Sentiment Analysis of Online Reviews with a Hierarchical Attention Network

Jingren Zhou[1], Peiquan Jin[1*], Jie Zhao[2*]
[1]School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
[2]School of Business, Anhui University, Hefei, China
jpq@ustc.edu.cn

***Abstract*—Sentiment analysis of online reviews has been an important task in online shopping and electronic commerce to understand customers' opinions and behavior. Generally, customers can express their opinions by posting comments and uploading images, therefore online reviews can be regarded as the combination of a textual document and a few related images. Images may not contain obvious sentimental information, but the visual aspects of images can augment the sentiment information of textual comments. Thus, it is a better way to consider both textual comments and visual images in sentiment analysis of online reviews. Following this idea, in this paper, we propose a hierarchical attention network that combines visual aspect attention, sentence attention, and self-attention to provide effective sentiment analysis of online reviews. With this mechanism, we can model the interactions among words within one sentence as well as the interactions between texts and images. We conduct experiments on a dataset about online restaurant reviews from Yelp.com and compare our model with five existing models. The results suggest the superiority of our proposal.**

Keywords-hierarchical attention network; online review; sentiment analysis

I. INTRODUCTION

The sentiment analysis of web data has been a widely-studied topic [1-5]. Sentiment analysis aims to detect the sentimental polarity, e.g., positive or negative, which is represented by the underlying data. Sentiment analysis is helpful for many web-based applications such as online reviews analysis, product recommendation, and personalized search. For example, in E-commerce platforms, vendors can know the public opinion toward a new product by analyzing the sentimental polarity of online reviews, which can be taken as the basis of market-promoting decisions.

Sentiment analysis used to be a research field in natural language processing, where the sentiment of a document is to be classified into different classes (e.g., positive, negative, and neutral), or a scaling factor (e.g., 1 to 5) is used to measure users' sentiment [2]. Researchers have proposed various ways to extract textual features, which are used as input for supervised learning. The recent works were mostly based on deep neural networks [3-5], which are proven to be highly effective for fitting nonlinear functions.

However, traditional sentiment analysis of online reviews mainly focused on the textual information and highly relied on natural language processing techniques. On the other hand, many E-commerce or online shopping sites have allowed users to post images as part of reviews. Visual images can contain complementary information that is highly connected to the textual information in a review. Thus, in this paper, we aim to figure out the sentiment of online reviews that involve both textual comments and images. One challenge is that images may only contain some objects but no explicit sentiment words like "great" and "wonderful". For example, an image may only show some food on the table in a restaurant. Thus, using visual features directly for sentiment analysis is not effective for online reviews. We noted that in many cases, the attached images in a review usually reflect the main topic of the review. For example, when a person wrote a review about a restaurant, if he posted an image of his meal, it was likely that his review mainly expressed his opinions about food. On the other side, if the image was about the surrounding environment, then we could assign high weights to the restaurant environment related texts in the textual part, which was expected to achieve high accuracy of the sentiment analysis of online reviews.

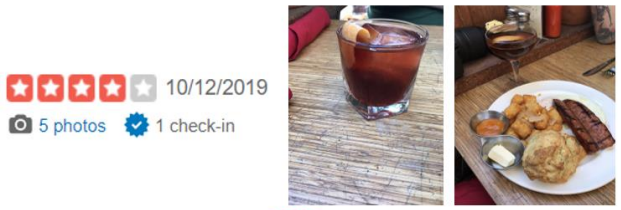

Figure1. Example of the visual-textual connection in *Yelp* reviews.

Following the above idea, we propose to use visual images and the attention mechanism to enhance the performance of sentiment analysis on textual reviews. For image analysis, we propose to adopt visual aspect attention [6], which can help each sentence to find some "aspects" that appear in an image and align the semantic information in the text with visual information in the image. Figure 1 shows an example of *Yelp* restaurant review, two images on the right of the document describe drinks and tater tots separately, while two related sentences state that the drinks are not good and the tater tots are fine. The visual aspect attention can bridge the gap between the visual part in the images with the aspect words in the textual reviews.

On the other hand, there are some limitations for visual aspect attention. As Fig. 2 shows, the content of two images is curry, fried chicken, and the beef soup, which are referred as

* Corresponding author
DOI reference number: 0.18293/SEKE2020-068

Figure 2. Example of deviation of images and texts.

"mediocre at best" in the first sentence, while the second sentence surrounded by a dashed green rectangle box indicated that the service was not good, but this was not directly reflected in the two images.

Briefly, this study aims to present a new method for the sentiment analysis of online reviews by combing visual aspect attention with sentence attention, yielding a hierarchical attention network model. The contributions of this study can be summarized as follows:

(1) We propose a hierarchical attention network for online-review-oriented sentiment analysis, which can both process images and texts in a review to generate effective sentiment polarity. Specially, we adopt self-attention as the base encoding layer to catch the interactions among words with long distances for information augment. We also use the visual aspect attention

(2) We present a dense layer to concatenate document representations generated by visual aspect attention and sentence attention to avoid the neglecting of some important sentences in sentiment analysis.

(3) We conduct experiments on a real dataset about online restaurant reviews from Yelp.com and compare our model with various baselines. The experimental results show that our method can improve the sentiment recognition accuracy of reviews, and can generalize to other scenarios where images reflex the main content of texts.

The remainder of the paper is structured as follows. Section II provides a brief literature review on recent research progress. Section III describes the proposed model. Section IV reports the experimental results, and finally, we conclude the entire paper in Section V.

II. Related Work

Previous works on sentiment analysis have focused on text classification [1], where variants of general classifying techniques are applied and deep learning also brings advancement. In recent years, neural networks like recurrent neural networks have achieved success by incorporating attention into natural language process tasks, these models are usually hierarchical, e.g., word encoding and sentence encoding.

Dimension based sentiment analysis [2] means not to analyze the general sentiment of some document but to detect the sentiment of each dimension or aspect mentioned in the document. The state-of-the-art approach for aspect-level sentiment analysis is attention based deep learning systems. We focus on the sentiment of the whole document instead of producing prediction for relevant aspects in images.

Recent works show that sentiment analysis can use information from more than one modality, e.g., text, acoustic, image, which is referred to as multimodal sentiment analysis, while this paper tries to work out sentiment analysis of online reviews involving text and image. Katsurai et al. [3] proposed mapping textual, and sentiment views into the latent embedding space, then mining correlations among these features. The visual features can be learned from color histograms of images and this method achieved success on Flickr dataset and Instagram dataset. Zhang et al. [4] tried to solve sentiment analysis on microblogging by integrating text features and image features into multiple kernel learning. You et al. [5] proposed to extract visual features with CNN and extract textual features from an unsupervised language model by learning distributed representations for documents and paragraphs, then to fuse these two modalities.

Truong et al. [6] proposed to incorporate images as attention for review-based sentiment analysis. They adopted an architecture of word encoder and sentence encoder, and used visual aspect attention to decide the weight of each sentence. Karpathy et al. [7] proposed a combination of CNNs over image regions, bidirectional RNNs, and a structured objective to align language and visual data into a multimodal embedding. Peng et al. [14] proposed using a visual-textual bi-attention mechanism for visual-textual alignment, their model tries to learn multi-level visual-textual correlation for enhancing the matched pairs of different media types. Xu et al. [8] proposed using soft deterministic attention and hard stochastic attention for image captioning. Lu et al. [9] proposed a model for name tagging in multimodal social media based on visual attention that provides deeper visual understanding of the decisions of the model. Lu et al. [10] proposed a mechanism that jointly reasons visual attention and question attention for visual question answering.

Differing from existing studies, in this paper we propose a hierarchical attention network that combines visual aspect attention, sentence attention, and self-attention, which can model the inter-word correlations among texts as well as the interactions between texts and images.

III. Architecture of the Hierarchical Attention Network

Reviews are comprised of a collection of documents C. Each document is a sequence of L sentences, s_i, $i \in [1, L]$. Each sentence consists of K words $x_{i,k}$, $k \in [1, K]$. Each document has a set of N images $g_j \in \{g_1, g_2, \cdots, g_N\}$, the vector representation of each image is noted as e_j. The goal of our study is to train a classification function to predict sentiment labels for unseen documents.

Our model is a four-layered hierarchical architecture, as shown in Fig. 3. The bottom layer is the self-attention layer that tries to encode each word vector. The next layer is the word encoding layer with soft attention that encodes word vectors into sentence vectors. The third layer is the sentence encoding layer with visual aspect attention. The top layer is the classification layer for the sentiment label.

The main difference of our model from previous models is that we present a layered attention mechanism based on visual

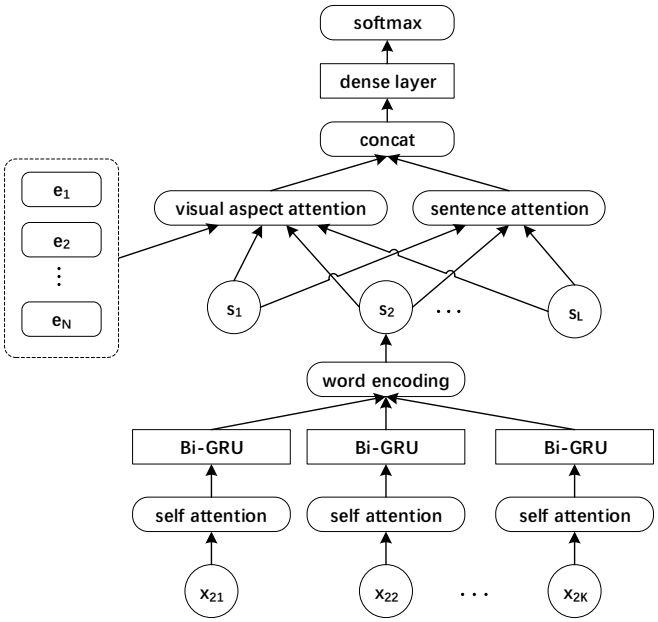

Figure 3. Architecture of the hierarchical attention model.

aspect attention, sentence attention, and self-attention, to integrate the texts and images to enhance the effectiveness of sentiment analysis of online reviews. With our design, both textual and imaginal information of online reviews can be reflected in the sentiment analysis process. By using the hierarchical attention network, especially the self-attention method, we can model the inter-word correlations among texts as well as the interactions between texts and images, resulting in the performance improvement of sentiment analysis of online reviews.

A. Self-Attention

First of all, words must be transferred into embedding vectors as input for the model. We use an embedding matrix W_e initialized from pre-trained word embedding models [17] to retrieve the embedding $x_{i,k}$ of each word $w_{i,k}$.

$$x_{i,k} = W_e w_{i,k}, \mathrm{k} \in [1, \mathrm{K}] \tag{1}$$

Self-attention is an important concept brought in the model Transformer [15] that helped improve the performance of neural machine translation applications. As shown in Fig. 4, self-attention tries to encode representations of other relevant words into the current one being processed, while the relevance degree varies for different words. We use this method to put word interactions into word embedding vectors. For each word, we create a query vector Q, a key vector K and a value vector V. The input of this layer is a sentence matrix composed of word vectors. There are three parameter matrices W^Q, W^K, W^V which are initialized randomly and updated during the training process.

$$Q = X_i W^Q \tag{2}$$

$$K = X_i W^K \tag{3}$$

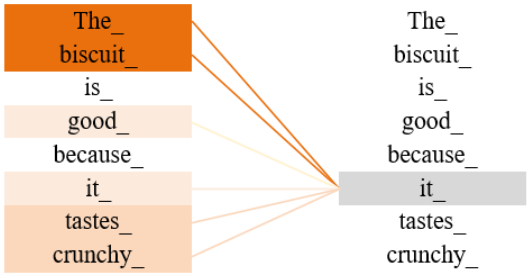

Figure 4. An example of how self-attention works.

$$V = X_i W^V \tag{4}$$

In the self-attention layer, the calculation process of the output Z includes dot product of Q and K, division for scaling, softmax for normalization and getting the weighted sum of V.

$$Z_i = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

The output vector of this layer, which will be used as the new word representation vector in the next layer, is the concatenation of original word vector $x_{i,k}$ and the vector $z_{i,k}$ generated by equation 5, as shown below.

$$y_{i,k} = [x_{i,k}, z_{i,k}]. \tag{6}$$

B. Word Encoding

In this part of the architecture, we try to encode a sentence matrix of word embedding vectors into a sentence embedding vector. The input to this layer is the output of self-attention layer.

We choose bidirectional recurrent neural networks (RNN) [16] with GRU (Gated Recurrent Unit) cell to encode the word embedding sequence, whose output $h_{i,k}$ is the concatenation of $\vec{h}_{i,k}$ generated by the forward GRU and $\overleftarrow{h}_{i,k}$ generated by the backward GRU. This component is noted as Bi-GRU (Bidirectional GRU) in this paper. GRU is a variant of RNN that costs less computation price and solves the problem of long term dependency.

$$h_{i,k} = \text{Bi-GRU}(y_{i,k}) \tag{7}$$

Words are not equally important, and attention helps to assign greater weight to more important words. We employ a soft attention mechanism [6] to distribute weights among words.

$$o_{i,k} = O^T \tanh\left(W_h h_{i,k} + b_h\right) \tag{8}$$

$$\alpha_{i,k} = \frac{\exp\left(o_{i,k}\right)}{\sum_k \exp\left(o_{i,k}\right)} \tag{9}$$

$$s_i = \sum_k \alpha_{i,k} h_{i,k} \tag{10}$$

We use a tangent function to project $h_{i,k}$ into its representation in the attention space. O is a context vector randomly initialized and updated during training, it is used to multiply the projection for the relative importance $o_{i,k}$ of $h_{i,k}$. Then we use softmax as normalization to obtain attention weight $\alpha_{i,k}$ of $h_{i,k}$. The embedding of sentence s_i is represented as the weighted summation of all its word representations $h_{i,k}$ by attention weights $\alpha_{i,k}$.

C. Visual Aspect Attention and Sentence Attention

This layer transfers the output of the word encoding layer into document-level representations using visual aspect attention and sentence attention, assigning greater weight to more salient sentences. We still use Bi-GRU taking sentence s_i as input to generate forward hidden states vector $\vec{h}_i$ and backward hidden states vector $\overleftarrow{h}_i$ and concatenate them as the output vector h_i.

$$h_i = \text{Bi-GRU}(s_i) \tag{11}$$

We take semantic connections between images and text as attention for sentences. A document is usually attached with several pictures, which are associated with sentences with varying degrees.

First, we need to encode visual images and VGG convolutional neural networks [18] have proven effective for learning image presentations in many similar situations. We employ VGG-16 to process image g_j and take the output of the last fully-connected layer (FC7) as the image representation e_j.

$$e_j = VGG(g_j) \tag{12}$$

Visual Aspect Attention. We learn the attention weights $\gamma_{j,i}$'s for sentence representations h_i's with respect to each image representation e_j.

$$p_j = \tanh\left(W_p e_j + b_p\right) \tag{13}$$

$$q_i = \tanh\left(W_q h_i + b_q\right) \tag{14}$$

$$m_{j,i} = M^T(p_j \odot q_i + q_i) \tag{15}$$

$$\gamma_{j,i} = \frac{\exp\left(m_{j,i}\right)}{\sum_i \exp\left(m_{j,i}\right)} \tag{16}$$

We project image representation e_j and sentence representation h_i into an attention space followed by a non-linear activation function to obtain output p_j and q_i. Then we try to find interactions between p_j and q_i by element-wise multiplication and summation. The learned vector M is a global attention vector similar to O in word encoder. Then we use softmax to normalize each attention value $m_{j,i}$ in M as $\gamma_{j,i}$.

With the visual-aspect attention weight $\gamma_{j,i}$, we aggregate sentence representations into document representation d_j as follows.

$$d_j = \sum_i \gamma_{j,i} h_i \tag{17}$$

Each document has a set of image-specific document representation d_j, $j \in [1, N]$. Attached images are not equally informative, thus we try to learn the importance weight τ_j of each document representation d_j. The visual attention-based document representation d' is the aggregation of image-specific document representation d_j.

$$a_j = A^T \tanh\left(W_a d_j + b_a\right) \tag{18}$$

$$\tau_j = \frac{\exp\left(a_j\right)}{\sum_j \exp\left(a_j\right)} \tag{19}$$

$$d' = \sum_j \tau_j d_j \tag{20}$$

Sentence Attention. We use sentence attention to generate a context vector U and reward sentences that are clues to classify a document correctly.

$$u_i = U^T tanh(W_u h_i + b_u) \tag{21}$$

$$\pi_i = \frac{exp(u_i)}{\sum_i exp(u_i)} \tag{22}$$

$$d'' = \sum_i \pi_i h_i \tag{23}$$

The concatenation of d' and d'' is fed into a dense layer, the output of which is the final document representation d.

$$d = \text{Dense}([d', d'']) \tag{24}$$

D. Sentiment Classification

The top layer treats the document representation d with a softmax based sentiment classifier, generating the probabilities distribution μ of sentiment classes.

$$\mu = \text{softmax}(W_\mu d + b_\mu) \tag{25}$$

The loss of this model is the cross-entropy error of sentiment classification:

$$\text{loss} = -\sum_d log\mu_{d,l}\,, \tag{26}$$

where l is the ground-truth label of review d.

IV. Performance Evaluation

A. Settings

Dataset. We use a dataset of restaurant reviews on Yelp.com [6], covering five US cities including Boston (BO), Los Angeles (LA), Chicago (CH), New York (NY), and San Francisco (SF). The dataset contains more than 44 thousand reviews and 244 thousand images with each review having at least 3 images. We split 80% of the dataset for training, 5% for validation, and 15% for tests. There are five classes of sentiment labels in this dataset, ranging from very negative to very positive.

Training. In the training process, we use NLTK [13] for sentence and word tokenization. In addition, we use the pre-trained Glove word embedding with dimensionality D = 200. The GRU cells are 50-dimensional in the encoding process, thus

TABLE I. ACCURACY COMPARISON OF SENTIMENT ANALYSIS (%)

Models	BO	CH	LA	NY	SF	Avg.	Improvement (compared with TFN-aVGG)
TFN-aVGG	46.35	43.69	43.91	43.79	42.81	43.89	-
TFN-mVGG	48.25	47.08	46.70	46.71	47.54	46.87	6.8%
HAN-aVGG	55.18	54.88	53.11	52.96	51.98	53.16	21.1%
HAN-mVGG	56.77	57.02	55.06	54.66	53.69	55.01	25.3%
VistaNet	63.81	65.74	62.01	61.08	60.14	61.88	41.0%
Our model	**66.67**	**68.31**	**60.83**	**61.10**	**61.05**	**63.59**	**44.9%**

the output of bidirectional cells is 100 dimensional. The model is implemented with Python 3.7 and TensorFlow 1.14. We select the Adam optimizer for gradient-based optimization and set the batch size to 32. The model is trained for 20 epochs and the result of the epoch with the least training loss is outputted as the final result.

B. Baselines

We compare our model with several baselines that use both textual and visual features, including TFN-aVGG, TFN-mVGG, HAN-aVGG, HAN-mVGG, and VistaNet. We focus on the accuracy of each model when evaluating the sentiment polarity of the online restaurant reviews.

(1) *HAN-aVGG* and *HAN-mVGG* [11]. HAN-aVGG and HAN-mVGG are composites of HAN-ATT for text and VGG for images. HAN-ATT uses a hierarchical architecture of word encoder and sentence encoder. HAN-aVGG and HAN-mVGG correspond to using average pooling and max pooling for image feature vectors respectively, which will be concatenated with textual feature vectors as the input vectors.

(2) *TFN-aVGG* and *TFN-mVGG* [12]. TFN-aVGG and TFN-mVGG are composites of Tensor Fusion Network. Textual features from HAN-ATT and visual features from VGG are combined using Tensor Fusion Layer and fed through Sentiment Inference Subnetwork for the final sentiment label. We use average pooling for TFN-aVGG and max pooling for TFN-mVGG.

(3) *VistaNet* [6]. VistaNet is a hierarchical architecture that adopts a soft-attention-based word encoding layer and a visual aspect attention based sentence encoding layer.

C. Results

Table I lists the comparative accuracy of our method with other baseline methods. The five columns, namely BO, CH, LA, NY, and SF represent the five cities, and the avg. column is the average accuracy of all the five cities.

As shown in Table I, our model outperforms these multimodal baselines in all five cities and average results. This result demonstrates that combing visual attention and sentence attention could effectively draw attention to more salient sentences of a review document. The second-best model VistaNet is ahead of other baselines, which proves that visual attention has significant effects in this experiment. Our model has a 1.71% accuracy improvement over VistaNet, showing that the self-attention layer is useful for encoding word vectors and a tradeoff between visual aspect attention and sentence attention could end up with better results.

We can also notice that TFN-aVGG and TFN-mVGG perform badly in this experiment even though TFN can provide rich interactions between textual features and visual features. In this experiment of online reviews, images seldom carry enough sentimental information, e.g., images of food cannot tell whether the customer likes the food or not. This is the reason why our model using visual features as attention for sentences can outperform models that use visual features as additional sentimental information.

D. Ablation Analysis

We conduct an ablation analysis to specifically analyze the contributions of each component of our architecture. We start from the most basic architecture and incrementally add a component until reaching the full architecture.

TABLE II. ABLATION ANALYSIS

Components					CITY (%)					
Bi-GRU	Hierarchical Structure	Visual Aspect Attention	Self-Attention	Sentence Attention	BO	CH	LA	NY	SF	Avg.
√	x	x	x	x	57.70	60.01	56.74	56.59	55.84	56.83
√	√	x	x	x	60.39	64.39	59.08	59.58	59.18	59.54
√	√	√	x	x	63.81	65.74	62.01	61.08	60.14	61.88
√	√	√	√	x	63.17	62.77	62.12	61.40	62.63	62.42
√	√	√	√	√	66.67	68.31	60.83	61.10	61.05	63.59

We first carry out experiments with the base model Bi-GRU using only text. Then, we implement a hierarchical structure with a word encoding layer and max-pooling sentence representations, the accuracy is 59.54%. By applying the visual aspect attention upon sentence-level representations, this structure has achieved an average accuracy of 61.88%. When a self-attention encoding layer is added to the hierarchical structure, the average accuracy has increased to 62.42%. Finally, the sentence attention is combined with the visual attention through a dense layer. We can see that the model has been improved to an average accuracy of 63.59%.

All the results in Table II show that our model outperforms other models in the average accuracy. We can also see that every component contributes to our model.

V. Conclusions and Future Work

In this paper, we propose a novel model for text-and-image-based bi-modal sentiment analysis. Our model utilizes visual aspect attention, sentence attention, and self-attention to form a hierarchical attention network, which has been experimentally demonstrated that it works well for the sentiment analysis of online reviews. In particular, we use self-attention as the base encoding layer and combine visual aspect attention with sentence attention to present a better attention mechanism. Compared with existing studies, the four-layered hierarchical attention model can encode the interactions among words within a sentence as well as the interactions between texts and images. It adopts a hierarchical attention mechanism by aggregating word representations into sentence representations, aggregating sentence representations into document representations, and finally generating the sentiment label. Our model also employs images as alignment to select important sentences within a document and employs a soft attention mechanism for sentences that may have few interactions with images. We conduct experiments on a real dataset about online restaurant reviews in five US cities. The results show that our model outperforms the other five baselines, indicating the effectiveness of our proposal.

Our future work will concentrate on building a more elastic attention mechanism, e.g., assigning higher weights to most influential words in a document and introducing most recent models of natural language processing for a better understanding of document content. We will also consider to apply the hierarchical attention model to sentiment analysis of multimodal social media such as microblogs [19, 20].

Acknowledgements

This work is supported by the National Science Foundation of China (no. 61672479 and 71273010) and the National Statistical Science Research Project (no. 2019LY66). Peiquan Jin and Jie Zhao are the joint corresponding authors of this paper.

References

[1] Pang, B. and Lee, L., Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2008, 2(1–2), pp.1-135.

[2] Zheng, L., Jin, P., Zhao, J., Yue, L. Multi-dimensional sentiment analysis for large-scale E-commerce reviews. In *Proceedings of the 25th International Conference on Database and Expert Systems Applications (DEXA)*, 2014, 449-463

[3] Katsurai, M. and Satoh, S.I., March. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, 2837-2841.

[4] Zhang, Y., Shang, L. and Jia, X., Sentiment analysis on microblogging by integrating text and image features. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2015, 52-63.

[5] You, Q., Luo, J., Jin, H. and Yang, J., Joint visual-textual sentiment analysis with deep neural networks. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM)*, 2015, 1071-1074.

[6] Truong, Q.T. and Lauw, H.W., VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, 305-312.

[7] Karpathy, A. and Fei-Fei, L., Deep visual-semantic alignments for generating image descriptions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 3128-3137.

[8] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, 2048-2057.

[9] Lu, D., Neves, L., Carvalho, V., Zhang, N. and Ji, H., Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, 1990-1999.

[10] Lu, J., Yang, J., Batra, D. and Parikh, D., Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*, 2016, 289-297.

[11] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E., Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, 1480-1489.

[12] Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L.P., Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, 1103-1114.

[13] Loper, E. and Bird, S., NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, 63-70.

[14] Peng, Y., Qi, J. and Zhuo, Y., MAVA: Multi-level Adaptive Visual-textual Alignment by Cross-media Bi-attention Mechanism. *IEEE Transactions on Image Processing*, 2020, 29: 2728-2741.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., Attention is all you need. *In Advances in Neural Information Processing Systems (NIPS)*, 2017, 5998-6008.

[16] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1724-1734.

[17] Pennington, J., Socher, R. and Manning, C.D., Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532-1543.

[18] Simonyan, K. and Zisserman, A., Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv: 1409.1556, 2014

[19] Jin, P., Mu, L., Zheng, L., Zhao, J., Yue, L. News feature extraction for events on social network platforms. In *Proceedings of the 26th International World Wide Web Conference (WWW)*, 2017, 69-78

[20] Mu, L., Jin, P., Zheng, L., Chen, E., Yue, L., Lifecycle-based event detection from microblogs. In *Proceedings of the 27th International World Wide Web Conference (WWW)*, 2018, 283-290