Sentiment and Emotion Analysis for Social Multimedia: Methodologies and Applications

Quanzeng You
Department of Computer Science, University of Rochester
Rochester NY, USA 14623
qyou@cs.rochester.edu

ABSTRACT

Online social networks have attracted the attention from both the academia and real world. In particular, the rich multimedia information accumulated in recent years provides an easy and convenient way for more active communication between people. This offers an opportunity to research people's behaviors and activities based on those multimedia content. One emerging area is driven by the fact that these massive multimedia data contain people's daily sentiments and opinions. However, existing sentiment analysis typically focuses on textual information regardless of the visual content, which may be as informative in expressing people's sentiments and opinions. In this research, we attempt to analyze the online sentiment changes of social media users using both the textual and visual content. Nowadays, social media networks such as Twitter have become major platforms of information exchange and communication between users, with tweets as the common information carrier. As an old saying has it, an image is worth a thousand words. The image tweet is a great example of multimodal sentiment. In this research, we focus on sentiment analysis based on visual and multimedia information analysis. We will review the state-of-the-art in this topic. Several of our projects related to this research area will also be discussed. Experimental results are included to demonstrate and summarize our contributions.

Keywords

Visual Sentiment Analysis; Joint Sentiment Analysis; Multimodal

1. INTRODUCTION

Online social networks are providing more and more convenient services to their users. Today, social networks have grown to be one of the most important sources for people to acquire information on all aspects of their lives. Meanwhile, every online social network user is a contributor to such large amounts of information. Online users love to share their experiences and to express their opinions on virtually all events and subjects.

Among the large amount of online user generated data, we are particularly interested in people's opinions or sentiments towards specific topics and events. There have been studies on using online users' sentiments to predict box-office revenues for movies, political elections and economic indicators. These results have suggested that online users' opinions or sentiments are closely correlated with our real-world activities. All of these results hinge on accurate estimation of people's sentiments according to their online generated content. Currently all of these studies only rely on sentiment analysis from textual content. However, multimedia content, including images and videos, has become prevalent over all online social networks. Indeed, online social network providers are competing with each other by providing easier access to their increasingly powerful and diverse services.

A picture is worth a thousand words. People with different backgrounds can easily understand the main content of an image or video. Apart from the large amount of easily available visual content, today's computational infrastructure is also much cheaper and more powerful to make the analysis of computationally intensive visual content analysis feasible. In this era of big data, it has been shown that the integration of visual content can provide us more reliable or complementary online social signals [12, 38].

On the other hand, images and text always come in pairs, which is becoming increasingly prevalent in online social networks. Intuitively, we can alleviate the discussed challenges in visual sentiment analysis by the integration of textual knowledge, which has been well studied.

To the best of our knowledge, little attention has been paid to the sentiment analysis of visual content as well as multi-modality sentiment analysis. Only a few recent works attempted to predict visual sentiment using features from images [26, 4, 3, 38] and videos [25]. Visual sentiment analysis is extremely challenging, as image sentiment involves a much higher level of abstraction and subjectivity in the human recognition process [13], on top of a wide variety of visual recognition tasks including object, scene, action and event recognition. However, Convolutional Neural Networks [19, 8, 16] have been proved to be very powerful in solving computer vision related tasks. In addition, the successes of deep learning make the understanding and jointly modeling of vision and language content feasible. In the context of deep learning, many related publications have proposed novel models that address image and text simultaneously. Starting with matching images with word-level concepts [9] and recently onto sentence-level descriptions [15, 27, 21, 22, 14], deep neural networks exhibit significant performance improvements on these tasks. These models have inspired the idea of joint feature learning [29], semantic transfer [9] and design of margin ranking loss [34].

Inspired by the recent advancement in deep learning, we are particularly interested in utilizing it to solve the challenging social multimedia sentiment analysis. To that end, we address in this thesis proposal the following challenges: 1) how to learn with large

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2971475

scale weakly labeled training data, 2) how to generalize and extend the learned model across domains, 3) we propose novel multi-modality models, which can integrate different modality features for sentiment analysis, and 4) we demonstrate a recently collected visual emotion dataset, which is publicly available for research purpose.

2. RELATED WORK

In this section, we review literature closely related to our study on social multimedia sentiment analysis.

2.1 Single modality sentiment analysis

For sentiment analysis of online user generated textual content, dictionaries based approaches [2, 10] have been widely used due to its efficiency and simplicity. Very recently, distributed representation of words started to attract research attention due to its ability in learning robust features for words [24]. Le and Mikolov [18] further proposed an approach to learn distributed representation for documents. They applied their document representations to sentiment analysis and achieve the best performance over existing competing algorithms.

There are also several recent works on visual sentiment analysis. Siersdorfer et al. [26] is a machine learning algorithm to predict the sentiment of images using pixel-level features. Motivated by the fact that sentiment involves high-level abstraction, which may be easier to explain by objects or attributes in images, both [3] and [38] propose to employ visual entities or attributes as features for visual sentiment analysis. In [3], 1200 adjective noun pairs (ANP), which may correspond to different levels of different emotions, are extracted. These ANPs are used as queries to crawl images from Flickr. Next, pixel-level features of images in each ANP are employed to train 1200 ANP detectors. The responses of these 1200 classifiers can then be considered as mid-level features for visual sentiment analysis. The work in [38] employed a similar mechanism of using scene attributes.

2.2 Joint visual and textual analysis

Computer vision and natural language processing are important application domains of machine learning. Recently, deep learning has made significant advances in tasks related to both vision and language [17]. Consequently, the task of higher-level semantic understanding, such as machine translation [1], image aesthetic analysis [20], and visual sentiment analysis [5, 35] have become tractable. A more interesting and challenging task is to bridge the semantic gap between vision and language, and thus help solve more challenging problems.

Notably, automatic image captioning is widely studied [7, 32, 14], which more intimately connects visual content and language semantics. In general, these models handle two primary tasks: 1) how to represent image and text, and 2) how to learn the model on top of visual and textual features. Indeed, Convolutional Neural Networks (CNNs) [17] become the common approaches for extracting visual features. Meanwhile, multimodal semantic mapping following a pairwise ranking loss is widely adopted for optimizing joint visual and textual models. Different approaches, including sequential [22, 14] and tree-structured models [27, 21], are selected to encode the text, among which Recurrent Neural Networks [30] and Recursive Neural Networks [28] are particularly popular. However, there are only a few publications on analyzing sentiment using both text and images, which are prevalent in user generated content. Both [33] and [6] 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 [3].

Inspired by these studies on learning joint visual and textual models. In this proposal, we are going to present two of our projects on using deep learning for joint visual-textual sentiment analysis.

3. APPROACHES

In this section, we describe our proposed models for visual sentiment analysis as well as joint visual-textual sentiment analysis. The proposed models intend to solve the discussed challenges in social multimedia sentiment analysis.

3.1 Visual sentiment analysis with progressive CNN

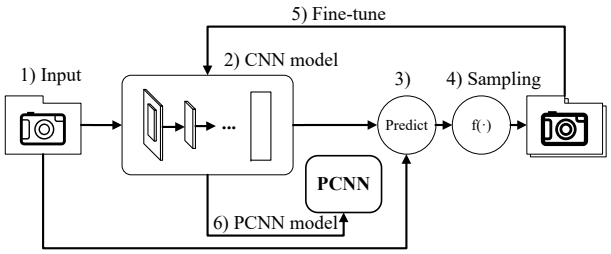

Figure 1: Progressive CNN (PCNN) for visual sentiment analysis.

To train a convolutional neural network, we need a large amount of high-quality labeled images. However, in visual sentiment analysis, this could be too challenging. It is difficult to collect such a large and high quality image database. Part of the reason can be attributed to the challenging nature of the task, which may lead inconsistent opinions for different people. Thus, we work on weakly labeled images. This can lead to the possibility that the neural network can get stuck in a bad local optimum. This may lead to poor generalizability of the trained neural network. But, the neural network is still able to correctly classify a large proportion of the training instances. Therefore, our idea is to progressively select a subset of the training instances to reduce the impact of noisy training instances. Figure 1 shows the overall flow of the proposed progressive CNN (PCNN). We first train a CNN on Flickr images. Next, we select training samples according to the prediction score of the trained model on the training data itself. Instead of training from the beginning, we further fine-tune the trained model using these newly selected, and potentially cleaner training instances. This fine-tuned model will be our final model for visual sentiment analysis.

In particular, we employ a probabilistic sampling algorithm to select the new training subset. Let $s_i = (s_{i1}, s_{i2})$ be the prediction sentiment scores for the two classes of instance i. We choose to remove the training instance i with probability p_i given by Eqn.(1).

$$p_i = \max\left(0, 2 - \exp(|s_{i1} - s_{i2}|)\right) \quad (1)$$

When the difference between the predicted sentiment scores of one training instance are large enough, this training instance will be kept in the training set. Otherwise, the smaller the difference between the predicted sentiment scores become, the larger the probability of this instance being removed from the training set.

3.2 Joint visual-textual sentiment analysis

In this section, we present two of our proposed models on joint visual-textual sentiment analysis. Cross-modality consistent re-

gression focuses on enforcing task consistency between textual and visual modalities. The second one focus on learning semantic mappings for visual and textual data.

3.2.1 Cross-modality consistent regression (CCR)

Our main idea is that the penalties between the predicted label distributions of different modality features need to be taken into consideration. To measure the penalty between any two different predicted label distributions, we employ KL divergence. We define $D(p \parallel q)$ as the sum of KL divergence between two probability distributions p and q.

The objective function is formulated in Eqn. (2). We denote by x_i^m (for $m \in \{1, \cdots, M\}$) the m-th modality features of the i-th instance and by x_i^c the concatenated features from all the M modality features of the i-th instance. $\Theta = \{\theta^c, \theta^1, \cdots, \theta^M\}$ are the parameters that needs to be learned.

$$\min_{\Theta} \sum_i (D(y_i \parallel p_{\theta^c}(x_i^c)) + \sum_{m=1}^{M} \gamma_m D(p_{\theta^c}(x_i^c) \parallel p_{\theta^m}(x_i^m))) \tag{2}$$

Let $p_\theta(x_i)$ be the prediction function for the label distribution of x_i given the parameter vector θ. We use softmax function to evaluate the probability distribution.

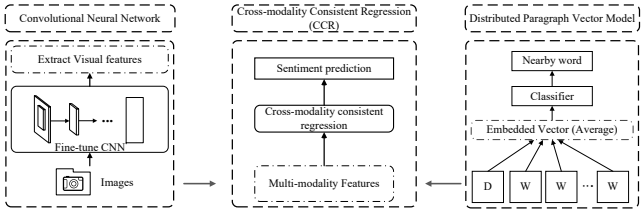

Figure 2: The framework for multi-modality sentiment analysis. *Left:* We fine-tune a CNN visual sentiment analysis model. *Right:* We train a distributed paragraph vector model to learn textual features. *Middle:* The proposed CCR is trained on the visual and textual features.

The first component of the objective function is the consistency constraint between the ground-truth label and the predicted label distribution using concatenated features. The last component considers the predicted distribution consistency between each single modality features and the concatenated features. In this way, we hope that knowledge is propagated to each other to improve the overall performance of the system.

There are two groups of parameters in our model, namely θ^c and $\{\theta^1, \theta^2, \cdots, \theta^M\}$. In our implementation, we learn those two groups of parameters iteratively. Specifically, in each iteration, the learning algorithm will try to update the two groups of parameters sequentially. Since, we built a large data set for our experiments, we employ mini-batch L-BFGS to learn the parameters[1].

3.2.2 Tree-structured LSTM for sentiment analysis

Different from CCR, which focuses on the output of the classifiers, we propose another framework trying to integrate their semantic content. In particular, we utilize tree-structured LSTM (T-LSTM), which is built on top of textual semantic parsing trees [31, 39]. Both of current studies build T-LSTM on text (single modality). We will generalize it to multimodality scenario where a pair of sentence and image (t^m, v^m) are taken into account jointly. We take the hidden state at the root node of the T-LSTM as the representation for the sentence and image pair. Next, this representation can be supplied as input to build a softmax classifier for sentiment analysis.

To learn and establish the joint representation $h[t^m, v^m]$ of a given pair of sentence and image (t^m, v^m), we focus on the leaf nodes, which directly accept textual words and visual representations as inputs. In particular, we want the *leaf* nodes of T-LSTM to jointly accept both individual words and image regions and produce its output based on both inputs. Eventually, the root node of the parsing tree will learn a joint embedding by receiving both the visual and the textual information propagated structurally from the *leaf* nodes. In such a way, we are able to integrate the visual information into the tree-structures. We use a bilinear attention model as the joint module to learn the alignments and produce the outputs of leaf nodes simultaneously.

Semantic Embedding Learning Inspired by the recent successes of deep visual-textual semantic embedding learning [14, 15, 21], we incorporate the semantic learning task to pilot the learning of attention model. Let (t^m,v^m) be a sentence and image pair, and v^n (randomly picked from the training set) be a contrasting image of t^m. We then use the previously introduced T-LSTM model with the bilinear attention mechanism to encode both (t^m,v^m) and (t^m,v^n). Next, the pairwise margin ranking function is optimized to learn the semantic embedding:

$$L'(t^m, v^m, v^n) = \max(0, \mu - g(h[t^m, v^m]) + g(h[t^m, v^n])) \tag{3}$$

where $g(\cdot)$ learns the embedding score given the hidden features $h[t^m, v^m]$ and $h[t^m, v^n]$ from T-LSTM. This objective function tries to make sure that the score defined by $g(\cdot)$ is at least greater than μ for correct pair (t^m, v^m) compared to the contrastive pair (t^m, v^n).

Similar to [21], we use a multi-layer perception (MLP) to learn the score of each sentence and image pair given their hidden state. In particular, we define $g(\cdot)$ as follows:

$$g(h_i) = W_2^e(\delta(W_1^e h_i)) \tag{4}$$

where W_2^e and W_1^e are the parameters of the MLP and $\delta(\cdot)$ is the activation function for the hidden state.

Multi-task Learning for Joint Sentiment Analysis Our model can be trained by a simultaneous optimization of both the sentiment analysis and the embedding tasks. Figure 3 shows the overall structure of the proposed model with the two tasks. The semantic embedding learning task is implemented using a Siamese network, where the parameters of T-LSTM (W_1 in Figure 3) and the parameters of the scoring MLP (W_2 in Figure 3) are shared. The loss of one training image-sentence pair is the summation of the sentiment classifier loss and the margin ranking loss (Eq.(3)). For each training pair (t^m, v^m), we randomly choose one image v^n to make a contrasting pair (t^m, v^n), which is also given as the input to the semantic embedding module. We use a mini-batch gradient descent algorithm with an adaptive learning rate to optimize the loss functions.

4. EXPERIMENTAL RESULTS

For visual sentiment analysis, we mainly focus on Visual Sentiment Ontology[2] dataset. In addition, we build several datasets for joint visual-textual sentiment analysis. Two of them are crawled from Flickr and Getty Images individually. They are weakly la-

[1]When the whole data set can fit into the machine's memory, it is also possible to employ full-batch L-BFGS.

[2]http://visual-sentiment-ontology.appspot.com

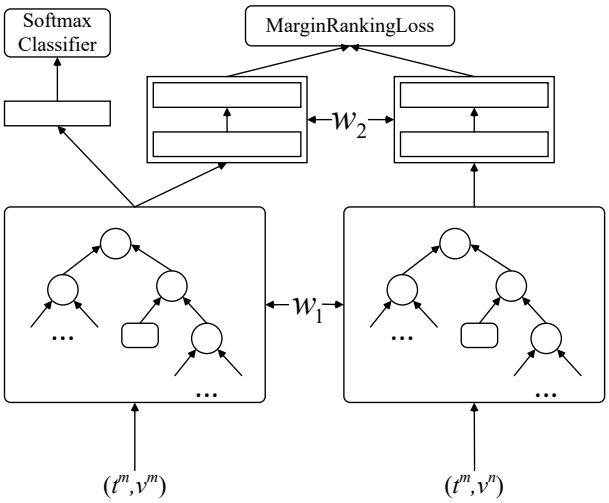

Figure 3: A multi-task learning framework for jointly training the sentiment classifier as well as the semantic embedding. (t^m, v^m) is the correct sentence and image pair and v^n is a randomly picked image from the training set.

belled. The third one is labelled by Amazon Mechanical Turks. Details on the dataests can be found in our work [37].

4.1 Performance analysis

Visual sentiment analysis Table 1 shows the performance of both CNN and PCNN on the 10% randomly chosen testing data. PCNN outperformed CNN in terms of *Precision*, *Recall*, *F1* and *Accuracy*. Both algorithms show better performance than results using lower-level and mid-level features (for more results see [35]).

Joint visual-textual sentiment analysis We compare the two proposed joint models for sentiment analysis. The results of the following baselines are also reported and analyzed. We also compare with these baselines: 1) single visual model, 2) single textual model, 3) early fusion and 4) late fusion. Due to the limit of space, we only compare the performance on the Twitter testing dataset for different algorithms. Table 2 summarizes the results. Both CCR and T-LSTM have improved the performance for joint sentiment analysis. The authors [37] have indicated that they collect the weak or noisy sentiment labels of the image tweets by analyzing the Tweets text only using a rule-based classifier VADER [11]. It is possible to generate overlap information with the parsing trees, which may explain the significant performance improvements of T-LSTM over previous models on this dataset.

Table 1: Performance on the Testing Dataset by CNN and PCNN.

Algorithm	Precision	Recall	F1	Accuracy
CNN	0.714	0.729	0.722	0.718
PCNN	**0.759**	**0.826**	**0.791**	**0.781**

4.2 Qualitative analysis of the results

Figure 4 shows two positive and negative examples for T-LSTM Attention. This indicates that the auxiliary embedding learning task helps the model to align words and image regions. Furthermore, the sentiment analysis task allows the attention mechanism to focus more on sentiment related regions. For instance, positive words *love* in Figure 4(a) and *sad* in Figure 4(b) of the negative example.

Table 2: Results of different T-LSTM variants and previously reported results on the Twitter testing dataset.

Model	Precision	Recall	F1	Accuracy
Textual	0.746	0.693	0.727	0.722
Visual	0.584	0.561	0.573	0.553
Early Fusion	0.730	0.744	0.737	0.717
Late Fusion	0.634	0.610	0.622	0.604
CCR	0.831	0.805	0.818	0.809
T-LSTM Embedding	**0.958**	**0.977**	**0.967**	**0.964**

(a) T-LSTM Attention (b) T-LSTM Attention

Figure 4: Visualization of *attention* on two examples with positive sentiment (left) and negative sentiment (right).

5. CONCLUSIONS AND FUTURE WORK

Social multimedia sentiment analysis is a challenging and interesting problem. In this thesis proposal, we adopt deep learning to solve this problem. We have designed a new architecture, as well as new training strategies to overcome the noisy nature of the large-scale training samples. We also present two frameworks for joint visual-textual sentiment analysis. Both of which are trying to integrate textual and visual information into a unified model. The experimental results suggest that convolutional neural networks that are properly trained can outperform both classifiers that use predefined low-level features or mid-level visual attributes for the highly challenging problem of visual sentiment analysis. Meanwhile, experimental results have demonstrated that the proposed joint models have significantly improved the performance of joint textual-visual sentiment analysis on several datasets. We hope our sentiment analysis results can encourage further research on online user generated content.

Future work We think there are several topics to explore to further promote the performance of machines on this task: 1) Localized visual sentiment. A more interesting task is to automatically identify the local regions of an image, which is mostly close to visual sentiment. 2) Fine-grained sentiment. We have been focusing on a binary classification. The next step would be analyzing the fine-grained sentiments. We have release a dataset on this topic [36]. 3) Sentiment for bridging visual-textual gap. We believe that the understanding of visual and textual sentiment will lead to better visual-textual semantics, which is helpful for tasks such as image captioning [23].

6. REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*, 2014.

[2] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.

[3] D. Borth, T. Chen, R. Ji, and S.-F. Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM MM*, pages 459–460. ACM, 2013.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, pages 223–232. ACM, 2013.

[5] D. Borth, R. Ji, T. Chen, T. M. Breuel, and S. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, pages 223–232, 2013.

[6] D. Cao, R. Ji, D. Lin, and S. Li. A cross-media public sentiment analysis system for microblog. *Multimedia Systems*, pages 1–8, 2014.

[7] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, June 2015.

[8] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *IJCAI*, pages 1237–1242, 2011.

[9] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2121–2129, 2013.

[10] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *WWW*, pages 607–618, 2013.

[11] C. J. Hutto and E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, 2014.

[12] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: using flickr for prediction and forecast. In *ACM MM*, pages 1235–1244. ACM, 2010.

[13] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

[14] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[15] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, volume 1, page 4, 2012.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, 2012.

[18] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *ACM MM*, pages 457–466. ACM, 2014.

[21] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[22] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015.

[23] A. P. Mathews, L. Xie, and X. He. Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580, 2016.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[25] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*, pages 169–176, 2011.

[26] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. Analyzing and predicting sentiment of images on the social web. In *ACM MM*, pages 715–718. ACM, 2010.

[27] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218, 2014.

[28] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, pages 129–136, 2011.

[29] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15(1):2949–2980, 2014.

[30] I. Sutskever. *Training recurrent neural networks*. PhD thesis, University of Toronto, 2013.

[31] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, pages 1556–1566, July 2015.

[32] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, June 2015.

[33] M. Wang, D. Cao, L. Li, S. Li, and R. Ji. Microblog sentiment analysis based on cross-media bag-of-words model. In *ICIMCS*, pages 76:76–76:80. ACM, 2014.

[34] J. Weston, S. Bengio, and N. Usunier. WSABIE: scaling up to large vocabulary image annotation. In *IJCAI*, pages 2764–2770, 2011.

[35] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, pages 381–388, 2015.

[36] Q. You, J. Luo, H. Jin, and J. Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, pages 308–314, 2016.

[37] Q. You, J. Luo, H. Jin, and J. Yang. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *WSDM*, pages 13–22, 2016.

[38] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM, 2013.

[39] X. Zhu, P. Sobhani, and H. Guo. Long short-term memory over recursive structures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1604–1612, 2015.