quad (4)$$

Then, we take the average of all hidden representations v_j as the final aspect feature vector $v^A \in \mathbb{R}^{2*D_h}$.

$$v^A = \frac{1}{N} \sum_j v_j \quad (5)$$

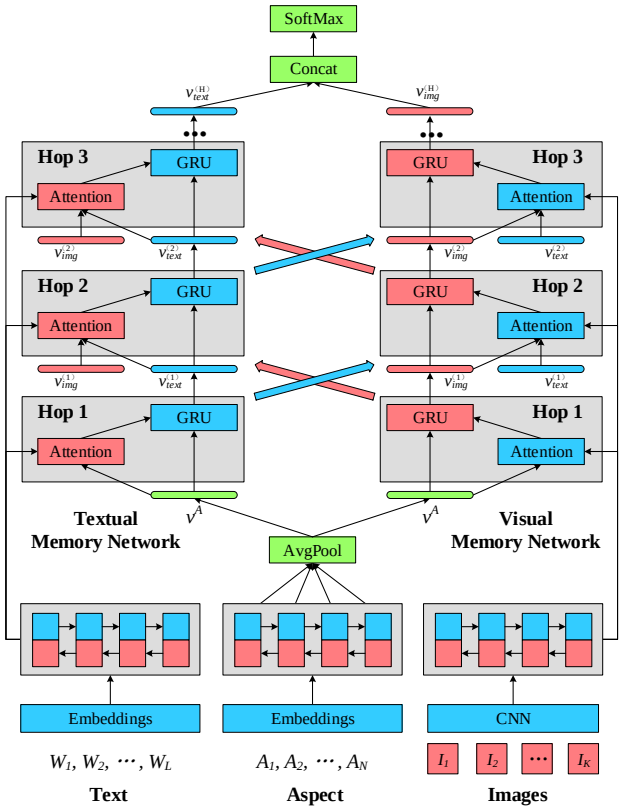

Figure 1: Overview of multi-interactive memory network for aspect based multimodal sentiment analysis.

Textual memory building. Unlike previous work that directly feeds sequence of word embeddings to the memory network (Tang, Qin, and Liu 2016), we adopt the Bi-LSTM to grasp phrase-like features and learn context information. Given a textual content $T = \{W_1, W_2, ..., W_L\}$, each word W_i is embedded to a word vector $w_i \in \mathbb{R}^{D_{text}}$ and initialized by the word representation method (Li et al., 2018). The LSTM unit takes an input word embedding w_i and output a hidden state $m_i^T \in \mathbb{R}^{2*D_h}$.

$$w_i = Embed(W_i), i \in [1, L] \quad (6)$$

$$\overrightarrow{m_i^T} = \overrightarrow{LSTM}(w_i), i \in [1, L] \quad (7)$$

$$\overleftarrow{m_i^T} = \overleftarrow{LSTM}(w_i), i \in [L, 1] \quad (8)$$

$$m_i^T = [\overrightarrow{m_i^T}, \overleftarrow{m_i^T}], \quad i \in [1, L] \quad (9)$$

We stack these hidden states and represent them as the external textual memory matrix M^T.

$$M^T = \{m_1^T, m_2^T, ..., m_L^T\} \quad (10)$$

Visual memory building. Images in multimodal data are usually arranged in sequence (e,g. Fig1, Fig2,...). For example, a multimodal review about cellphone has three paragraphs of text content, with the first paragraph mainly about aspect 'appearance and feeling' and the second and third

paragraphs about 'photographing effect'. The customer may add images about appearance of cellphone after the first paragraph and/or images about photographing at the end of the review.

To model this pervasive image sequence information and synthesize subset images for different aspects, we also adopt the Bi-LSTM model. Given an image set $I = \{I_1, I_2, ..., I_K\}$, we firstly adopt a pre-trained convolutional neural network and remove the top fully connected layer to extract D_{img}-dimensional visual feature vector x_k from each image I_k. Then, the LSTM unit takes visual feature vector x_k into hidden space.

$$x_k = CNN(I_k), k \in [1, K] \tag{11}$$

$$\overrightarrow{m_k^I} = \overrightarrow{LSTM}(x_k), k \in [1, K] \tag{12}$$

$$\overleftarrow{m_k^I} = \overleftarrow{LSTM}(x_k), k \in [K, 1] \tag{13}$$

$$m_k^I = [\overrightarrow{m_k^I}, \overleftarrow{m_k^I}], \quad k \in [1, K] \tag{14}$$

Each hidden state $m_k^I \in \mathbb{R}^{2*D_h}$ is stacked to build the visual memory M^I, which records all the visual information.

$$M^I = \{m_1^I, m_2^I, ..., m_K^I\} \tag{15}$$

Multimodal Memory Network

We use two memory networks for textual and visual feature extraction to focus on informative parts for the given aspect and suppress less important parts. At the first hop of each memory network, we propose the aspect-guided attention mechanism which supervises the generation of textual and visual attention vectors with aspect information.

Textual Memory Network. The textual memory network extracts important words for sentiment and aggregates textual memory with the representation of the given aspect to account for the influence that the aspect brings to texts. It takes the external textual memory matrix M^T and aspect feature vector v^A as the inputs and combines each piece of textual memory m_i^T with aspect feature vector v^A through a multi layer perceptron network to generate the textual hidden representation $h_i^{(1)}$.

$$h_i^{(1)} = tanh(w_{text}^{(1)}[m_i^T, v^A] + b_{text}^{(1)}) \tag{16}$$

It then calculates and normalizes the attention weight as follows:

$$\alpha_i^{(1)} = \frac{exp(h_i^{(1)})}{\sum_i exp(h_i^{(1)})} \tag{17}$$

Finally, the attention layer outputs the textual feature vector $v_{text}^{(1)}$ by weighted average of those textual memory pieces using attention weight $\alpha_i^{(1)}$.

$$v_{text}^{(1)} = \sum_i \alpha_i^{(1)} m_i^T \tag{18}$$

Visual Memory Network. We also propose the visual memory network to extract informative images. It combines image features with the representation of the given aspect to fully capture the influence that the aspect brings to images. We feed the visual memory M^I and the aspect feature vector v^A as the input of our visual memory network to extarct the visual feature vector $v_{img}^{(1)}$.

$$p_k^{(1)} = tanh(w_{img}^{(1)}[m_k^I, v^A] + b_{img}^{(1)}) \tag{19}$$

$$\beta_k^{(1)} = \frac{exp(p_k^{(1)})}{\sum_k exp(p_k^{(1)})} \tag{20}$$

$$v_{img}^{(1)} = \sum_k \beta_k^{(1)} m_k^I \tag{21}$$

Multi-Interactive Attention Mechanism

We have considered the influences of aspect on text and image respectively via our aspect guided attention mechanism. The corresponding operation is unidirectional: aspect to text or aspect to image. In fact, for multimodal data, textual and visual information mutually reinforce and complement each other in sentiment analysis. To fully capture the bidirectional interactions between image and text, we propose a multi-interactive attention mechanism. The previous interaction attention mechanism used in (Xu, Mao, and Chen 2018) only consider the influence one modality information brings to the other modality (e.g. image to text or text to image). Our multi-interactive attention mechanism consists of both cross-modality attention and single-modality attention. With this design, our model could learn not only the interactive influences caused by cross-modality data, but also the self influences caused by single-modality data (i.e. Text to text and image to image).

Textual Attention. Using the textual and visual memory network, we get the original textual feature vector $v_{text}^{(1)}$ and visual feature vector $v_{img}^{(1)}$. Next, in the t-th hop of the textual memory network, to learn the self influence caused by textual data, we feed the textual feature vector $v_{text}^{(t-1)}$ to query the textual memory M^T to generate the textual modality attentional feature $v_{text2text}^{Att(t)}$. Then, to learn the interactive influence that image bring to text, we use the visual feature vector $v_{img}^{(t-1)}$ to query the textual memory M^T again to get the cross-modality attentional feature $v_{img2text}^{Att(t)}$. Last, we finally average them as the textual attentional feature vector $v_{text}^{Att(t)}$.

$$v_{text2text}^{Att(t)} = Att([m_i^T, v_{text}^{(t-1)}]) \tag{22}$$

$$v_{img2text}^{Att(t)} = Att([m_i^T, v_{img}^{(t-1)}]) \tag{23}$$

$$v_{text}^{Att(t)} = \frac{v_{text2text}^{Att(t)} + v_{img2text}^{Att(t)}}{2} \tag{24}$$

where $t \in [2, H]$, and H is the number of memory hops. Att is the attention layer , which refers to the operation mentioned above (Eqs. 16-18, or 19-21).

Visual Attention. Synchronously, we also combine both the visual feature vector $v_{img}^{(t-1)}$ and textual feature vector $v_{text}^{(t-1)}$ with the visual memory M^I in the t-th hop of the visual memory network. The intermediate results are averaged as the visual attentional feature vector $v_{img}^{Att(t)}$ for the next operation.

$$v_{img2img}^{Att(t)} = Att([m_k^I, v_{img}^{(t-1)}]) \quad (25)$$

$$v_{text2img}^{Att(t)} = Att([m_k^I, v_{text}^{(t-1)}]) \quad (26)$$

$$v_{img}^{Att(t)} = \frac{v_{img2img}^{Att(t)} + v_{text2img}^{Att(t)}}{2} \quad (27)$$

Recurrent Memory Combination. Previous work has demonstrated the capability of deep neural network with multiple layers to learn the deep representations of data with multi-level abstractions (LeCun, Bengio, and Hinton 2015). Thus, we also stack our interactive attention mechanism with several memory hops to learn the deep abstraction of multimodal data. Unlike the linear layer used in previous work (Tang, Qin, and Liu 2016; Xu and Mao 2017) which simply transfers textual and visual hidden representations to the next memory hop, we adopt GRUs to combine all interactive attention memory results $v_{img}^{(*)}$ and $v_{text}^{(*)}$ non-linearly. The GRU has fewer parameters and is simpler than other recurrent models, which is also used in (Chen et al. 2017). This operation enhances the memory capability of our model for global memory abstraction. Formally, at the t-th memory hop, we first get the textual and visual attentional feature vectors using our multi-interactive attention mechanism. Then the GRU unit updates the new textual and visual feature vectors for the next operation.

$$v_{text}^{(t)} = GRU(v_{text}^{Att(t)}, v_{text}^{(t-1)}), t \in [2, H] \quad (28)$$

$$v_{img}^{(t)} = GRU(v_{img}^{Att(t)}, v_{img}^{(t-1)}), t \in [2, H] \quad (29)$$

Sentiment Classification

After H interactive memory hops, we extract the last outputs of GRUs as final textual and visual feature vectors and concatenate them as the input of a softmax layer to predict the aspect sentiment score.

$$Pred = Softmax(w_{multi}[v_{text}^{(H)}, v_{img}^{(H)}] + b_{multi}) \quad (30)$$

We train our model by minimizing the cross-entropy loss with the Adam (Kingma and Ba 2014) optimization algorithm. To avoid overfitting, the dropout (Hinton et al. 2012) is employed. During the training process, we also adopted the early stop strategy, i.e. stop training if the loss at the development set had gone down for several successive epochs.

Experiments

Dataset

There was no publicly available dataset for aspect based multimodal sentiment analysis. Hence, we provide a new publicly available multimodal aspect-level sentiment dataset. The ZOL.com is the leading IT information and business web portal in China. It consists of 40 large channels, including news, shopping malls, hardware, downloads, games, mobile phones etc. We crawl the top hot mobile phones' reviews from page 1 to 50 in mobile phones channel. For each mobile phone, only reviews in top 20 pages were crawled. The meta-dataset has 12587 reviews (7359 single-modal reviews, 5288 multimodal reviews), covering 114 brands and 1318 kinds of mobile phone. It can also be applied to textual aspect-level sentiment analysis or multimodal sentiment analysis task. The dataset is available at https://github.com/xunan0812/MIMN.

We evaluate our model on these 5288 multimodal reviews, namely the Multi-ZOL Dataset. In this dataset, each multimodal review contains a textual content, an image set, and at least one but no more than six aspects. The six aspects are price-performance ratio, performance configuration, battery life, appearance and feeling, photographing effect, and screen. We pair each aspect with multimodal review, getting 28469 aspect-review pairs of samples. For each aspect, the review has an integer sentiment score from 1 to 10, which is regarded as the sentiment label in our experiment. Table 1 shows the detailed statistics for this dataset.

Table 1: Statistics of Multi-ZOL Dataset

Attribute	Statistic
#Review	5228
#Label	10
#Aspect-Review Pair	28,469
Avg. of #Aspect / Review	5.45
Avg. text length / Review	315.11
Max text length / Review	8511
Min text length / Review	5
Avg. of #image / Review	4.5
Max of #image / Review	111
Min of #image / Review	1

Baseline Methods

We compare our model with several baseline methods, including the representative textual aspect-level sentiment analysis methods (LSTM, AEAT-LSTM, MemNet, IAN and RAM) [2] and a variant of the state-of-the-art multimodal sentiment analysis method (Co-Memory+ Aspect).

(1) **LSTM** (Wang et al. 2016) adopts an LSTM model to learn the context information of text sequence. It outputs the hidden state of each word and takes the average of these hidden states as the final representation of the whole text.

(2) **AEAT-LSTM** (Wang et al. 2016) appends the aspect embeddings with each word embedding to strengthen the effect of aspect in the process of generating hidden

[2] As the models of TD-LSTM (Tang et al. 2016) and Cabasc (Liu et al. 2018) require extra structure (i.e. aspect words surrounded by left context and right context), we did not compare our model with these two methods.

states. After that, it also focuses on the keywords associated with the given aspect by combining word hidden states with aspect embedding in the attention layer.

(3) **MemNet** (Tang, Qin, and Liu 2016) is a memory model which regards aspect embedding as the query vector to generate deep memory using multiple attention mechanisms on those memories stacked by input word embeddings. The last output of attention layer is fed into a softmax layer for aspect-level sentiment prediction.

(4) **IAN** (Ma et al. 2017) is an LSTM based model which includes two LSTMs to represent aspect and text context respectively and the interactive attention mechanism to interactively learn attentions in the aspect and contexts. It concatenates the aspect attention's output and context attention's output for aspect-level sentiment classification.

(5) **RAM** (Chen et al. 2017) is a memory based model which builds memory on the hidden states of a Bi-LSTM and generates aspect representation also based on a Bi-LSTM. The outputs of its multiple attention layers are non-linearly combined with a recurrent neural network to strengthen the expressive power for global memory abstraction.

As there is no related work on aspect-based multimodal sentiment analysis, we introduce the aspect information to Co-Memory (Xu, Mao, and Chen 2018), the state-of-the-art model for multimodal sentiment analysis, and construct its variant to further evaluate the performance of our model.

(6) **Co-Memory+Aspect** is the variant of Co-Memory. In addition to the co-memory attentional mechanism to interactively model the interaction between textual and visual memories, it introduces the average of aspect embeddings as the input of textual and visual memory networks.

Implementation Details

We compare our model and baseline methods based on our Multi-ZOL dataset. We randomly divide this dataset into training set (80%), development set (10%) and test set (10%). Each sample consists of a textual content, several images, and one aspect. For textual contents and aspects, we first use Jieba[3] Chinese Word segmentation tool for word segmentation. Then all the word embeddings are initialized to the 300-dimension vectors by SGNS (Li et al., 2018) pre-trained on Baidu Encyclopedia corpus and will be fine-tuned during training process to adapt the domain of our aspect based multimodal sentiment analysis. We set the max padding length of textual content L to 320, the max padding length of aspect words N to 4. For images, we resize them to 224*224 and feed them into a pre-trained conventional neural network ResNet50 (He et al. 2016) to extract the 2048-dimension visual feature embeddings. The max padding number of images K is 5. We set the dimension of the LSTM hidden representation D_h to 100, the probability dropout to 0.5, the learning rate to 0.005 and the batch size to 128. Our model has approximately 26M parameters in total. It takes about 40 seconds to train it for each epoch with one Titan X GPU.

[3]https://github.com/fxsjy/jieba

Experimental Results

The evaluation metrics used in our experiment are accuracy and macro-F1. Table 2 illustrates the performance comparison of our MIMN mdoel with the baseline methods. For fair comparison those text based methods (LSTM, AEAT-LSTM, MemNet, IAN, RAM) only use the textual data in Multi-ZOL, and Co-Memory+Aspect and our model use multimodal data in Multi-ZOL. We can see that our proposed model outperforms all the comparative methods with the best accuracy 61.59% and macro-F1 60.51%.

LSTM performs poorly, because it makes no clear distinction between aspects and context words by treating them equally. Besides, averaging context hidden representations may ignore those informative sentiment words.

The LSTM based models ATAE-LSTM and IAN both exceed LSTM with a significant improvement. For ATAE-LSTM, the introduction of the aspect guided attention mechanism as well as the incorporation of the input word embedding and aspect embedding emphasizes more reasonable hidden representation for aspect-level sentiment analysis. Compared with ATAE-LSTM, IAN performs better because it not only represents aspect and context respectively in different recurrent neural networks, but also interactively learns attentions in context and aspect.

The memory network based models MemNet and RAM also achieve better results than LSTM. MemNet absorbs aspect and word memory at the input memory hop and uses multiple memory hops to extract deeper attentional representation. The RAM performs best among all those traditional baseline methods based on textual modality data. It combines the advantages of recurrent network in considering context information and multiple attention mechanism in recording informative information. Besides, the non-linear combination of multiple attentions takes full account of all the memory results.

The Co-Memory+Aspect is similar to MemNet but introduces another modality data, i.e. image, and fully considers the interaction between text and image. Thus, it performs better than all the baseline methods mentioned above. The result shows the effectiveness of introducing image modality data into traditional text based aspect-level sentiment analysis task.

Table 2: Comparative Results of MIMN and Baselines

Method	Accuracy	Macro-F1
LSTM	58.92	57.29
MemNet	59.51	58.73
ATAE-LSTM	59.58	58.95
IAN	60.08	59.47
RAM	60.18	59.68
Co-Memory+Aspect	60.43	59.74
MIMN	**61.59**	**60.51**

Compared with the traditional text based methods, our model uses multimodal data, text and image, for aspect-level sentiment analysis and fully captures the impacts that aspect brings to text and image, as well as the interactions in and between text and image. Compared with the Co-Memory+Aspect method, our multi-interactive attention mechanism learns not only the interactive influence caused by cross-modality data, but also the self influence caused by single-modality data. In addition, our model builds multimodal memories by recurrent neural network to grasp phrase-like features in texts and the subset of images. The recurrent memory network could also strengthen the memory attribute for global memory abstraction. Therefore, our MIMN model obtains the best performance among all the baseline methods.

Effects of Memory Hops

Our model is a memory network based method. Hence, the number of memory hops is one major hyper-parameter affecting the performance. Table 3 shows the experimental results for our model with 1 to 5 memory hops, where MIMN (t) means MIMN using t memory hops. Here, MIMN (1) has none recurrent memory operation for only using one memory hop. Thus, it is not sufficient to represent multimodal data adequately, and leads to the worst performance. As the number of memory hops increases, the performance of MIMN gets better. The results show that our model with 3 memory hops achieves the best result. However, the performance does not continue to increase because of the incremental complexity and decreasing generalization capability of our model with the growing number of hops.

Table 3: Results of Different Memory Hops

# Hops	Accuracy	Macro-F1
MIMN (1)	60.08	59.44
MIMN (2)	60.43	60.10
MIMN (3)	**61.59**	**60.51**
MIMN (4)	60.78	60.05
MIMN (5)	60.11	59.92

Effects of Different Variants of Multi-Interactive Attention Mechanism

To verify the effectiveness of our multi-interactive attention mechanism, we design a series of variants by replacing the multi-interactive attention mechanism used in our model. Table 4 shows the performances of the following variants of our attention mechanism. For fair comparison, all these variants have 3 memory hops.

- **MIMN-self** does not consider the interaction between texts and images by removing the cross-modality attention and just use self modality information to guide the attention mechanism.
- **MIMN-text2image** removes the image2text attention on the basis of MINM model and models the influence that text brings to image using the text2image attention.
- **MIMN-image2text** removes the text2image attention on the basis of MINM model and models the influence that image brings to text using the image2text attention.

The MIMN-self model gets the worst results among all the variants. It removes the bidirectional interaction between text and image and ignores the influences that textual and visual have on each other. As for MIMN-text2img and MIMN-img2text models, they both outperform MIMN-self and are worse than MIMN for only considering unidirectional interaction between texts and images. Our MIMN synchronously learns not only the interactive influences caused by cross-modality data but also the self influences of single-modality data. The results show that the self and unidirectional attentions cannot fully capture the interactive correlations between text and image, and are insufficient for multimodal data representation.

Table 4: Results of Different Variants of Multi-Interactive Attention Mechanism

Method	Accuracy	Macro-F1
MIMN-self	60.50	59.48
MIMN-text2image	61.17	60.27
MIMN-image2text	61.23	60.36
MIMN	**61.59**	**60.51**

Conclusions

In this paper, we propose the new task of aspect based multimodal sentiment analysis by firstly introducing the image modality data to the traditional text based aspect-level sentiment analysis. To capture the impacts that aspect brings to text and image, as well as the multiple interactions associated with text and image, we propose a novel multi-interactive memory network for this task. Our model uses two memory networks to separately model text and image data and learns not only the interactive influences between cross-modality data but also the self influences in single-modality data. Multiple memory hops are used for multi-interactive attentions extraction, which is non-linearly combined by the recurrent neural network to learn the global memory abstraction. We also build a new multimodal aspect-level sentiment dataset to evaluate our model. The results show that our proposed model outperforms the comparative baseline methods and demonstrate the capability of our model in capturing multiple correlations on aspect based multimodal sentiment data.

Acknowledgments

This work is supported in part by the Ministry of Science and Technology of China under Grant #2016QY02D0305, National Natural Science Foundation of China under Grants #71621002, #11832001, #61671450, and Chinese Academy of Sciences under Grant #ZDRW-XH-2017-3.

References

Borth, D.; Ji, R.; Chen, T.; Breuel, T.; and Chang, S.-F. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, 223–232.

Chen, P.; Sun, Z.; Bing, L.; and Yang, W. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 452–461.

Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; and Xu, K. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, 49–54.

Dong, L.; Wei, F.; Zhou, M.; and Xu, K. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 260–269.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 151–160.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiritchenko, S.; Zhu, X.; Cherry, C.; and Mohammad, S. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, 437–442.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436.

Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; and Du, X. 2018. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Liu, Q.; Zhang, H.; Zeng, Y.; Huang, Z.; and Wu, Z. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 World Wide Web Conference*, 1023–1032.

Ma, D.; Li, S.; Zhang, X.; and Wang, H. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4068–4074.

Mohammad, S. M.; Kiritchenko, S.; and Zhu, X. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Proceedings of the 16th International Conference on Data Mining*, 439–448.

Poria, S.; Cambria, E.; and Gelbukh, A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2539–2544.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 3104–3112.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2016. Effective lstms for target-dependent sentiment classification. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, 3298–3307.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 214–224.

Vo, D.-T., and Zhang, Y. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of 24th International Joint Conference on Artificial Intelligence*, 1347–1353.

Wagner, J.; Arora, P.; Cortes, S.; Barman, U.; Bogdanova, D.; Foster, J.; and Tounsi, L. 2014. Dcu: Aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, 223–229.

Wang, Y.; Huang, M.; Zhao, L.; et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615.

Xu, N., and Mao, W. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2399–2402.

Xu, N.; Mao, W.; and Chen, G. 2018. A co-memory network for multimodal sentiment analysis. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 929–932.

Yu, Y.; Lin, H.; Meng, J.; and Zhao, Z. 2016. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* 9(2):41.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zhang, M.; Zhang, Y.; and Vo, D.-T. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 3087–3093.