Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Visual Sentiment Analysis by Attending on Local Image Regions

Quanzeng You
Department of Computer Science
University of Rochester
Rochester, NY 14627
qyou@cs.rochester.edu

Hailin Jin
Adobe Research
345 Park Avenue
San Jose, CA 95110
hljin@adobe.com

Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY 14627
jluo@cs.rochester.edu

Abstract

Visual sentiment analysis, which studies the emotional response of humans on visual stimuli such as images and videos, has been an interesting and challenging problem. It tries to understand the high-level content of visual data. The success of current models can be attributed to the development of robust algorithms from computer vision. Most of the existing models try to solve the problem by proposing either robust features or more complex models. In particular, visual features from the whole image or video are the main proposed inputs. Little attention has been paid to local areas, which we believe is pretty relevant to human's emotional response to the whole image. In this work, we study the impact of local image regions on visual sentiment analysis. Our proposed model utilizes the recent studied attention mechanism to jointly discover the relevant local regions and build a sentiment classifier on top of these local regions. The experimental results suggest that 1) our model is capable of automatically discovering sentimental local regions of given images and 2) it outperforms existing state-of-the-art algorithms to visual sentiment analysis.

Introduction

Visual sentiment analysis studies the emotional response of humans on visual stimuli such as images and videos. It is different from textual sentiment analysis (Pang and Lee 2008), which focus on human's emotional response on textual semantics. Recently, visual sentiment analysis has achieved comparable performance with textual sentiment analysis (Borth et al. 2013; Jou et al. ; You et al. 2015). This can be attributed to the success of deep learning on vision tasks (Krizhevsky, Sutskever, and Hinton 2012), which makes the understanding of high-level visual semantics, such as image aesthetic analysis (Lu et al. 2014), and visual sentiment analysis (Borth et al. 2013), tractable.

The studies on visual sentiment analysis have been focused on designing visual features, from pixel-level (Siersdorfer et al. 2010a), to middle attribute level (Borth et al. 2013) and to recent deep visual features (You et al. 2015; Campos, Jou, and Giro-i Nieto 2016). Thus, the performance of visual sentiment analysis systems has been gradually improved due to more and more robust visual features. However, almost all of these approaches have been trying to reveal the high-level sentiment from the global perspective of the whole images. Little attention has been paid to research from which local regions have we obtain the sentimental response and how is the local regions towards the task of visual sentiment analysis. In this work, we are trying to solve these two challenging problems. We employ the recent proposed attention model (Mnih et al. 2014; Xu et al. 2015) to learn the correspondence between local image regions and the sentimental visual attributes. In such a way, we are able to identify the local image regions, which is relevant to sentiment analysis. Subsequently, a sentiment classifier is built on top of the visual features extracted from these local regions.

To the best of our knowledge, our work is the first to automatically discover the relevant local images and build a sentiment classifier on top of the features from these local image regions. Indeed, Chen *et al.* (Chen et al. 2014) has been trying to identify the local regions corresponding sentiment related adjective noun pairs. However, their approach is limited to hand-tuned small number of adjectives and nouns. The work in (Campos, Jou, and Giro-i Nieto 2016) tries to visualize the sentiment distribution over a given image using a fine-tuned fully convolutional neural network on the given images. Their results are obtained by using the global images and the localization is only used for visualization purpose.

We evaluate the proposed model on the publicly available Visual Sentiment Ontology dataset[1], which is the largest available dataset for visual sentiment analysis. We will learn both the attention model and the sentiment classifier simultaneously. The performance on sentiment analysis using local visual features will be reported. Meanwhile, we will also quantitatively validate the attention model on discovering sentiment relevant local image regions.

Related work

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 (Krizhevsky, Sutskever, and Hinton 2012). Consequently, the task of higher-level

[1] http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html

semantic understanding, such as machine translation (Bahdanau, Cho, and Bengio 2014), image aesthetic analysis (Lu et al. 2014), and visual sentiment analysis (Borth et al. 2013; You et al. 2015) 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 problem.

The successes of deep learning make the understanding and jointly modeling vision and language content a feasible and attractive research topic. 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 (Frome et al. 2013) and recently onto sentence-level descriptions (Kiros, Salakhutdinov, and Zemel 2014; Socher et al. 2014; Ma et al. 2015; Karpathy and Li 2015), deep neural networks exhibit significant performance improvements on these tasks. Despite of the fact that there are no semantic and syntactic structures, these models have inspired the idea of joint feature learning (Srivastava and Salakhutdinov 2014), semantic transfer (Frome et al. 2013) and design of margin ranking loss (Weston, Bengio, and Usunier 2011).

In this work, we focus on visual sentiment analysis, which is different from the widely studied textual sentiment analysis (Pang and Lee 2008). It is quite new and challenging. There are several recent works on visual sentiment analysis using initially pixel-level features (Siersdorfer et al. 2010b), then mid-level attributes (Borth et al.), and more recently deep visual features (You et al. 2015) and unsupervised framework (Wang et al. 2015). These approaches have achieved acceptable performance on visual sentiment analysis. However, due to the complex nature of visual content, the performance of visual sentiment analysis still lags behind textual sentiment analysis.

There are also several publications on analyzing sentiment using multi-modalities, such as text and image. Both (Wang et al. 2014) and (Cao et al. 2014) 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 (Borth et al.). More recently, You *et al.* (You et al. 2016b) proposed a cross-modality consistent regression (CCR) scheme for joint textual-visual sentiment analysis. Their approach employed deep visual and textual features to learn a regression model. Their model achieved the best performance over other fusion models, however, overlook the mapping between image regions and words.

Our work is first to consider the local visual regions induced by sentiment related visual attributes. We build our model on the recent proposed attention model, which is capable of learning the context semantics (Bahdanau, Cho, and Bengio 2014; Xu et al. 2015) or semantic mappings (You et al. 2016a) between two representations.

The Model

We study the problem of predicting sentiment label of a given image. Beyond this main task, we are particularly interested in studying the mechanism behind visual sentiment response. We want to show where and how localized image regions awake people's sentiment response towards a given image. To achieve that goal, we need to discover image regions related

Attention model

Recently, attention model (Bahdanau, Cho, and Bengio 2014; Mnih et al. 2014) is employed to solve various tasks in natural language processing and computer vision. In particular, attention model is able to learn the mappings between different inputs at a given context. In sequence-to-sequence learning for machine translation (Bahdanau, Cho, and Bengio 2014) and semantic parsing (Vinyals et al. 2015), attention mechanism is employed to learn the context vector on the encoder's hidden state. We denote the encoder's hidden state as (h_1,h_2, $\cdots$, h_T) and the decoder's hidden state as (h'_1,h'_2, $\cdots$, h'_T). The attention vector at decoder's t-th time step is computed as:

$$\beta_i^t = v^T \tanh(W_1 h_i + W_2 h'_t) \tag{1}$$

$$\alpha_i^t = \text{softmax}(\beta_i^t) \tag{2}$$

$$c_t = \sum_{i=1}^{T} \alpha_i^t h_i \tag{3}$$

Next, the attention vector c_t is usually concatenated with h_t to produce the input for the next layer in the network. In such a way, attention model is able to find the relevant information for the current state, and hence promote the performance of the overall model.

Attention model is also able to bridge the gap between data from different modalities (Xu et al. 2015; You et al. 2016a). Given a image and one descriptive word of the image, we assume that the attribute word is likely associated with some local regions in the image. Our goal is to automatically find such kind of connections between the descriptive word and the image regions. Let t_i denote the i-th given word and let $V_i = \{v_{1i}, v_{2i}, \ldots, v_{ni}\}$ denote the regions of the corresponding i-th image, and n is the number of image regions. In attention model, a score α_j ($1 \leq j \leq n$) is assigned to each image region v_j based on its relevance with the content of v_j. As a common approach to model relevance in vector space, a bilinear function is used to evaluate α_{ij}:

$$\alpha_{ij} \propto \varphi\left(t_i^T U v_{ji}\right), \tag{4}$$

where the α_i.s are taken to normalize over all the $\{v_j\}$, $\varphi(\cdot)$ is a smooth function, and U is the weight matrix to be learned. One popular choice for $\varphi(\cdot)$ is the $\exp(\cdot)$ as in the softmax function.

In such a way, we can calculate the attention score to modulate the strength of relatedness between the descriptive word and different image regions. We are also able to calculate the weighted sum of all candidate local regions, which is a mapped visual features for the image.

$$v^i = \sum_{k=1}^{n} \alpha_{ik}(U v_{ki}). \tag{5}$$

We obtain a weighted visual feature mapping v^i for the given descriptive word t. Next, we can supply v^i as input to a sentiment classifier, which is based on a weighted sum of the local visual features.

Localization of visual regions for visual sentiment analysis

We present our model in this section. We assume that we are given one image and one descriptive *attribute* of the image. Since we are interested in visual sentiment analysis, the given attribute is sentiment related visual attribute. The overall framework is shown in Figure 1. The overall end to end system accepts image and attribute pairs. Local image regions are represented using convolutional layer features in order to learn the attention module (Xu et al. 2015). We follow the same strategy to represent local image regions using convolutional layers.

At the same time, these convolutional layers are also shared with the task of attribute detector. Following several fully-connected layers, the main goal of the attribute detector is to learn a attribute classifier for a given image. In the training stage, the ground-truth attribute is given for each image, which is utilized to learn the attribute classifier. In the testing state, the attribute classifier can be employed to predict the attribute for any given image. The negative log-likelihood (NLL) is employed to calculate the cost for the attribute detector.

$$L(t_i, V_i) = -\log\left(p(F(V_i), y_{t_i})\right) \tag{6}$$

where $F(V_i)$ is the output of fully connected layer and y_{t_i} is the attribute label for the i-th image.

The inputs to the attention model are pairs of image and its attribute. Using the bilinear attention model introduced in previous section, we are able to produce a weighted representation of the images' local features. Next, this representation can be supplied as input to build a softmax classifier for sentiment analysis. In such a way, we can solve the problem of visual sentiment analysis. We employ the negative log-likelihood (NLL) to define the cost:

$$p(h[t_i, V_i]) = \text{softmax}\left(W_s \tanh(W_h v^i)\right) \tag{7}$$

$$L'(t_i, V_i) = -\log\left(p(h[t_i, V_i], l_i)\right) \tag{8}$$

where $v^i = h[t_i, V_i]$ is the output produced by the attention module, W_h and W_s are the parameters for the multi-layer perceptron, and l_i is the sentiment label for the i-th image and attribute pair. The overall network can be trained using back propagation.

Experiments

We evaluate the proposed model on the publicly available benchmark dataset visual sentiment ontology (VSO). This dataset is collected by querying Flickr with adjective noun pairs (ANPs). These adjectives are considered to be sentiment related. Thus, each image is related to one ANP and each image is labelled according to the sentiment label of its ANP. In total, there are 3244 ANPs and about 1.4 million images.

We crawl all the images according to the provided URLs. After removing invalidated URLs, we obtain a total of 1.3 million images. However, the dataset is imbalanced. Because there are more positively labelled images, we randomly sample the same number of positive images with the negative images to manually build a balanced dataset. In the end, we have 1.1 million images, half of them is positive and the remaining is negative. We randomly split them into 80% for training, 10% for testing and 10% for validating.

Model settings

In our implementation, the convolutional layers, step b) in Figure 1, are initialized using the VGG-16 (Simonyan and Zisserman 2014) convolutional layers, which is pre-trained on the ImageNet classification challenge. The feature map of the last convolutional layer is $14 \times 14 \times 512$. In other words, the attention model will operate on these 196 flattened features, which is the same with (Xu et al. 2015).

Next, we need to choose feature representations for the attribute words. There are two popular approaches to represent words. The first is one-hot representation with an embedding layer. In particular, the goal is to map a word w_i with representation $w_i = [0, \ldots, 1_i, \ldots, 0] \in R^{|V|}$ (only the i-th position is one in the one-hot representation) to $e_i \in R^m$, where $|V|$ is the size of the vocabulary and m is the size of embedding layer. The second approach is to directly employ the pre-trained distributed representations of words, such as Word2Vec (Mikolov et al. 2013) and GloVe (Pennington, Socher, and Manning 2014). We use the pre-trained 300-dimensional GloVe features to represent words, which has been employed for sentiment analysis (Tai, Socher, and Manning 2015) and textual-visual semantic learning (You et al. 2016a). This is particular helpful since insufficient text data may not lead to well learned word features in the one-hot representation setting.

All the parameters are automatically learned by minimizing the two loss functions over the training split. We use a mini-batch gradient descent algorithm with an adaptive learning rate to optimize the loss functions.

Preliminary experiments

Before conducting the experiments using the proposed model in Figure 1, we first experiment with the GloVe features and test the upper bound of the system. The entire number of adjectives is 269. 127 of them are labelled positive and the remaining are negative. We extract the 300-dimensional pre-trained GloVe[2] vectors to represent these 269 adjectives. Next, we build a logistic classifier on top of these features. We only use a linear model to transform the given GloVe feature vectors. The model achieves an accuracy of $95\%(\pm 0.08)$ using a 5-fold cross-validation. On the other hand, the current state-of-the-art performance of visual sentiment analysis is around 80%. This result implies that the semantic meaning of the pre-trained GloVe feature vectors is helpful for the task of sentiment analysis. We would expect an extra advantage by matching these textual semantic embedding using local attended image regions.

Later, we train our model using the ground truth adjective for training split and we also test the model using the ground truth adjectives. We do not train the attribute detector. In such a way, we can produce the performance *upper bound* of our framework. Table 1 shows the performance

[2]http://nlp.stanford.edu/projects/glove/

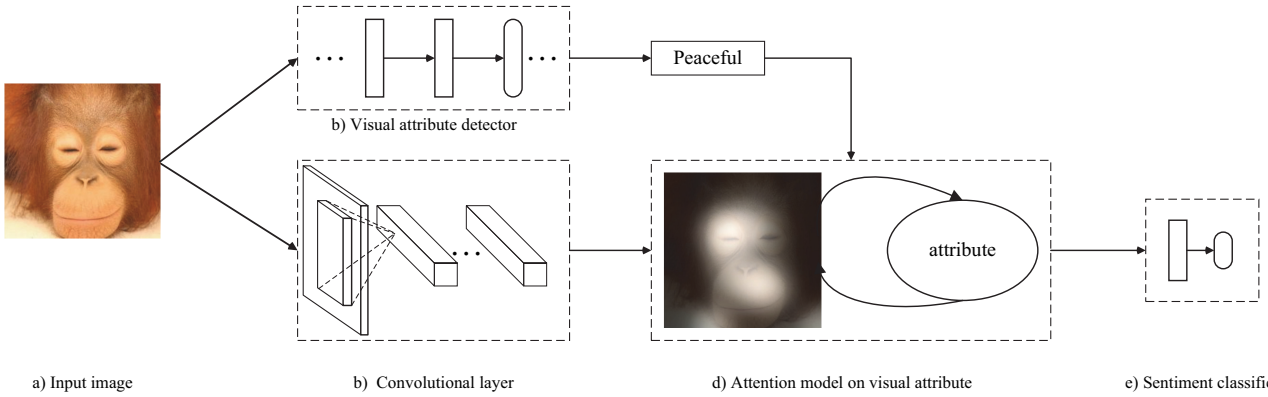


Figure 1: Overall end to end architecture for localized visual sentiment analysis. The system has several different modules. It accepts a image as input. Visual features extracted from convolutional layers and visual attribute which is the output of a visual attribute detector are supplied as inputs to the attention module. The attention model discovers the correspondence between local image regions and textual visual attribute. Sentiment classifier accepts the weighted sum of semantic local visual features produced by the attention model as inputs to train a multi-layer perceptron.

Split	Precision	Recall	F1	Accuracy
Validating	1.000	1.000	1.000	1.000
Testing	0.998	0.997	0.998	0.998

Table 1: Performance upper bound of our model.

Metric	Accuracy
Top-1	27.7%
Top-5	60.3%

Table 2: Accuracy of the visual attribute detector.

of our model on both the validating and testing splits. After only 2 epochs over the training split, the classification on validating split has been all correct. By providing the ground truth attribute, the model shows significant performance improvement on the visual sentiment analysis.

Quantitative analysis of attentions

We also try to visualize the attention weights. In particular, Xu, et al. (Xu et al. 2015) have employed upsampling and Gaussian filtering to visualize attention weights. In this section, we follow the same steps to visualize the attention weights of the ground truth visual attributes.

Figure 2 show several positive examples. Overall, the proposed model tends to learn accurate attention given the ground truth visual attributes. This helps us understand why the model has almost perfect performance on the visual sentiment analysis task. Localized visual regions extract robust and accurate visual representations, which lead to the significant improvement of sentiment classifier.

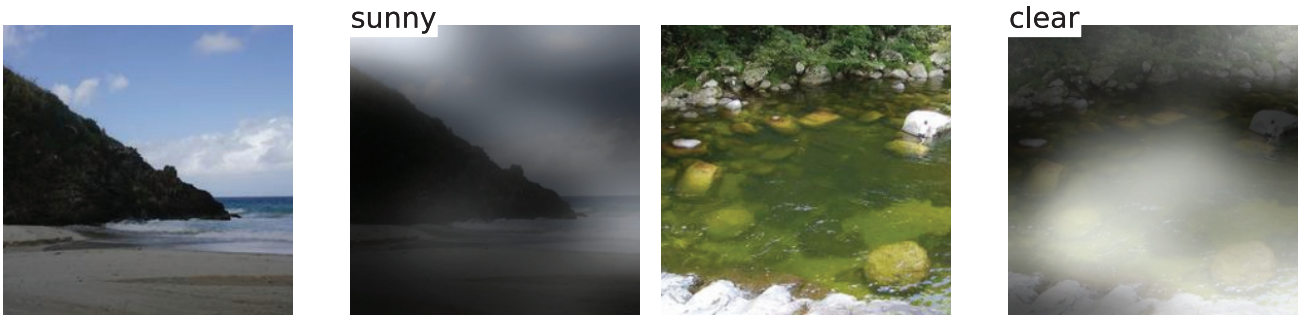
Training attribute detector

The results in previous sections suggest that attention model is able to find the matching local image regions given the ground truth visual attribute. Subsequently, we can obtain a robust visual sentiment classifier trained on these attended local regions. However, instead of using the ground truth visual attribute, a more interesting approach is to automatically discover the visual attributes and thus build a visual sentiment classifier on these attributes. Indeed, visual attribute detection is one of the most challenging problems in computer vision. Recent work (Escorcia, Niebles, and Ghanem 2015) has studied on utilizing CNN for visual attribute detection. In particular, we follow the study from Jou and Chang (Jou and Chang 2016), which has compared different architectures on the performance of Visual Ontology dataset.

Because the number of images in each ANP of VSO follows a long tail distribution, we follow the steps in (Jou and Chang 2016) to preprocess the data set. We keep adjective noun pairs with at least 500 images and filtered out some abstract and general nouns. Next, we keep those adjectives which has at least 6,000 images. To build a relatively balanced dataset, we randomly select $6,000$ images for those ANPs with more than $6,000$ images. In total, we obtain a dataset with 32 visual attributes and $6,000$ images for each attribute. Figure 3 shows the architecture for train the visual attribute detector. We fine-tuned on top of the pre-trained VGG-16 (Simonyan and Zisserman 2014) by adding one an adaptation fully-connected layer (Oquab et al. 2014). The total $192,000$ images are split into 80% for training, 10% for validating and 10% for testing.

We train the detector using Caffe with mini-batch stochastic gradient descent. We use the validating split of the dataset for early stopping and hyper-parameter selection. Table 2 shows the top-1 accuracy and top-5 accuracy on the testing split of the dataset. This performance is comparable with Jon and Chang (Jou and Chang 2016). Next, we use this model as the visual attribute detector in Figure 1 to train a visual sentiment classifier. Specifically, we use this visual attribute detector to predict at all the training, validating and testing splits of the dataset. However, since the top-5 accuracy is

(a) Angry (b) Abandoned



(c) Sunny (d) Clear

Figure 2: Visualization of *attention* on several selected examples.

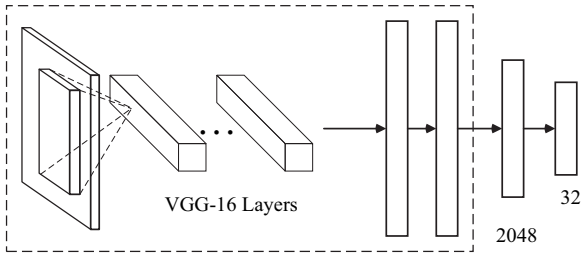


Figure 3: Deep architecture for training visual attribute detector.

Algorithm	Accuracy	F1
Global Multi-task CNN	70.3%	70.4%
Local Attention	69.4%	69.3%

Table 3: Performance of the two sentiment classifier using global and local visual features respectively.

much better than the top-1 accuracy, we use 5 attributes instead of 1 attribute to train our sentiment classifier. For each of the 5 predicted attribute, we use Eq.(5) to compute the attention local visual features. Then, these local visual features are concatenated and passed to the sentiment classifier as inputs.

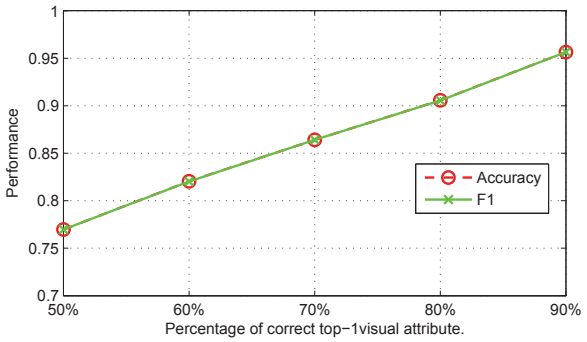
To compare the performance of this sentiment classifier, we also train another deep CNN model using global visual features. Specifically, we follow the mask-task settings proposed in (Jou and Chang 2016) to train the global visual sentiment classifier. In our settings, there are three tasks: prediction of the visual attribute (adjective), prediction of the object (noun) and prediction of the sentiment. All these tasks share the same lower layers of VGG-16 (see Figure 3). Meanwhile, each task has their own adaptive layer targeted for each individual task. Both the local and global sentiment classifier are trained using the same splits with the visual attribute detector task. Table 3 indicates that the performance of the two models are comparable. Global CNN in a multi-task settings show a relatively better performance than the local attention model. Considering the relatively poor performance of the visual attribute detector, the performance of local features on visual sentiment analysis is acceptable.

Manually curated visual attributes

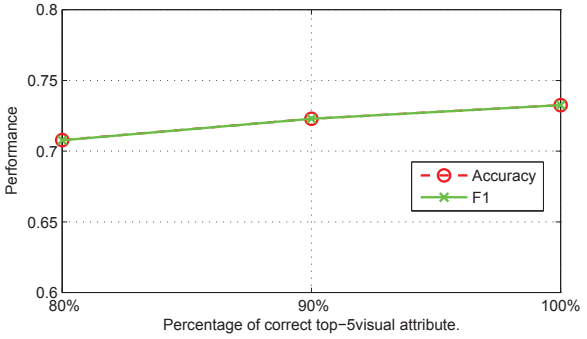
In previous section, we study the performance of our model using a relative poor attribute detector. It is interesting to check the performance of our model by providing more accurate visual attributes. Indeed, in most of the current image networks, such as Flickr (http://www.flickr.com) and Adobe Stock (https://stock.adobe.com), users are allowed to add tags and descriptions to their uploaded images. Most of the time, users are likely to carefully choose these text data for their images to create high quality albums and share with other users. In this section, we simulate users' curated visual attributes by randomly selecting different level of correct visual attributes.

This experiment follows the same steps in previous section. However, we manually change the predicted visual attributes of the previously trained attribute detector. We study

the performance of two strategies: 1) For the incorrectly predicted visual attribute (top-1), we randomly replace some of them with the ground truth visual attribute. We study the performance of providing different percentages of correctly top-1 visual attributes[3]. 2) Instead of provide correct top-1 visual attribute, we provide the correct attribute to randomly replace one of the top-5 predicted attributes. Specifically, for samples where all top-5 attributes are incorrect, we just randomly replace one of them with the ground truth visual attribute. In such a way, we are able to manually curate visual attributes for all the images in the three splits.



(a) Manually curated on top-1



(b) Manually curated on top-5

Figure 4: Performance of the proposed model on visual sentiment analysis with different level of manually curated visual attributes.

Next, we train a local sentiment prediction model using the two curated datasets individually. Figure 4(a) shows the accuracy and the F1 score of the proposed model given different percentages of correct top-1 visual attributes. As expected, the model performs better when more correct visual attributes are provided. In particular, the performance almost linearly increases with the percentage of correct top-1 visual attributes. Meanwhile, the performance of our model is also increased with more correct top-5 manually curated visual attributes. However, the increase is not as significant as the top-1 case. This is expected given the fact that the top-1 accuracy can only achieve 35.8% even when we manually curate the top-5 accuracy to 100%. These results indicate that the proposed attention model needs good attributes in order to have better visual sentiment analysis results. However, it is interesting to see that the proposed attention mechanism make the localization of sentiment related image regions possible, which is another interesting and challenging research problem.

[3]The samples with correct visual attributes include both the correctly predicted samples by the visual attribute detector and the randomly replaced samples

Conclusions

Visual sentiment analysis is a challenging and interesting problem. Current state-of-the-art approaches focus on using visual features from the whole image to build sentiment classifiers. In this paper, we adopt attention mechanism to discover sentiment relevant local regions and build sentiment classifiers on these localized visual features. The key idea is to match local image regions with the descriptive visual attributes. Because visual attribute detector is not our main problem to solve, we have experimented with different strategies of generating visual attributes to evaluate the effectiveness of the proposed model. The experimental results suggest that more accurate visual attributes will lead to better performance on visual sentiment analysis. In particular, the studied attribute detector, which is a basic and direct fine-tuning strategy on CNN, could lead to comparable performance of CNN using global visual features. More importantly, the utilization of attention model enables us to match the local regions in an image, which is much more interesting. We hope that our work on using local image regions can encourage more studies on visual sentiment analysis. In the future, we plan to incorporate visual context and large scale user generated images for building rich and robust attribute detector, localizing sentiment relevant local image regions and learning robust visual sentiment classifier.

Acknowledgment

This work was generously supported in part by Adobe Research and New York State through the Goergen Institute for Data Science at the University of Rochester.

References

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

Borth, D.; Chen, T.; Ji, R.; and Chang, S.-F. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content.

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

Campos, V.; Jou, B.; and Giro-i Nieto, X. 2016. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *arXiv preprint arXiv:1604.03489*.

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

Chen, T.; Yu, F. X.; Chen, J.; Cui, Y.; Chen, Y.-Y.; and Chang, S.-F. 2014. Object-based visual sentiment concept

analysis and application. In *Proceedings of the ACM International Conference on Multimedia*, 367–376. ACM.

Escorcia, V.; Niebles, J. C.; and Ghanem, B. 2015. On the relationship between visual attributes and convolutional networks. In *CVPR 2015*, 1256–1264. IEEE.

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

Jou, B., and Chang, S.-F. 2016. Deep cross residual learning for multitask visual recognition. *arXiv preprint arXiv:1604.01335*.

Jou, B.; Chen, T.; Pappas, N.; Redi, M.; Topkara, M.; and Chang, S.-F. Visual affect around the world: A large-scale multilingual visual sentiment ontology.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137.

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

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, 457–466. ACM.

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

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 3111–3119.

Mnih, V.; Heess, N.; Graves, A.; and kavukcuoglu, k. 2014. Recurrent models of visual attention. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 2204–2212.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1717–1724.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1-2):1–135.

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

Siersdorfer, S.; Minack, E.; Deng, F.; and Hare, J. 2010a. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*, 715–718. ACM.

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

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

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

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

Tai, K. S.; Socher, R.; and Manning, C. D. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1556–1566.

Vinyals, O.; Kaiser, Ł.; Koo, T.; Petrov, S.; Sutskever, I.; and Hinton, G. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, 2755–2763.

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

Wang, Y.; Wang, S.; Tang, J.; Liu, H.; and Li, B. 2015. Unsupervised sentiment analysis for social media images. In *24th International Joint Conference on Artificial Intelligence*. IJCAI.

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

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 381–388.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016a. Image captioning with semantic attention. In *CVPR 2016*.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016b. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM)*, 13–22.