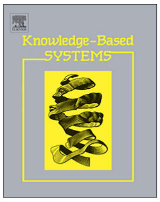
Knowledge-Based Systems 167 (2019) 26–37

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Image–text sentiment analysis via deep multimodal attentive fusion

Feiran Huang [a], Xiaoming Zhang [b,*], Zhonghua Zhao [c], Jie Xu [a], Zhoujun Li [a]

[a] *State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China*
[b] *School of Cyber Science and Technology, Beihang University, Beijing, 100191, China*
[c] *National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029, China*

ARTICLE INFO

Article history:
Received 10 May 2018
Received in revised form 4 January 2019
Accepted 12 January 2019
Available online 16 January 2019

Keywords:
Multimodal learning
Sentiment analysis
Attention model
Fusion

ABSTRACT

Sentiment analysis of social media data is crucial to understand people's position, attitude, and opinion toward a certain event, which has many applications such as election prediction and product evaluation. Though great effort has been devoted to the single modality (image or text), less effort is paid to the joint analysis of multimodal data in social media. Most of the existing methods for multimodal sentiment analysis simply combine different data modalities, which results in dissatisfying performance on sentiment classification. In this paper, we propose a novel image–text sentiment analysis model, i.e., Deep Multimodal Attentive Fusion (DMAF), to exploit the discriminative features and the internal correlation between visual and semantic contents with a mixed fusion framework for sentiment analysis. Specifically, to automatically focus on discriminative regions and important words which are most related to the sentiment, two separate unimodal attention models are proposed to learn effective emotion classifiers for visual and textual modality respectively. Then, an intermediate fusion-based multimodal attention model is proposed to exploit the internal correlation between visual and textual features for joint sentiment classification. Finally, a late fusion scheme is applied to combine the three attention models for sentiment prediction. Extensive experiments are conducted to demonstrate the effectiveness of our approach on both weakly labeled and manually labeled datasets.

1. Introduction

With the increasing popularity of social networks, multimodal data containing both image and text has become increasingly available in social websites (e.g., Flickr, Instagram, Twitter). More and more individuals express their opinions through these media. Sentiment analysis of such large-scale multimodal data can help better understand people's attitude or opinion toward certain events or topics. For example, companies are interested in understanding how their products or brands is perceived among their customers [1–3]. Shareholders are interested in the Twitter mood for predicting the stock market [4–6]. Therefore, how to automatically detect the sentiment of the multimodal data has been increasingly attracting attention in both academia and industry.

However, it is still a challenging task to deal with the multimodal data for sentiment analysis. First, in contrast to traditional single modality sentiment analysis, diverse patterns of manifestation are contained in multimodal sentiment analysis. For example, the visual content and text description are heterogeneous in the feature spaces. Therefore, the sentiment analysis method should effectively bridge the gap between different modalities. Second, the visual content and text description may cover some semantic information different from each other. It is necessary to distill the comprehensive and discriminative information that is most related to sentiment classification from each modality. Third, it is a common phenomenon where one modality is missing in the multimodal data. For example, many users post a tweet without any images and some photographers may upload an image without any text description. It is another challenge to deal with the incomplete multimodal data for sentiment analysis.

Multimodal sentiment analysis has gained increasing attention in recent years. Based on the combination strategies for the multimodal contents, these methods can be categorized into three groups, namely early fusion [7–9], intermediate fusion [10–12] and late fusion [13–15]. The early fusion-based methods integrate multiple sources of data into a single feature vector before being used as the input to a machine learning algorithm. However, early fusion of multimodal data is not effective to capture the complementary nature of the modalities involved and may lead to large input vectors that may contain redundancies. Late fusion refers to the aggregation of decisions from multiple sentiment classifiers, each trained on separate modalities. This method cannot effectively capture the correlation between different modalities. Intermediate fusion, which is mainly implemented with neural networks, refers to that the fusion process is conducted in the intermediate layers of the whole networks. Most of the intermediate fusion

[*] Corresponding author.
E-mail address: yolixs@buaa.edu.cn (X. Zhang).

https://doi.org/10.1016/j.knosys.2019.01.019
0950-7051/© 2019 Elsevier B.V. All rights reserved.

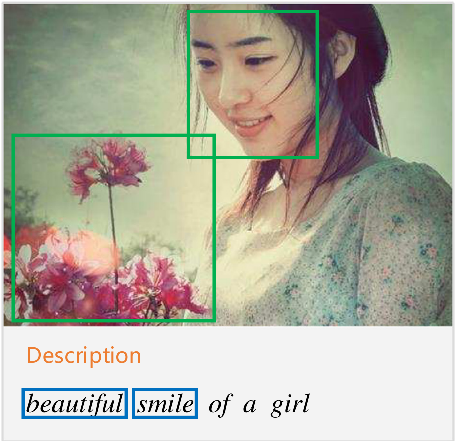


Fig. 1. An example of multimodal data with an image–text pair in Flickr: (1) some regions of image and words of text are more important and discriminative for sentiment analysis. Smiling face and flowers (green box) are more emotional regions. "beautiful" and "smile" (blue box) are more affective words; (2) complementary information exists in respective modality. "beautiful" is text-private information which can hardly be acquired from the image. While flowers and sunshine are image-private information which cannot be captured in the description.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

models adopt a shared representation layer to merge units with connections from multiple modality-specific paths. For example, You et al. [11] proposed a cross-modality consistent regression (CCR) scheme for joint textual-visual sentiment analysis. A tree-structured LSTM (T-LSTM) integrated with visual attention mechanism is also proposed to capture the correlation between image and text for social image sentiment [12]. TFN [16] was proposed for multimodal sentiment analysis by modeling intra-modality and inter-modality dynamics together into a joint framework. Though the intermediate fusion-based methods have achieved state-of-the-art performance in multimodal sentiment analysis, it cannot effectively perform sentiment analysis when a portion of multimodal contents are incomplete.

Meanwhile, the sequential and correlated characteristics of social media data provide clues for multimodal sentiment analysis. First, it is observed that different words in the text content play different roles for sentiment analysis. For example, the words "beautiful" and "smile" in Fig. 1 carry strong emotion, while "of" and "a" can hardly be perceived any sentiment in them. Similarly, not all image regions contribute equally to the presentation of sentiment. As can be seen in Fig. 1, the regions of "smiling face" and "flower" are distinctly more important zones for sentiment analysis. On the other side, the attention mechanism is effective to allow salient features of sequential data dynamically come to the forefront for further applications [17–19]. It is reasonable to use the attention mechanism to automatically focus on the most salient regions and emotional words for more effective sentiment analysis. Second, different data modalities usually contain complementary information. As shown in Fig. 1, "beautiful" is text-private information which can hardly be acquired from the image, while "smiling face" and "flower" are image-private information which cannot be captured in text tags. Obviously, the two kinds of features are complementary to the emotion analysis. It is reasonable to combine both the shared information and modal-specific information for multimodal sentiment analysis.

To tackle the challenges, we take full advantage of attention mechanism and deep fusion scheme for multimodal sentiment analysis. In particular, we investigate: (1) How to automatically discover the most discriminative text words and visual regions for sentiment classification; (2) How to capture the complementary information from multiple modalities for sentiment analysis. Our solutions to these questions result in a novel approach named Deep Multimodal Attentive Fusion (DMAF) for multimodal sentiment analysis. In particular, two separate unimodal attention models are proposed to capture the most discriminative features in image and text for sentiment analysis respectively. The visual attention mechanism is used to automatically focus on the affectional regions, while the semantic attention mechanism is used to highlight the emotion-related words. To exploit the complementary and non-redundant information in different modalities, a deep intermediate fusion-based multimodal attention model is proposed to mine the correlation between the features of different modalities. Then, a late fusion scheme upon the three models, i.e., visual attention model, semantic attention model, and multimodal attention model, is proposed to derive the final decision of sentiment classification. The main contributions are summarized as follows:

- We investigate the problem of multimodal sentiment analysis by exploiting the discriminative features in texts and images and excavating the internal correlations between different modalities.
- We propose a novel multimodal sentiment analysis approach named Deep Multimodal Attentive Fusion (DMAF). Our method automatically draws focuses on affectional regions and words, which can capture the complementary and non-redundant information for more effective sentiment classification.
- We propose to integrate intermediate fusion and late fusion into a holistic framework for multimodal sentiment analysis. Our approach is more effective to handle the incomplete contents of multimodal data.
- The proposed method is extensively evaluated on 3 real-world datasets. Experiment results demonstrate the superiority of DMAF.

The remainder of this paper is organized as follows. In the next section, the related work is summarized. Section 3 presents the details of DMAF. Then we present the experimental results in Section 4. Finally, Section 5 concludes the paper and outlines the future work.

2. Related work

Sentiment analysis is an important task which has been rapidly developed in recent years. It has been applied to a broad set of applications, including political election prediction [20–22], stock market predicting [4–6], product evaluation [1–3] and movie box-office performance prediction [23,13]. The related works include text sentiment analysis, image sentiment analysis, and multimodal sentiment analysis.

2.1. *Text sentiment analysis*

Text sentiment analysis is a well-studied research area in NLP. These methods can be divided into two groups: lexicon-based methods [24–26,20] and machine learning-based methods [27–30]. Semantic Orientation CALculator (SO-CAL) [25] built a dictionary of words annotated with their semantic orientation (polarity and strength) to analyze text sentiment. In [31], three pruning strategies was proposed to automatically build a word-level emotional dictionary for social emotion detection. In [28], the unsupervised and supervised techniques were combined to leverage both continuous and multi-dimensional sentiment information as well as non-sentiment annotations. In [30], a graph model was proposed to incorporate the co-occurrence information of hashtags to perform the graph-based sentiment classification. Recently, with the development of deep neural network, the deep learning-based methods have attracted great attention [32–34]. For example, Coooolll [33] was proposed for message-level Twitter sentiment classification by combining hand-crafted features with the sentiment-specific word embedding features.

2.2. Image sentiment analysis

Image sentiment analysis is relatively more challenging due to that image is more abstract and subjective. Usually, there are three types of visual features used for image sentiment analysis, i.e., low-level features [35,36], middle-level features [37,38], and high-level features [39–41]. In [36], a topic-based emotion learning method was proposed to integrate low-level visual features with social information. In [37], a traditional machine learning-based method was employed to detect 1200 adjective–noun pairs (ANP) which was considered as the middle-level features, and then a visual sentiment ontology was produced. In [38], a set of mid-level attributes were generated from scene and facial expression dataset. Then, the facial emotion information was combined to describe the visual phenomena in a scene perspective. Motivated by the powerful performance of deep models on extracting high-level image features, Xu et al. [39] transferred VGG networks trained on ImageNet dataset into visual sentiment analysis on the sentiment datasets. PCNN [40] was proposed for image sentiment analysis with progressive training, which leveraged a progressive training strategy and a domain transfer strategy to fine-tune the neural network. You et al. [41] adopted attention mechanism to automatically discover the related visual regions for the detected attributes for image sentiment analysis.

2.3. Multimodal sentiment analysis

Multimodal sentiment analysis has gained increasing attention in recent years. There are mainly three types of combination strategies for multimodal sentiment analysis, i.e., early fusion [7–9], intermediate fusion [10–12] and late fusion [13–15,42].

Early fusion integrates multiple sources of data into a single feature vector directly. [8] identified the sentiment expressed in utterance-level visual data streams. It first extracted the linguistic, acoustic, and visual features from multi-view dataset and then combined them directly for sentiment classification. Poria et al. [9] extracted features from visual and textual modalities using deep convolutional neural networks and then fuse the features with a multiple kernel learning classifier for sentiment analysis. However, the early fusion methods cannot fully exploit the complementary nature of the modalities involved and may produce very large input vectors that may contain redundancies.

Late-fusion refers to the aggregation of decisions from multiple sentiment classifiers, each trained on separate modalities. [14] and [15] employed both text and images for sentiment analysis, where late fusion is employed to combine the prediction results of using n-gram textual features and mid-level visual features. Late fusion is based on the assumption that separate modalities are independent in the feature space. This may not be true in practice, as multiple modalities tend to be highly correlated.

Intermediate fusion, which is mainly used with neural networks, refers that the fusion process is conducted in the intermediate layers of the whole networks. A shared representation layer is constructed by merging units with connections from multiple modality-specific paths. In [10], a gated multimodal embedding model was proposed by employing gated mechanism and temporal attention layer to perform sentiment comprehension of text, video, and audio. You et al. [11] proposed a cross-modality consistent regression (CCR) scheme for joint textual-visual sentiment analysis. You et al. [12] proposed an image–text joint sentiment analysis model which integrates tree-structured LSTM (T-LSTM) with visual attention mechanism to capture the correlation between image and text. TFN [16] was proposed for multimodal sentiment analysis by modeling intra-modality and inter-modality dynamics together into a joint framework. Memory Fusion Network (MFN) [43] was recently proposed to associate a cross-view relevance score to each LSTM for multimodal sentiment analysis. However, it used LSTMs to model the interactions within multi-view sequential data, which was not suitable for image feature encoding. Intermediate fusion-based methods have achieved state-of-the-art performance, but their performance may be affected when a portion of multimodal contents are incomplete.

3. Deep multimodal attentive fusion

The details of the overall architecture of the proposed model Deep Multimodal Attentive Fusion are shown in Fig. 2. First, two separate unimodal attention models are proposed to learn the most discriminative features in image and text respectively. The visual attention mechanism is used to automatically focus on the affectional regions, while the semantic attention mechanism is used to highlight the most emotional words. Then, a deep intermediate fusion-based multimodal attention model is proposed to exploit the complementary and non-redundant information in different modalities. It employs a multi-layer perceptron to mine the non-linear correlation between different modalities of features. Finally, a late fusion scheme upon the three models, i.e., visual attention model, semantic attention model, and multimodal attention model, is proposed to obtain the final decision of sentiment classification.

In the following subsections, we first propose two separate attention model, i.e., visual attention model and semantic attention model. Then we introduce the multimodal attention model to combine the two modalities with an intermediate fusion scheme. Based on them, a joint deep model is proposed to integrate the three models via late fusion.

3.1. Visual attention model for image sentiment analysis

Usually, a part of visual regions are more related to the emotion. If the features of these visual regions are highlighted, the sentiment classification could be more effective. Due to the success of fine-grained visual representation, visual attention has been proven to be beneficial for many vision related tasks, such as image caption [17,18], visual sentiment analysis [12,41], and representation learning [44]. You et al. [41] adopted attention mechanism to automatically discover the related visual regions of the detected attribute for image sentiment analysis. Different from their works, our visual attention model is formulated in an end-to-end format, which can directly focus on the most emotional regions of images.

Let $\mathcal{V} = \{V_1, V_2, \dots, V_i, \dots, V_n\}$ denote a set of n images. For each image V_i, we use the deep Convolution Neural Networks (CNN) to obtain the image region maps $R_i = \{r_i^1, r_i^2, \dots, r_i^j, \dots, r_i^D\} \in \mathbb{R}^{D\times M}$ as follows:

$$R_i = f_c(V_i; \theta_c), \ R_i \in \mathbb{R}^{D\times M} \tag{1}$$

where θ_c denotes the parameters of the CNN layers, D is the number of image regions, and M is the dimension of the map. In the attention model, a score α_i^j between 0 and 1 is assigned to each image region r_i^j based on its relevance to the sentiment. We empirically use a softmax function to calculate α_i^j as follows:

$$\alpha_i^j = \frac{\exp(e_i^j)}{\sum_{j=1}^{D}\exp(e_i^j)} \tag{2}$$

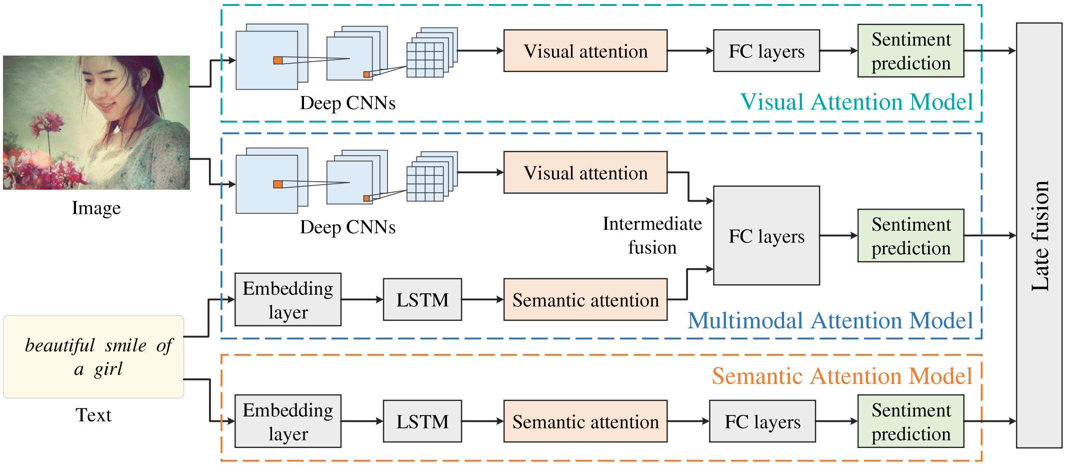


Fig. 2. The framework of DMAF for image–text sentiment analysis.

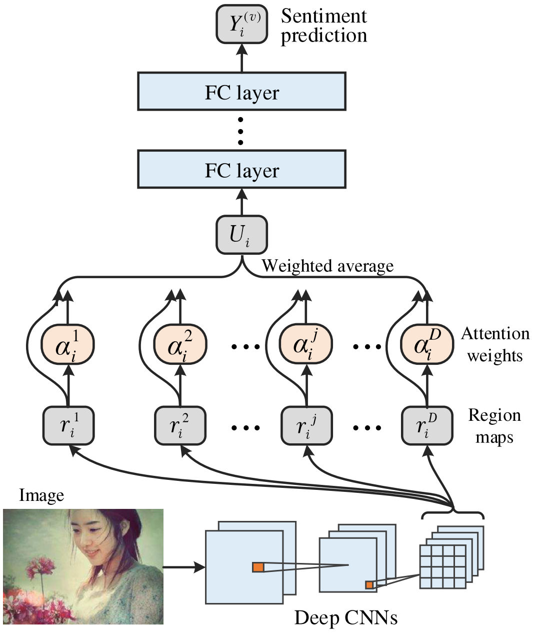


Fig. 3. The architecture of the visual attention model.

where

$$e_i^j = \varphi(Wr_i^j + b) \tag{3}$$

is the unnormalized attention score which measures how well the image region $\{r_i^j\}$ presents the sentiment. The weight matrix W and the bias term b are the parameters to be learned. $\varphi(\cdot)$ is a nonlinear activation function to capture the nonlinear correlation, and usually a *tanh* function is adopted. α_i^js are used to normalize the attentions over all the regions $\{r_i^j\}_{1\leqslant j\leqslant D}$. The attended visual features can be calculated as a weighted average over all the regions:

$$U_i = \sum_{1\leqslant j\leqslant D} \alpha_i^j r_i^j,\ U_i \in \mathbb{R}^M \tag{4}$$

We denote the attention process to automatically generate the attended image features as follows:

$$U_i = f_a(R_i; \theta_a^{(v)}),\ U_i \in \mathbb{R}^M \tag{5}$$

where $\theta_a^{(v)}$ is the weight parameters which consists of W and b in Eq. (3). The equation above behaves somewhat like a weighted average pooling layer over all region features. Compared with the originally independent visual features, the weighted visual feature mapping U_i is more effective to represent the emotion related features. Next, this representation can be supplied as input to a sentiment classifier. The multiple fully-connected (FC) layers is built to perform sentiment classification:

$$Y_i^{(v)} = f_l(U_i; \theta_l^{(v)}),\ Y_i \in \mathbb{R}^C \tag{6}$$

where $\theta_l^{(v)}$ is the parameters of the FC layers and C is the number of sentiment classes. We employ the negative log-likelihood (NLL) to define the cross-entropy loss:

$$\begin{aligned}\mathcal{L}^{(v)}(V; \theta_c, \theta_a^{(v)}, \theta_l^{(v)}) &= \sum_{i=1}^{n} -\log(Y_i^{(v)}) \\ &= \sum_{i=1}^{n} -\log(f_l(f_a(f_c(V_i; \theta_c); \theta_a^{(v)}); \theta_l^{(v)}))\end{aligned} \tag{7}$$

The whole architecture of the visual attention model is illustrated in Fig. 3. The parameters θ_c, $\theta_a^{(v)}$ and $\theta_l^{(v)}$ are automatically learned by minimizing the cross-entropy loss over the training set. We use the stochastic gradient descent (SGD) over the shuffled mini-batches with an adaptive learning rate to optimize the loss function.

3.2. Semantic attention model for text sentiment analysis

Similar to image regions, some words in the text are usually more important to sentiment presentation compared to other words. Recently, semantic attention mechanism has been proven to be beneficial for many natural language processing related tasks, such as machine translation [45,46], text sentiment analysis [47,48]. Different from these work, our semantic attention model for sentiment classification is also formulated in an end-to-end process, which can directly highlight the most important words.

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_i, \ldots, T_n\}$ denotes a set of n texts, where T_i is the one hot vector representation. For each text T_i, we first use embedding matrix to embed the words into a vector space $S_i = \{s_i^1, s_i^2, \ldots, s_i^j, \ldots, s_i^L\}$ as follows:

$$S_i = T_i W_e,\ S_i \in \mathbb{R}^{L\times E} \tag{8}$$

where W_e is the parameter matrix, L denotes the length of the text, and E is the dimension of the word embedding. Then for each step

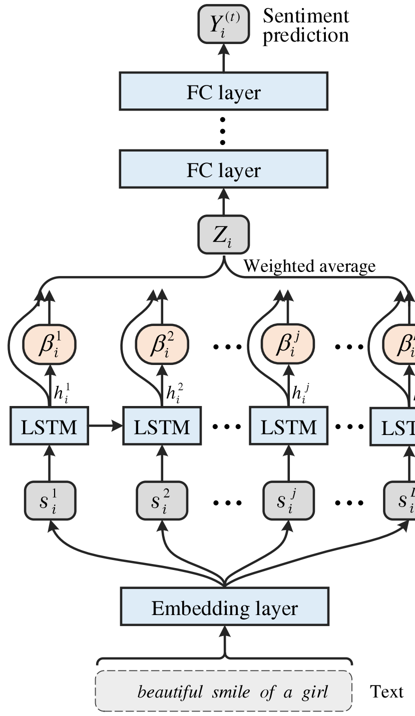


Fig. 4. The architecture of the semantic attention model.

j, we feed the word embedding vector s_i^j into an LSTM to learn the higher level features $H_i = \{h_i^1, h_i^2, \dots, h_i^j, \dots, h_i^L\}$:

$$H_i = f_s(Z_i, \theta_s),\ H_i \in \mathbb{R}^{L\times B} \tag{9}$$

where θ_s is the parameters of LSTM and B is the dimension of LSTM cells. Similar to visual attention, an attention score β_i^j is assigned to each word h_i^j based on its importance to the sentiment classification, which is calculated as follows:

$$\beta_i^j = \frac{\exp(e_i^j)}{\sum_{j=1}^{L}\exp(e_i^j)} \tag{10}$$

where

$$e_i^j = \varphi(Wh_i^j + b) \tag{11}$$

is the unnormalized attention score which measures how well the word $\{h_i^j\}$ is related to the sentiment. The weight matrix W and the bias term b are the parameters to be learned. $\varphi(\cdot)$ is a nonlinear activation function and usually a *tanh* function is adopted. β_is are used to normalize the attention over all the words $\{h_i^j\}_{1\leqslant j\leqslant L}$. The attended semantic features can be calculated as a weighted average over the word features:

$$z_i^k = \sum_{1\leqslant j\leqslant L} \beta_i^j h_i^j,\ Z_i \in \mathbb{R}^B \tag{12}$$

We denote the whole process to generate the attended features of text as follows:

$$Z_i = f_a(H_i; \theta_a^{(t)}),\ Z_i \in \mathbb{R}^B \tag{13}$$

where $\theta_a^{(t)}$ is the weight parameters which consist of W and b in Eq. (11). Similarly, the attended semantic feature mapping Z_i is more effective to represent the emotion related text features. Next, the multiple fully-connected (FC) layers are also built as the sentiment classifier:

$$Y_i^{(t)} = f_l(Z_i; \theta_l^{(t)}),\ Y_i \in \mathbb{R}^C \tag{14}$$

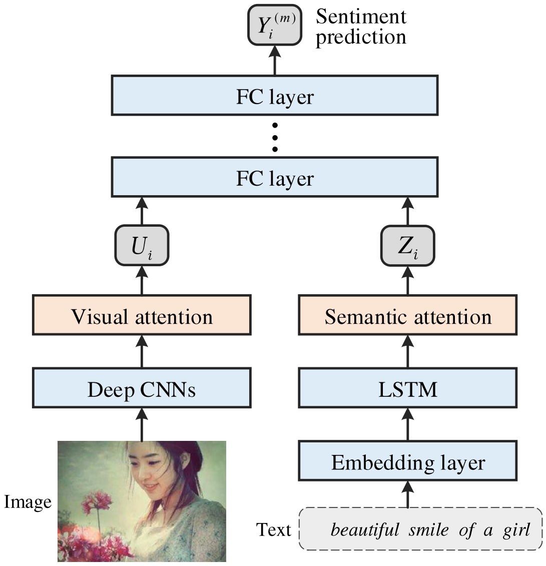


Fig. 5. The architecture of the intermediate fusion-based multimodal sentiment analysis model.

where $\theta_l^{(t)}$ is the parameters of the FC layers. The cross-entropy loss is used to learn the classifier:

$$\begin{aligned}\mathcal{L}^{(t)}(T; W_e, \theta_s, \theta_a^{(t)}, \theta_l^{(t)}) &= \sum_{i=1}^{n} -\log(Y_i^{(t)}) \\ &= \sum_{i=1}^{n} -\log(f_l(f_a(f_s(T_iW_e; \theta_s); \theta_a^{(t)}); \theta_l^{(t)}))\end{aligned} \tag{15}$$

The whole architecture of the semantic attention model is illustrated in Fig. 4. The parameters $W_e, \theta_s, \theta_a^{(t)}$ and $\theta_l^{(t)}$ are automatically learned by minimizing the cross-entropy loss over the training set.

3.3. Multimodal attention model via intermediate fusion

As discussed above, two independent models have been proposed to learn sentiment classification for image and text respectively. Then, a multimodal deep model is proposed to combine the two separate attention models for multimodal sentiment analysis. A straightforward method to combine the two models is using a late fusion scheme over the results derived from the two models. However, this kind of fusion is not effective to exploit the internal correlation between images and texts. Inspired by the recent studies on intermediate fusion [10–12], we propose to employ the intermediate fusion scheme to fuse the visual and semantic attention models for multimodal sentiment analysis.

In order to integrate the two separate models, we extract the attended visual features U_i and textual features Z_i from Eqs. (5) and (13) respectively. Then U_i and Z_i are fed into multiple fully-connected (FC) layers to encode the internal correlation between image and text features for sentiment classification:

$$Y_i^{(m)} = f_l(U_i, Z_i; \theta_l^{(m)}),\ Y_i^{(m)} \in \mathbb{R}^C \tag{16}$$

where θ_l denotes the parameters of the FC layers and C denotes the number of classes. The whole loss can be written as:

$$\mathcal{L}^{(m)}(V, T; \theta_c, \theta_a^{(v)}, W_e, \theta_s, \theta_a^{(t)}, \theta_l^{(m)}) = \sum_{i=1}^{n} -\log(Y_i^{(m)}) \tag{17}$$

where θ_c and $\theta_a^{(v)}$ are the parameters learned to obtain the attended visual features, and W_e, θ_s, and $\theta_a^{(t)}$ are the parameters learned to obtain the attended semantic features. The whole architecture of the multimodal attention model is illustrated in Fig. 5.

3.4. Improved sentiment analysis via late fusion

Since some multimodal data is incomplete, we combine the three aforementioned models, i.e., visual attention model, semantic attention model, and multimodal attention model, to derive the final result. Late fusion is reputed as a simple and effective way to fuse the features of different nature for machine-learning problems, and it is employed upon the three models. Then, DMAF is proposed to classify the multimodal sentiment via a late fusion scheme as follows:

$$Y_i = \begin{cases} \frac{1}{1+\gamma+\delta}(Y_i^{(m)} + \gamma Y_i^{(v)} + \delta Y_i^{(t)}) & input : (V_i, T_i) \\ Y_i^{(v)} & input : (V_i) \\ Y_i^{(t)} & input : (T_i) \end{cases} \tag{18}$$

where γ and δ are the hyper-parameters used to weight the importance of different classifiers. When the multimodal input is incomplete, e.g., only image modality (V_i) is possible, the final sentiment score (Y_i) is calculated directly as the sentiment of the visual attention model ($Y_i^{(v)}$).

4. Experiments

In this section, a set of experiments are conducted to analyze the effectiveness of DMAF. The proposed model is tested on several real-world datasets and the performance is qualitatively presented.

4.1. Datasets

The evaluation is conducted on four large-scale social image datasets collected from Getty, Twitter, and Flickr. Details of the datasets are described below.

- **Getty Image**. The first dataset used in this work is Getty Images. Similar to [11], we use 37 positive keywords and 64 negative keywords in the Balanced Affective Word List Project[1] to retrieve images in Getty Image. At most 2000 images are collected for each keyword. The returned images are labeled using the sentiment labels of these keywords. In this way, we collect a large weakly labeled dataset containing more than 500,000 image and text pairs.
- **Twitter**. We use Twitter API[2] to collect a large number of Tweets, i.e., about 10 million Tweets. Then, only the Tweets containing both image and English text are kept, and 181,643 image tweets are preserved. Similar to [11], VADER [49] is adopted to weakly label these tweets. The top-ranked positive and negative tweets are selected according to the VADER score. We manually filter out duplicated, low-quality, porn, and all-text images. Finally, a total of nearly 20,000 weakly labeled image tweets are retained.
- **Flickr**. The third dataset is collected from Flick. We use the 1200 adjective and noun pairs (ANPs) in [37] to collect images from Flickr. Each ANP is used to retrieve the top 500 images and the corresponding descriptions are also collected. The images with too long (more than 100 words) or too short (less than 5 words) descriptions are removed. After that, a dataset named **Flickr-w** is obtained. It contains 312,796 weakly labeled image–description pairs. Besides, in order to obtain more accurate labels, a portion of this weakly labeled dataset are further manually labeled to form a new dataset. This is, we randomly select 12,000 positive image–text pairs and 12,000 negative image–text pairs to be sent to 5 annotators for sentiment annotation. After that, we keep those that have at least 4 agreements on the sentiment label. In this way, a manually labeled dataset called **Flickr-m** containing 19,356 image and text pairs is collected.

Table 1
Statistics of the datasets.

Dataset	Labeling	Positive	Negative	Total
Getty Image	Weakly	271,786	240,925	512,711
Twitter	Weakly	10,734	8,960	19,694
Flickr-w	Weakly	162,768	150,028	312,796
Flickr-m	Manually	9,924	9,432	19,356

The final statistics of these datasets are shown in Table 1. We randomly separate all the image–text pairs in each dataset into a training set, a validation set and a test set with the proportion of 70%, 10%, and 20% respectively.

4.2. Model settings

The images with size 224 × 224 and channel RGB are used as the visual input, and deep CNNs are used to extract the visual features. Our CNN layers employ the VGG19 networks [50] pre-trained on ImageNet 2012 classification challenge dataset [51] with tensorflow[3] framework. Similar to [17], the output of the convolutional layer "conv5_4" is used as the region features, of which the dimension is 196 × 512. Thus each image has a total of 196 regions for the attention model. To improve the image features, the parameters of CNNs are then fine-tuned during the training of the proposed model.

As for the textual content, the pre-trained word features built on GloVe [52] are applied for the initialization of the parameters of the embedding layer, and each word is represented by a 300-dimensional vector. We adopt the zero padding to set a max length of the text descriptions with 30, and the sequences which are longer than 30 are truncated.

The LSTM for semantic feature learning is set to have 256 hidden neurons in each cell. The network structure of FC layers in the three attention models are all set as $1024(tanh)-512(tanh)-256(tanh)-1(sigmoid)$. In the training procedure, Stochastic gradient descent (SGD) is used to solve the object function, with learning rate 0.01, momentum 0.9 and nesterov True. Dropout technique [53] is employed for FC layers with probability 0.5 to reduce overfitting during training. As for the hyper-parameters in Eq. (18), we tune the parameters γ and δ using grid search (in both cases from 0 to 1, 0.1 times per step). The best results of DMAF are reported for the optimal values of γ and δ per dataset. All of the implementations of our model are trained on 2× NVIDIA Geforce GTX 1080.

4.3. Compared methods and baselines

We evaluate the performance of our model DMRLM by comparing it with the state-of-the-art approaches of sentiment analysis. The compared methods are introduced below:

- **Single Visual Model**: Logistic regression model on deep visual features from pre-trained VGG19 [50] model.

[1] http://www.sci.sdsu.edu/CAL/wordlist/origwordlist.html.

[2] https://developer.twitter.com/.

[3] https://github.com/tensorflow/tensorflow.

Table 2
Results of compared methods and our approaches on the Getty Image testing dataset.

Model	Precision	Recall	F1	Accuracy
Single Visual Model	0.739	0.639	0.686	0.709
Single Textual Model	0.768	0.730	0.749	0.746
Early Fusion	0.782	0.776	0.779	0.788
Late Fusion	0.786	0.780	0.783	0.791
CCR	0.829	0.766	0.796	0.792
T-LSTM-E	0.839	0.804	0.821	0.837
TFN	0.840	0.820	0.830	0.841
VAM	0.786	0.764	0.775	0.774
SAM	0.826	0.783	0.804	0.804
MAM	0.879	0.847	0.863	0.858
DMAF	**0.882**	**0.851**	**0.866**	**0.869**

Table 3
Results of compared methods and our approaches on the Twitter testing dataset.

Model	Precision	Recall	F1	Accuracy
Single Visual Model	0.612	0.590	0.601	0.602
Single Textual Model	0.702	0.674	0.687	0.683
Early Fusion	0.715	0.721	0.718	0.705
Late Fusion	0.653	0.639	0.646	0.656
CCR	0.737	0.711	0.724	0.724
T-LSTM-E	0.763	0.738	0.750	0.749
TFN	0.756	0.725	0.740	0.742
VAM	0.682	0.657	0.669	0.662
SAM	0.743	0.714	0.728	0.734
MAM	0.750	**0.771**	0.760	0.756
DMAF	**0.778**	0.760	**0.769**	**0.763**

- **Single Textual Model**: Logistic regression model on text feature vectors [54].
- **Early Fusion**: Logistic regression on the concatenation of visual [50] and textual features [54].
- **Late Fusion**: Average of logistic regression sentiment score on Single Visual Model and Single Textual Model.
- **CCR** [11]: A cross-modality consistent regression model, which uses progressive CNN to extract image feature and title information to represent the text information, for joint textual-visual sentiment analysis.
- **CCR+V**: A mixed method by introducing the visual attention of [41] into CCR [11]. Note that CCR+V can be only implemented on the datasets Flickr-w and Flickr-m, which provide ANPs for attribute detection in [41].
- **T-LSTM-E** [12]: A image–text joint sentiment analysis model which integrates tree-structured LSTM (T-LSTM) with visual attention mechanism to capture the correlation between image and text.
- **TFN** [16]: A deep multimodal sentiment analysis method modeling intra-modality and inter-modality dynamics together into a joint framework.

We also implement some simplified versions of DMAF to evaluate the effectiveness of different components:

- **VAM**: Only the visual attention model built for image sentiment analysis.
- **SAM**: Only the semantic attention model built for text sentiment analysis.
- **MAM**: The multimodal attention model combining the intermediate fusion for image–text sentiment analysis.

4.4. *Experimental results*

4.4.1. *Results on getty image*

Experimental Results in Table 2 show that the proposed DMAF outperforms state-of-the-art methods on the dataset Getty Image. Since Single Visual Model and Single Textual Model classify sentiment with only unimodal data, these two approaches show relatively low performance. In comparison, Early Fusion and Late Fusion improves the performance largely, which confirms combining visual and textual contents is effective for sentiment classification. One can see that T-LSTM-E and TFN are state-of-the-art methods, but the propose DMAF still makes a considerable improvement. It validates the effectiveness of our method on multimodal sentiment analysis.

On the other side, it can be seen that the proposed visual attention model (VAM) and semantic attention model (SAM) outperform Single Visual Model and Single Textual Model respectively on all metrics. This is because VAM can automatically draw attention on the affectional regions, and SAM can automatically focus on the emotional words. With deep intermediate fusion, our multimodal attention model MAM significantly improves the performance compared to VAM and SAM. It confirms that MAM is effective to exploit the complementary information from different modalities for sentiment prediction. DMAF further improves the performance with a late fusion to combine the confidence scores of the three separate attention models.

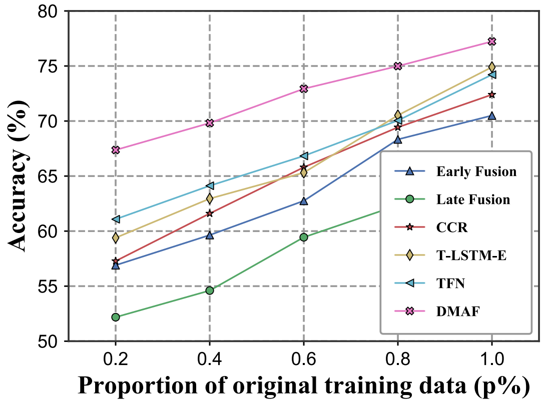


Fig. 6. Results of different multimodal sentiment classification methods training on varied proportion of original training data.

4.4.2. *Results on Twitter*

We also test the proposed model on the dataset of Twitter. Since images from Twitter are much more diverse and Tweets are also much more informal, the performance on Twitter is not as good as it on Getty Image. Table 3 shows the results of different methods on Twitter. It can be concluded that the proposed method DMAF outperforms the baseline methods and the simplified versions. It is interesting to find that the textual features perform much better than the visual features. This may be due to the fact that the weak labels are constructed by the text-based method VADER. However, the proposed multimodal attention model MAM still obtains slight improvement compared to the state-of-the-art baseline T-LSTM-E. The reason is that MAM is effective to combine the multiple modality contents for joint image–text sentiment analysis.

To fine-grained evaluate our method, DMAF is further compared with other multimodal baselines on different size of the training set. Specifically, We extracted different ratios (from 20% to 100%) of original training data as a new training dataset. Fig. 6 shows the results of accuracy of different methods trained on the new training sets. one can see that our approach DMAF consistently outperforms the compared methods. When the size of training data is smaller, the margin between DMAF and other methods becomes larger. It indicates that our method is more effective to handle the scene where there is a lack of sufficient training data.

Table 4
Results of compared methods and our approaches on the weakly labeled Flickr-w testing dataset.

Model	Precision	Recall	F1	Accuracy
Single Visual Model	0.720	0.760	0.739	0.732
Single Textual Model	0.685	0.753	0.718	0.722
Early Fusion	0.715	0.767	0.740	0.753
Late Fusion	0.735	0.779	0.756	0.764
CCR	0.781	0.798	0.790	0.796
CCR+V	0.789	0.806	0.797	0.801
T-LSTM-E	0.804	0.809	0.806	0.797
TFN	0.808	0.812	0.810	0.803
VAM	0.784	0.780	0.782	0.784
SAM	0.799	0.776	0.787	0.781
MAM	0.835	0.829	0.832	0.843
DMAF	**0.855**	**0.845**	**0.850**	**0.859**

Table 5
Results of compared methods and our approaches on the manually labeled Flickr-m testing dataset.

Model	Precision	Recall	F1	Accuracy
Single Visual Model	0.746	0.808	0.776	0.761
Single Textual Model	0.733	0.799	0.765	0.753
Early Fusion	0.785	0.801	0.793	0.801
Late Fusion	0.791	0.806	0.798	0.798
CCR	0.810	0.833	0.821	0.817
CCR+V	0.821	0.835	0.828	0.822
T-LSTM-E	0.851	0.837	0.844	0.832
TFN	0.850	0.846	0.848	0.841
VAM	0.809	0.803	0.806	0.803
SAM	0.801	0.819	0.810	0.808
MAM	0.876	0.860	0.868	0.866
DMAF	**0.882**	**0.870**	**0.876**	**0.880**

4.4.3. Results on Flickr

On Flickr, we first report the results of different methods on weakly labeled Flickr-w in Table 4. One can see that the proposed DMAF has shown considerable improvements over compared methods. Especially, DMAF improves the performance by an increase of 6% accuracy against T-LSTM-E. This proves that our method is effective to learn discriminative visual and semantic features for sentiment prediction. Meanwhile, the weighted late fusion is also helpful to make a better sentiment classification decision than separate attention models.

In addition, we also report the experimental results on the manually labeled Flickr-m in Table 5. Considering that Flickr-w has much more training data, we pre-train different methods on Flickr-w and then fine-tune them on Flickr-m. From Table 5, it is worth noting that our two uni-modal models VAM and SAM show slightly better performance than the methods based on early fusion and late Fusion. This proves that the visual and semantic attention models with end-to-end process allow the salient features to dynamically come to the forefront for more effective sentiment classification. Both MAM and DMAF have shown better performance than T-LSTM-E and TFN, which indicates that the combination of two attention models could lead to more effective performance.

4.5. Qualitative analysis

In this section, we present qualitative analysis on the impact of the proposed DMAF approach by comparing with its variant methods VAM, SAM, and MAM. Fig. 7 shows several examples of image–description pairs on Flickr-m dataset and the sentiment scores predicted by different methods. The scores range from 0 (negative) to 1 (positive).

We can see from Fig. 7 that, VAM and SAM are manageable to handle the scene where the sentiment is obvious in the corresponding modalities. However, these two unimodal methods is not effective for the images or texts with no obvious sentiment inclination (e.g., the description in the second example). In comparison, MAM is able to capture the internal correlation between visual and semantic contents and thus improves the performance. With a late fusion to combine the confidence scores, DMAF gives more reasonable sentiment scores. For the last sample, DMAF gives a wrong sentiment prediction, we speculate that this is because the fog in the image misguides our model for a wrong classification.

4.6. Visualization of learned attention

One advantage of including the attention mechanism is the ability to visualize what the model "sees". To better understand the interpretability of meaningful attention drawn for image regions and semantic words, we present several examples of visualization of the attention weights learned by our visual attention model (VAM), semantic attention model (SAM), and multimodal attention model (MAM) in Fig. 8.

For visual attention, the visualization approach is similar to [17]. That is, the image input to the convolutional network is resized to 224 × 224. Consequently, with 4 max-pooling layers, the dimension of the output on the top convolutional layer is 14 × 14. We simply up-sample the scores by a factor of $2^4 = 16$ and apply a Gaussian filter. Different from [17], we further draw a heat map for the up-sampled attention scores for brighter and more colorful visualization. The finally attended images are the original images masked by the heat map with transparency of 0.7. Therefore, if the attention scores of the regions are greater, the regions are colored redder. As for semantic attention, it is easier to visualize because the attention weights are related to the textual words directly. Thus, we put various background color on the words based on the attention weights. Specifically, the larger the weight of the word is, the stronger its background color is.

4.6.1. Visual attention on VAM

From the third column of Fig. 8, it can be seen that our model VAM generally draws the right attention on the image regions. For the first image, VAM pays attention on the more affective positions, such as the lake, trees and white clouds, which describe a magnificent scenery. Attention is scattered in the third image, because the whole image is colored mainly in dark, which all seem to contribute to negative sentiment. For the fourth image, attention is focused on the lovely dog and its food which infers that the image is positive.

4.6.2. Semantic attention on SAM

We present the semantic attention visualization results in the fourth column of Fig. 8. It can be found that SAM can capture important words, which could lead to more effective sentiment classification. For the first example, the words "magic", "gorgeous", and "scene" are especially focused on, which are all regarded as positive words. Similarly, the emotional words "decorated" and "colorful" in the second example and the words "sad", "broken", and "flattened" in the third example are drawn strong attention by SAM. Note that the attended words "ohh" and "busy" in the fourth example are more likely be negative words. However, combining with the fourth image, this image and description pair totally presents positive sentiment. It confirms that capturing complementary information from both the image and text modalities can make more reasonable sentiment prediction.

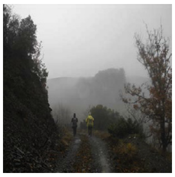
Original image-text pair		VAM	SAM	MAM	DMAF
	Good morning sunshine at Surf City USA.Huntington Beach, California	0.79	0.94	0.83	0.84
	Canon Eos 5D Mark IV, Canon EF 70-200mm f/2.8 L IS II USM	0.13	0.58	0.43	0.41
	Giant storm roars in the CapeOne of the worst storms in seven years hit the city of Cape Town on the weekend of 30 August 2008…...	0.63	0.05	0.37	0.36
	hiking at Pieria mountains Greece .	0.31	0.65	0.47	0.47

Fig. 7. Several examples of qualitative analysis results on Flickr-m.

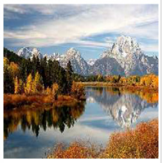
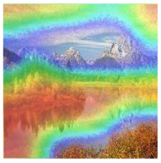
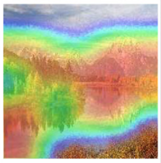
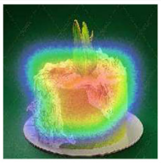
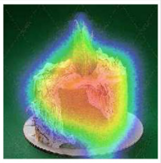
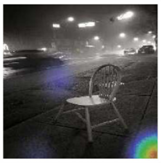
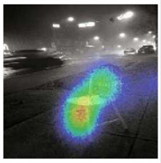
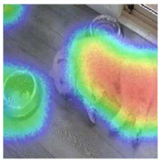
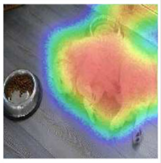
No.	Original image-text pair		Attended image (VAM)	Attended text (SAM)	Attended image and text pair (MAM)	
1		Located in Grand Teton National Park. A magic moment on our tour. We arrived here on the right day and at the right time to encounter this gorgeous scene. It's at Oxbow Bend on the Snake River just downstream from Jackson Lake.		Located in Grand Teton National Park. A magic moment on our tour. We arrived here on the right day and at the right time to encounter this gorgeous scene. It's at Oxbow Bend on the Snake River just downstream from Jackson Lake.		Located in Grand Teton National Park. A magic moment on our tour. We arrived here on the right day and at the right time to encounter this gorgeous scene. It's at Oxbow Bend on the Snake River just downstream from Jackson Lake.
2		Green birthday cake decorated with burned meringue and colored pear slices on green background. Colorful cake with Italian meringue		Green birthday cake decorated with burned meringue and colored pear slices on green background. Colorful cake with Italian meringue		Green birthday cake decorated with burned meringue and colored pear slices on green background. Colorful cake with Italian meringue
3		A sad, broken chair sitting curbside on Queen Street on a Saturday night. By Sunday morning, it was completely flattened		A sad, broken chair sitting curbside on Queen Street on a Saturday night. By Sunday morning, it was completely flattened		A sad, broken chair sitting curbside on Queen Street on a Saturday night. By Sunday morning, it was completely flattened
4		ohh it's been a busy day		ohh it's been a busy day		ohh it's been a busy day

Fig. 8. Four examples of visualization of the learnt visual and semantic attentions learned by VAM, SAM, and MAM.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.6.3. *Visual and semantic attention on MAM*

The last column of Fig. 8 shows the attended images and text descriptions of MAM. It can be seen that the visual attention weights are generally similar to them in VAM, and the semantic attention weights are similar to them in SAM. However, since MAM combines multiple modalities to capture complementary information in different data modalities, there exists a certain difference in the attended regions (and words) between MAM and VAM (and SAM). Taking the second image–text pair as an example, VAM puts attention on the green birthday cake, while MAM focuses on the

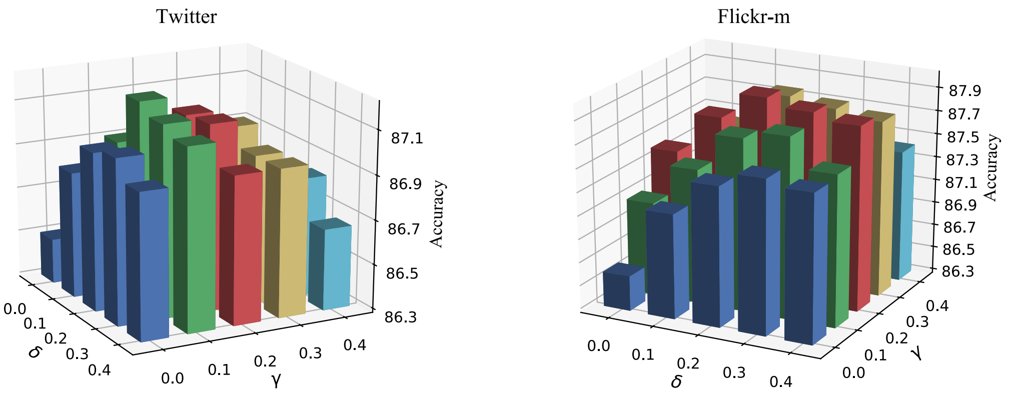


Fig. 9. Multimodal sentiment analysis performance of DMAF on dataset Twitter and Flickr-m with different values of model parameters γ and δ.

regions of both the colored pear slices and the green birthday cake. The reason may be that MAM exploits the correlation between image and text, which results in that more visual attention is drawn on the regions of pear slices learned from "colored pear slices" in the text.

4.7. Parameter sensitivity

In previous experiments, we empirically set the parameters γ and δ in the sentiment prediction function (i.e., Eq. (18)) for late fusion. As γ and δ control the contributions of visual attention model (VAM) and semantic attention model (SAM) respectively, we use Twitter and Flickr-m to test the sensitivity of parameters.

When $\gamma = 0$ and $\delta = 0$, DMAF degenerates into multimodal attention model (MAM). The evaluation is conducted by changing one parameter (e.g., δ) while fixing the other (e.g., γ). Fig. 9 shows the performance of DMAF on different values of γ and δ. It can be observed that DMAF obtains the best performance when $\gamma = 0.1$ and $\delta = 0.2$ on Twitter, and $\gamma = 0.2$ and $\delta = 0.2$ on Flickr-m. It can be inferred that SAM has more contribution to the overall performance compared to VAM on Twitter, which is in accordance with the accuracy scores of SAM and VAM in Table 3. While on Flick-m, SAM and VAM have a similar contribution to the final performance of DMAF.

Note that since the three models of VAM, SAM, and MAM have been trained before late fusion, the selection of γ and δ is actually very simple to be implemented. If there is one available validation or testing dataset, the best parameter values can be found with random search or grid search where DMAF has best test performance. This process is very fast to implement since it has only feed-forward calculation with several iterations. If it is not capable to find a validation or testing dataset, we can just set $\gamma = 0.2$ and $\delta = 0.2$. Because DMAF can nearly get to the optimal results around 0.2 for γ and δ on most datasets.

5. Conclusions

In this paper, we exploit the visual and semantic attention mechanism with a mixed fusion framework for image–text sentiment analysis. A Deep Multimodal Attentive Fusion (DMAF) method is proposed, in which two separate unimodal attention models are proposed to learn effective emotion classifiers for visual and textual modality respectively. Then an intermediate fusion-based multimodal attention model is proposed to exploit the internal correlations between different modalities. Finally, a late fusion scheme is applied to the outputs of the three models to derive the final decision. The experiment results indicate that our method outperforms state-of-the-art baselines on four real-world datasets.

In the future, we will design a more reasonable deep model to make the learned multimodal feature more effective for image–text sentiment analysis. Especially, The fine-granularity relation between image and text pairs is worth excavated to fuse the two types of modalities more comprehensively and non-redundantly. We also want to explore how to generalize our method to other multimodal data such as music and videos.

Acknowledgment

This work was supported in part by Beijing Natural Science Foundation of China (No. 4182037), in part by the National Natural Science Foundation of China (No. U1636211), and in part by the Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2017ZX-19).

References

[1] H. Cui, V.O. Mittal, M. Datar, Comparative experiments on sentiment classification for online product reviews, in: Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16–20, 2006, Boston, Massachusetts, USA, AAAI Press, 2006, pp. 1265–1270, URL http://www.aaai.org/Library/AAAI/2006/aaai06-198.php.

[2] W. Wei, J.A. Gulla, Sentiment learning on product reviews via sentiment ontology tree, in: J. Hajic, S. Carberry, S. Clark (Eds.), ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11–16, 2010, Uppsala, Sweden, The Association for Computer Linguistics, 2010, pp. 404–413, URL http://www.aclweb.org/anthology/P10-1042.

[3] D. Tang, B. Qin, T. Liu, Learning semantic representations of users and products for document level sentiment classification, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26–31, 2015, Beijing, China, in: Long Papers, vol. 1, The Association for Computer Linguistics, 2015, pp. 1014–1023, URL http://aclweb.org/anthology/P/P15/P15-1098.pdf.

[4] R. Feldman, B. Rosenfeld, R. Bar-Haim, M. Fresko, The stock sonar - sentiment analysis of stocks based on a hybrid approach, in: D.G. Shapiro, M.P.J. Fromherz (Eds.), Proceedings of the Twenty-Third Conference on Innovative Applications of Artificial Intelligence, August 9–11, 2011, San Francisco, California, USA, AAAI, 2011, URL http://www.aaai.org/ocs/index.php/IAAI/IAAI-11/paper/view/3506.

[5] X. Li, H. Xie, L. Chen, J. Wang, X. Deng, News impact on stock price return via sentiment analysis, Knowl.-Based Syst. 69 (2014) 14–23, http://dx.doi.org/10.1016/j.knosys.2014.04.022.

[6] T.H. Nguyen, K. Shirai, Topic modeling based sentiment analysis on social media for stock market prediction, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26–31, 2015, Beijing, China, in: Long Papers, vol. 1, The Association for Computer Linguistics, 2015, pp. 1354–1364, URL http://aclweb.org/anthology/P/P15/P15-1131.pdf.

[7] L. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: H. Bourlard, T.S. Huang, E. Vidal, D. Gatica-Perez, L. Morency, N. Sebe (Eds.), Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14–18, 2011, ACM, 2011, pp. 169–176, http://dx.doi.org/10.1145/2070481.2070509.

[8] V. Pérez-Rosas, R. Mihalcea, L. Morency, Utterance-Level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4–9 August 2013, Sofia, Bulgaria, in: Long Papers, vol. 1, The Association for Computer Linguistics, 2013, pp. 973–982, URL http://aclweb.org/anthology/P/P13/P13-1096.pdf.

[9] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: F. Bonchi, J. Domingo-Ferrer, R.A. Baeza-Yates, Z. Zhou, X. Wu (Eds.), IEEE 16th International Conference on Data Mining, ICDM 2016, December 12–15, 2016, Barcelona, Spain, IEEE, 2016, pp. 439–448, http://dx.doi.org/10.1109/ICDM.2016.0055.

[10] M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, L. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: E. Lank, A. Vinciarelli, E.E. Hoggan, S. Subramanian, S.A. Brewster (Eds.), Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13–17, 2017, ACM, 2017, pp. 163–171, http://dx.doi.org/10.1145/3136755.3136801.

[11] Q. You, J. Luo, H. Jin, J. Yang, Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia, in: P.N. Bennett, V. Josifovski, J. Neville, F. Radlinski (Eds.), Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22–25, 2016, ACM, 2016, pp. 13–22, http://dx.doi.org/10.1145/2835776.2835779.

[12] Q. You, L. Cao, H. Jin, J. Luo, Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks, in: A. Hanjalic, C. Snoek, M. Worring, D.C.A. Bulterman, B. Huet, A. Kelliher, Y. Kompatsiaris, J. Li (Eds.), Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15–19, 2016, ACM, 2016, pp. 1008–1017, http://dx.doi.org/10.1145/2964284.2964288.

[13] M. Wöllmer, F. Weninger, T. Knaup, B.W. Schuller, C. Sun, K. Sagae, L. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53, http://dx.doi.org/10.1109/MIS.2013.34.

[14] M. Wang, D. Cao, L. Li, S. Li, R. Ji, Microblog sentiment analysis based on cross-media bag-of-words model, in: H. Wang, L. Davis, W. Zhu, S. Kopf, Y. Qu, J. Yu, J. Sang, T. Mei (Eds.), International Conference on Internet Multimedia Computing and Service, ICIMCS '14, Xiamen, China, July 10–12, 2014, ACM, 2014, p. 76, http://dx.doi.org/10.1145/2632856.2632912.

[15] D. Cao, R. Ji, D. Lin, S. Li, A cross-media public sentiment analysis system for microblog, Multimedia Syst. 22 (4) (2016) 479–486, http://dx.doi.org/10.1007/s00530-014-0407-8.

[16] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, Association for Computational Linguistics, 2017, pp. 1103–1114, URL https://aclanthology.info/papers/D17-1115/d17-1115.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F.R. Bach, D.M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, in: JMLR Workshop and Conference Proceedings, 37, JMLR.org, 2015, pp. 2048–2057, URL http://jmlr.org/proceedings/papers/v37/xuc15.html.

[18] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing When to Look: Adaptive attention via a visual sentinel for image captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 3242–3250, http://dx.doi.org/10.1109/CVPR.2017.345.

[19] J. Song, Q. Yu, Y. Song, T. Xiang, T.M. Hospedales, Deep spatial-semantic attention for fine-grained sketch-based image retrieval, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 5552–5561, http://dx.doi.org/10.1109/ICCV.2017.592.

[20] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting Elections with Twitter: What 140 characters reveal about political sentiment, in: W.W. Cohen, S. Gosling (Eds.), Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23–26, 2010, The AAAI Press, 2010, URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441.

[21] J.E. Chung, E. Mustafaraj, Can collective sentiment expressed on twitter predict political elections? in: W. Burgard, D. Roth (Eds.), Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7–11, 2011, AAAI Press, 2011, URL http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3549.

[22] H. Wang, D. Can, A. Kazemzadeh, F. Bar, S. Narayanan, A system for real-time twitter sentiment analysis of 2012 U.S. Presidential Election Cycle, in: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea, The Association for Computer Linguistics, 2012, pp. 115–120, URL http://www.aclweb.org/anthology/P12-3020.

[23] T.T. Thet, J. Na, C.S.G. Khoo, Aspect-based sentiment analysis of movie reviews on discussion boards, J. Inf. Sci. 36 (6) (2010) 823–848, http://dx.doi.org/10.1177/0165551510388123.

[24] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17–23 May 2010, Valletta, Malta, European Language Resources Association, 2010, URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/769.html.

[25] M. Taboada, J. Brooke, M. Tofiloski, K.D. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307, http://dx.doi.org/10.1162/COLI_a_00049.

[26] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domain-oriented sentiment analysis, in: D. Jurafsky, É. Gaussier (Eds.), EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22–23 July 2006, Sydney, Australia, ACL, 2006, pp. 355–363, URL http://www.aclweb.org/anthology/W06-1642.

[27] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25–26 July 2004, Barcelona, Spain, ACL, 2004, pp. 412–418, URL http://www.aclweb.org/anthology/W04-3253.

[28] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19–24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 142–150, URL http://www.aclweb.org/anthology/P11-1015.

[29] R. Remus, Asvuniofleipzig: sentiment analysis in twitter using data-driven machine learning techniques, in: M.T. Diab, T. Baldwin, M. Baroni (Eds.), Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14–15, 2013, The Association for Computer Linguistics, 2013, pp. 450–454, URL http://aclweb.org/anthology/S/S13/S13-2074.pdf.

[30] X. Wang, F. Wei, X. Liu, M. Zhou, M. Zhang, Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach, in: C. Macdonald, I. Ounis, I. Ruthven (Eds.), Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24–28, 2011, ACM, 2011, pp. 1031–1040, http://dx.doi.org/10.1145/2063576.2063726.

[31] Y. Rao, J. Lei, L. Wenyin, Q. Li, M. Chen, Building emotional dictionary for sentiment analysis of online news, World Wide Web 17 (4) (2014) 723–742, http://dx.doi.org/10.1007/s11280-013-0221-9.

[32] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, in: L. Getoor, T. Scheffer (Eds.), Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28–July 2, 2011, Omnipress, 2011, pp. 513–520.

[33] D. Tang, F. Wei, B. Qin, T. Liu, M. Zhou, Coooolll: A deep learning system for twitter sentiment classification, in: P. Nakov, T. Zesch (Eds.), Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23–24, 2014, The Association for Computer Linguistics, 2014, pp. 208–212, URL http://aclweb.org/anthology/S/S14/S14-2033.pdf.

[34] A. Severyn, A. Moschitti, Twitter sentiment analysis with deep convolutional neural networks, in: R.A. Baeza-Yates, M. Lalmas, A. Moffat, B.A. Ribeiro-Neto (Eds.), Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9–13, 2015, ACM, 2015, pp. 959–962, http://dx.doi.org/10.1145/2766462.2767830.

[35] S. Siersdorfer, E. Minack, F. Deng, J.S. Hare, Analyzing and predicting sentiment of images on the social web, in: A.D. Bimbo, S. Chang, A.W.M. Smeulders (Eds.), Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25–29, 2010, ACM, 2010, pp. 715–718, http://dx.doi.org/10.1145/1873951.1874060.

[36] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, J. Tang, How do your friends on social media disclose your emotions? in: C.E. Brodley, P. Stone (Eds.), Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, Québec City, Québec, Canada, AAAI Press, 2014, pp. 306–312, URL http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8393.

[37] D. Borth, R. Ji, T. Chen, T.M. Breuel, S. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D.A. Shamma, M. Worring, R. Zimmermann (Eds.), ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21–25, 2013, ACM, 2013, pp. 223–232, http://dx.doi.org/10.1145/2502081.2502282.

[38] J. Yuan, S. Mcdonough, Q. You, J. Luo, Sentribute: image sentiment analysis from a mid-level perspective, in: E. Cambria, B. Liu, Y. Zhang, Y. Xia (Eds.), Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013, ACM, 2013, pp. 10:1–10:8, http://dx.doi.org/10.1145/2502069.2502079.

[39] C. Xu, S. Cetintas, K. Lee, L. Li, Visual sentiment prediction with deep convolutional neural networks, CoRR abs/1411.5731 (2014) arXiv:1411.5731.

[40] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, in: B. Bonet, S. Koenig (Eds.), Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA, AAAI Press, 2015, pp. 381–388, URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9556.

[41] Q. You, H. Jin, J. Luo, Visual sentiment analysis by attending on local image regions, in: S.P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 231–237, URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14964.

[42] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L. Morency, Context-Dependent sentiment analysis in user-generated videos, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, in: Long Papers, vol. 1, Association for Computational Linguistics, 2017, pp. 873–883, http://dx.doi.org/10.18653/v1/P17-1081.

[43] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L. Morency, Memory fusion network for multi-view sequential learning, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, 2018, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17341.

[44] F. Huang, X. Zhang, C. Li, Z. Li, Y. He, Z. Zhao, Multimodal network embedding via attention based multi-view variational autoencoder, in: K. Aizawa, M.S. Lew, S. Satoh (Eds.), Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR 2018, Yokohama, Japan, June 11–14, 2018, ACM, 2018, pp. 108–116, http://dx.doi.org/10.1145/3206025.3206035.

[45] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR abs/1409.0473 (2014) arXiv:1409.0473.

[46] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, Y. Marton (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, The Association for Computational Linguistics, 2015, pp. 1412–1421, URL http://aclweb.org/anthology/D/D15/D15-1166.pdf.

[47] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016, The Association for Computational Linguistics, 2016, pp. 606–615, URL http://aclweb.org/anthology/D/D16/D16-1058.pdf.

[48] H. Chen, M. Sun, C. Tu, Y. Lin, Z. Liu, Neural sentiment classification with user and product attention, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016, The Association for Computational Linguistics, 2016, pp. 1650–1659, URL http://aclweb.org/anthology/D/D16/D16-1171.pdf.

[49] C.J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: E. Adar, P. Resnick, M.D. Choudhury, B. Hogan, A.H. Oh (Eds.), Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014, The AAAI Press, 2014, URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

[50] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014).

[51] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, ImageNet: A large-scale hierarchical image database, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 2009, pp. 248–255.

[52] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[53] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958, URL http://dl.acm.org/citation.cfm?id=2670313.

[54] Q.V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014, in: JMLR Workshop and Conference Proceedings, vol. 32, JMLR.org, 2014, pp. 1188–1196, URL http://jmlr.org/proceedings/papers/v32/le14.html.