A Co-Memory Network for Multimodal Sentiment Analysis

Nan Xu
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences, Beijing, China
xunan2015@ia.ac.cn

Wenji Mao
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences, Beijing, China
wenji.mao@ia.ac.cn

Guandan Chen
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences, Beijing, China
chenguandan2014@ia.ac.cn

ABSTRACT

With the rapid increase of diversity and modality of data in user-generated contents, sentiment analysis as a core area of social media analytics has gone beyond traditional text-based analysis. Multimodal sentiment analysis has become an important research topic in recent years. Most of the existing work on multimodal sentiment analysis extracts features from image and text separately, and directly combine them to train a classifier. As visual and textual information in multimodal data can mutually reinforce and complement each other in analyzing the sentiment of people, previous research all ignores this mutual influence between image and text. To fill this gap, in this paper, we consider the interrelation of visual and textual information, and propose a novel co-memory network to iteratively model the interactions between visual contents and textual words for multimodal sentiment analysis. Experimental results on two public multimodal sentiment datasets demonstrate the effectiveness of our proposed model compared to the state-of-the-art methods.

KEYWORDS

Multimodal Sentiment Analysis; Co-Memory Network; Attention

ACM Reference Format:
Nan Xu, Wenji Mao, and Guandan Chen. 2018. A Co-Memory Network for Multimodal Sentiment Analysis. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210093

1 INTRODUCTION

Social media has become the main platform for people to share information and express their personal opinions. The diversity and modality of data in user-generated contents are increasing rapidly in recent year, leading to the ever-increasing of the proportion of visual contents in social media. Sentiment analysis as a core area of social media analytics has gone beyond traditional text-based analysis, and image information has proven an important source to analyze the sentiment of people [12].

In contrast to traditional single modality based sentiment analysis, recent studies attempt to recognize sentiment expressed in multimodal data streams [1–3, 8, 10, 11, 13]. Although multimodal sentiment analysis is still in its infancy, it has demonstrated great potential in research and industry investment [8]. Early work adopts feature-based methods. For example, Borth et al. [2] extract 1200 adjective-noun pairs from image and calculates the textual sentiment scores based on English grammar and spelling style of text. Feature engineering required in these methods is painstakingly detailed, biased, and labor-intensive.

With the booming of deep learning, neural network based models [1, 3, 10, 11, 13] have been proposed for multimodal sentiment analysis, with significant progress in performance. Baecchi et al. [1] train word embeddings by continuous bag-of-words and concatenate them with the mid-level representation generated by denoising autoencoder for multimodal sentiment classification. Inspired by the performance of convolutional neural network in image and text classification tasks, the work in [3] and [13] employs pre-trained text CNN and image CNN to extract feature representations of text and image separately for sentiment analysis. To incorporate the semantic information contained in image, Xu et al. [10] extracts image captions which include detailed semantic visual information, and then processes tweet texts and image captions synchronously based on RNN. They further consider the influence of image to text by extracting visual features of scene and object from image, and absorb text words with these visual features in an attentional LSTM model [11].

These previous work on neural network based models typically extracts features from image and text separately and then combine them directly to train the multimodal sentiment classifier, with the only exception of [11]. Although the work in [11] considers the single-direction influence of image to text, it still ignores the mutual reinforcing and complementary characteristics between visual and textual information, and in general lacks a fine-grained architectural framework to handle the interactions of multimodal contents. To address the above issue, in this paper, we propose a co-memory network for multimodal sentiment analysis, the key structure of which models the bi-directional interactions of image and text. The contributions of our work are as follows:

- We propose a novel co-memory network, which is the first time to model the mutual influence of visual and textual information in multimodal sentiment analysis.
- We design a co-memory attentional mechanism to capture the interactions of visual contents and textual words, and iteratively feed text information for finding visual key contents and image information for locating textual keywords.
- Experimental results on two public multimodal sentiment datasets demonstrate that our proposed model outperforms the state-of-the-art methods.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210093

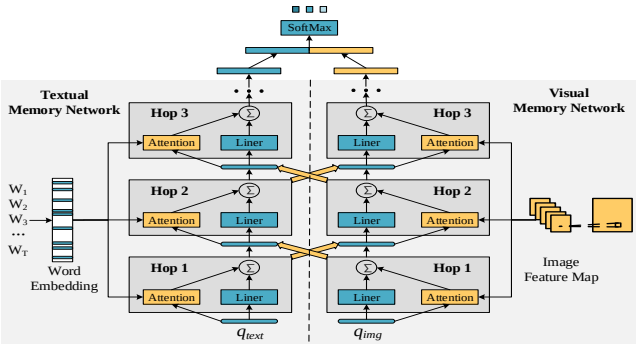

Figure 1: Architecture of co-memory network with memory hops for multimodal sentiment analysis

2 PROPOSED MODEL

The overall architecture of our co-memory network with several memory hops for multimodal sentiment analysis is shown in Fig. 1. In the bottom level of memory hop, we extract feature representation from text and image data respectively. Then in next memory hop, we feed image feature to query keywords in textual memory network and text feature to query key contents of image in visual memory network. The interactions of textual and visual features go up through the co-memory network. After several memory hops, we concatenate the last queried textual and visual feature representation for sentiment classification.

2.1 Feature Extraction

2.1.1 Image Features. Images are highly associated with sentiments. For example, a picture with love sign, rose, wedding ring or chapel would bring about positive sentiment, while a picture describing ghost, war or disaster might lead to negative sentiment. In fact, when people look at an image, they typically focus on the interesting parts instead of the whole image. In other words, not all pixels in a picture contribute equally to image representation for sentiment analysis. Thus we propose a visual memory network to focus on such crucial contents of image for sentiment and aggregate the representations of these key contents by our attention model.

We firstly use a pre-trained convolutional neural network in order to extract a set of visual feature vectors which we refer to as visual memory pieces. In order to obtain a correspondence between the feature vectors and portions of the 2-D image, we extract visual feature maps from a lower convolutional layer unlike previous work using the output of top fully connected layer [11]. The visual extractor generates L feature maps, each of which is a $N * N$ tensor. We flatten each feature map into D_{img}-dimensional feature vector m_i which is corresponded to a part of an image.

$$M = \{m_1, m_2, ..., m_L\}, m_i \in \mathbb{R}^{D_{img}} \tag{1}$$

Then in our visual memory network, we stack these feature map vectors and represent them as the external memory matrix M. Taking the external visual memory M and the visual query vector q_{img} as the input of our visual memory network, we feed each piece of visual memory m_i through a single-layer perceptron to get the image hidden representation h_i.

$$h_i^1 = tanh(w_{img}^1 m_i + b_{img}^1) \tag{2}$$

We measure the interest of visual feature maps as the similarity of h_i^1 with the visual query vector q_{img}. Then we normalize it to compute visual attention weight α_i through a softmax function as

$$\alpha_i^1 = \frac{exp({h_i^1}^T q_{img})}{\sum_{j=1}^{L} exp({h_j^1}^T q_{img})} \tag{3}$$

where the visual query vector $q_{img} \in^{D_i mg}$ is randomly initialized and jointly learned during the training process.

Lastly, our attention model outputs a continuous image representation vector $v_{img}^1 \in \mathbb{R}^{D_{img}}$ to selectively focus on certain contents of an image by weighting a subset of all visual feature vectors.

$$v_{img}^1 = \sum_{i=1}^{L} \alpha_i^1 m_i \tag{4}$$

2.1.2 Text Features. Similarly, not all words contribute equally to the representation of text for sentiment analysis. Therefore, we propose a textual memory network to extract such important words for sentiment and aggregate the representation of those informative words by our attention model. Given a sentence $s = \{w_1, w_2, ..., w_T\}$, each word w_t is initialized by the word representation method Glove [7] to generate a D_{text}-dimensional word embedding x_t.

$$X = \{x_1, x_2, ..., x_T\}, x_t \in \mathbb{R}^{D_{text}} \tag{5}$$

These word embeddings are stacked and regarded as the external memory matrix X. We input the external textual memory X and the textual query vector q_{text} to our textual memory network. We get the word hidden representation g_t using each piece of textual memory x_t through a single-layer perceptron.

$$g_t^1 = tanh(w_{text}^1 x_t + b_{text}^1) \tag{6}$$

We then quantify the importance of word embeddings as the similarity of g_t^1 with the textual query vector q_{text}. The textual attention weight β_t is computed by

$$\beta_t^1 = \frac{exp({g_t^1}^T q_{text})}{\sum_{r=1}^{T} exp({g_r^1}^T q_{text})} \tag{7}$$

where the textual query vector $q_{text} \in \mathbb{R}^{D_{text}}$ is also randomly initialized and jointly learned during the training process.

The attention model eventually outputs the textual representation vector $v_{text}^1 \in \mathbb{R}^{D_{text}}$ to pay attention to these informative words in text by weighted average of the word embedding vectors based on textual attentional weight.

$$v_{text}^1 = \sum_{t=1}^{T} \beta_t^1 x_t \tag{8}$$

2.2 Co-Memory Network

Visual and textual information mutually reinforce and complement each other in sentiment analysis. To capture the mutual influence between these two modality data, we propose a co-memory network that fully considers the interactions between image and text. Similar to [5] use two memory networks to handle two domain data, our model includes two memory networks to conduct two modality data. Unlike previous works simply concatenating image and text features [3, 13], our co-memory attentional mechanism uses text information to facilitate the finding of key feature maps, and image information for locating textual keywords at each memory hop.

2.2.1 Text-guided Visual Memory Network (TgVMN). We absorb text representation vector to find key feature maps of image. Firstly, we concatenate text representation vector v_{text}^1 with each visual feature map vector m_i. Then we compute text-guided visual hidden

representation h_i^2 by taking new visual memory $[m_i, v_{text}^1]$ as the input of a single-layer perceptron. A softmax function is used to output a new normalized visual attentional weight α_i^2. After that, we update image representation vector v_{img}^2 by weighting a subset of all visual feature vectors based on the new attentional weight.

$$h_i^2 = tanh(w_{img}^2[m_i, v_{text}^1] + b_{img}^2) \tag{9}$$

$$\alpha_i^2 = \frac{exp(h_i^2)}{\sum_{j=1}^{L} exp(h_j^2)} \tag{10}$$

$$v_{img}^2 = \sum_{i=1}^{L} \alpha_i^2 m_i \tag{11}$$

2.2.2 Image-guided Textual Memory Network (IgTMN). Meanwhile, we also incorporate image representation vector to query keywords of text. We first concatenate image representation vector v_{img} with each word embedding x_t. Then we feed each piece of new textual memory $[x_t, v_{img}^1]$ through a single-layer perceptron to get the image-guided textual hidden representation g_t^2. We normalize new textual attention weight β_t^2 through a softmax function. Finally, we compute the weighted average of word embedding vectors using the new attentional weight to update the text representation vector v_{text}^2.

$$g_t^2 = tanh(w_{text}^2[x_t, v_{img}^1] + b_{text}^2) \tag{12}$$

$$\beta_t^2 = \frac{exp(g_t^2)}{\sum_{r=1}^{T} exp(g_r^2)} \tag{13}$$

$$v_{text}^2 = \sum_{t=1}^{T} \beta_t^2 x_t \tag{14}$$

2.2.3 Stacked Co-Memory Network. It has been demonstrated that neural network models with multiple processing layers are capable to learn deep representations of data with multiple levels of abstraction [4]. Hence, we propose the stacked co-memory network to explore subtle relationships between text and image by iteratively querying the original memory matrices of image and text by the newly updated representations in next multiple memory hops.

Formally, for the k-th memory hop, we incorporate previous text representation vector v_{text}^{k-1} with image feature map vectors and previous image representation vector v_{img}^{k-1} with word embedding vectors respectively as the inputs of attention layers.

$$v_{text}^k = IgTMN([x_t, v_{img}^{k-1}]) \tag{15}$$

$$v_{img}^k = TgVMN([m_i, v_{text}^{k-1}]) \tag{16}$$

where $k \in [2, K]$, K is the number of memory hops.

2.3 Sentiment Classification

After several co-memory hops, we combine the final feature representation vectors of image and text as the input of a softmax layer for sentiment classification.

$$y = Softmax(w_s[v_{text}^K, v_{img}^K] + b_s) \tag{17}$$

Our model minimizes cross entropy loss with RMSProp update rule. During the training process, the dropout and early-stopping tricks are also employed to avoid model overfitting.

3 EXPERIMENTS

3.1 Datasets and Setup

We evaluate our model using two public multimodal sentiment datasets: MVSA-Single and MVSA-Multi [6].

MVSA-Single has 5129 samples labeled by a single annotator. MVSA-Multi has 19600 samples labeled by three annotators. Each sample of the both datasets is a text-image pair tweet collected from Twitter. For fair comparison, we process the original MVSA datasets based on the same approach used in [11] to remove noise tweets, in which the textual label and visual label are inconsistent. The datasets are randomly divided into training set, development set and test set by using the split ratio 8:1:1. For texts, each word is embedded into 100-dimensionality and each sentence is padded into 50. For images, we resize them to 224*224 and feed them into a pre-trained InceptionV3 [9] network to extract 2048 visual feature maps with the size 5*5. We train our model with learning rate 0.0005 and mini-batch 128.

3.2 Baselines

We compare our model with the following baseline methods. All these methods have been shown to be superior to traditional single modality sentiment analysis methods.

(1) **SentiBank+SentiStrength** [2] extracts 1200 adjective-noun pairs of image based on SentiBank and calculates the sentiment scores of text by SentiStrength.

(2) **CBOW+DA+LR** [1] concatenates word embeddings with the mid-level representation visual encoder generated at each context every window for multimodal sentiment classification.

(3) **CNN-Multi** [3] uses text CNN and image CNN to extract the feature representations of text and image and joints the feature vectors as the input of multi CNN to classification.

(4) **DNN-LR** [13] utilizes pre-trained CNNs to extract text and image feature vectors respectively and combines them as the input of logistics regression.

(5) **HSAN** [10] gets image semantic captions and proposes a hierarchical attentional network to process tweet texts and image captions synchronously.

(6) **MultiSentiNet** [11] extracts semantic visual information, scene and object, from image and proposes a visual feature guided attention LSTM model to absorb the text words with these visual semantic features. It performs the state-of-the-art on multimodal sentiment analysis task.

(7) **MN-Hop1, MN-Hop2, MN-Hop2+text2img** and **MN-Hop 2+img2text** are four variants of our proposed model. MN-Hop1 directly concatenate the output of hop1 for sentiment classification. MN-Hop2 has two memory hops without using co-memory mechanism. MN-Hop2+text2img only uses TgVMN in hop2 and MN-Hop2+img2text only uses IgTMN in hop2.

3.3 Experimental Results

We choose accuracy and F1-measure as evaluation measures. Table 1 gives the experimental results of baselines and our model (and the variants). Our approach is abbreviated to CoMN-Hop K, where K is the number of using co-memory hops.

We can see that feature-based model SentiBank+SentiStrength gets the worst performance than other neural network based methods. CBOW+DA+LR performs better than SentiBank+ SentiStrength by using denoising autoencoder to automatically generate image encoder feature and combine it with word embedding. The CNN-Multi and DNN-LR are both convolutional neural network based model and pre-train CNNs for text and image features extraction respectively. They also perform better than SentiBank+SentiStrength.

Table 1: Experimental Results of Comparative Methods

Method	MVSA-Single		MVSA-Multi	
	Acc	*F1*	*Acc*	*F1*
SentiBank+SentiStrength	52.05	50.08	65.62	55.36
CBOW+DA+LR	63.86	63.52	64.22	63.73
CNN-Multi	61.20	58.37	66.39	64.19
DNN-LR	61.42	61.03	67.86	66.33
HSAN	66.83	66.9	68.16	67.76
MultiSentiNet	69.84	69.63	68.86	68.11
MN-Hop1	64.31	63.12	67.16	66.48
MN-Hop2	64.84	63.96	67.32	66.57
MN-Hop2+text2img	65.19	64.37	67.80	67.01
MN-Hop2+img2text	68.07	65.19	67.92	67.16
CoMN-Hop2	70.07	68.03	68.68	68.06
CoMN-Hop3	69.62	65.95	69.39	68.57
CoMN-Hop4	69.18	68.29	**69.92**	**69.83**
CoMN-Hop5	69.40	69.71	69.68	69.31
CoMN-Hop6	**70.51**	**70.01**	68.92	68.83

When dealing with the bigger dataset MVSA-Multi, these CNN based models work better compared to the denoising autoencoder used in CBOW+DA+LR. The HSAN and MultiSentiNet are recurrent neural network based models and both consider the visual semantic information. They reach higher performance, which demonstrates the effectiveness of RNN in modeling the word sequence as well as the importance of visual semantic information in sentiment analysis. Especially, MultiSentiNet takes into account the influence of visual information to text, and achieves the best performance among all the baselines.

We can also see that our model outperforms all the baselines. Our co-memory network absorbs image representation vector to query keywords in text and incorporates text representation vector to locate key feature maps in image. Moreover, we stack co-memory network to learn deeper feature representation with multiple level hops. The results show that our co-memory network with 6 hops achieves the best performance on MVSA-Single dataset, and co-memory network with 4 hops achieves best on MVSA-Multi dataset. To further verify the importance of modeling interactions between image and text, we generate four variants of our model for comparison. The MN-Hop2 performs better than MN-Hop1 for stacking two memory hops. MN-Hop2+text2img and MN-Hop2+img2text outperform MN-Hop2 for considering the single-direction interaction of image and text, but they still perform worse than Co-MN Hop2. Moreover, compared to MN-Hop2+text2img, the higher performance of MN-Hop2+img2text indicates that the influence of image to text is more effective than that text to image for multimodal sentiment. In general, the experimental results demonstrate the effectiveness of our co-memory in modeling the mutual influence of visual and textual information in multimodal sentiment analysis.

3.4 Attention Visualization

As an example, we visualize the outputs of the last co-memory attention layer with a negative multimodal tweet in Fig. 2. Our co-memory attentional mechanism works well for multimodal sentiment analysis by accurately finding the car bomb and fire in the image and localizing the informative words: "car boom", "martyrs", "wounded" in text.

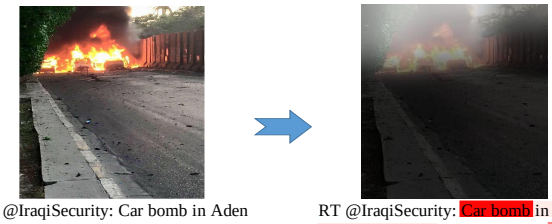

Figure 2: Visualization of our co-memory attention

4 CONCLUSIONS

Visual and textual information in multimodal data can mutually reinforce and complement each other in analyzing the sentiment of people. Compared to previous work which directly concatenates image and text feature vectors, we develop a stacked co-memory network that captures the interactions between visual contents and textual words for multimodal sentiment classification. Our model iteratively uses text information to facilitate the finding of visual key contents, and image information for locating textual keywords. Experimental results demonstrate that our proposed model outperforms the state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grants #2016QY02D0305 and #2016YFC1200702, the National Natural Science Foundation of China under Grants #71621002 and #91546112, and CAS Key Grant #ZDRW-XH-2017-3.

REFERENCES

[1] Claudio Baecchi, Tiberio Uricchio, Marco Bertini, and Alberto Del Bimbo. 2016. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications* 75, 5 (2016), 2507–2525.
[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*. ACM, 223–232.
[3] Guoyong Cai and Binbin Xia. 2015. Convolutional neural networks for multimedia sentiment analysis. In *NLPCC*. Springer, 159–167.
[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.
[5] Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*. 2237.
[6] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MMM*. Springer, 15–27.
[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.
[8] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.
[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*. IEEE, 2818–2826.
[10] Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *ISI*. IEEE, 152–154.
[11] Nan Xu and Wenji Mao. 2017. MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis. In *CIKM*. ACM, 2399–2402.
[12] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*. AAAI Press, 381–388.
[13] Yuhai Yu, Hongfei Lin, Jiana Meng, and Zhehuan Zhao. 2016. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* 9, 2 (2016), 41.