DISCOVERING AFFECTIVE REGIONS IN DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VISUAL SENTIMENT PREDICTION

Ming Sun, Jufeng Yang, Kai Wang, Hui Shen*

College of Computer and Control Engineering, Nankai University, China
* corresponding author: yangjufeng@nankai.edu.cn

ABSTRACT

In this paper, we address the problem of automatically recognizing emotions in still images. While most of current work focus on improving whole-image representations using CNNs, we argue that discovering affective regions and supplementing local features will boost the performance, which is inspired by the observation that both global distributions and salient objects carry massive sentiments. We propose an algorithm to discover affective regions via deep framework, in which we use an off-the-shelf tool to generate N object proposals from a query image and rank these proposals with their objectness scores. Then, each proposal's sentiment score is computed using a pre-trained and fine-tuned CNN model. We combine both scores and select top K regions from the N candidates. These K regions are regarded as the most affective ones of the input image. Finally, we extract deep features from the whole-image and the selected regions, respectively, and sentiment label is predicted. The experiments show that our method is able to detect the affective local regions and achieve state-of-the-art performances on several popular datasets.

***Index Terms*—** Visual sentiment prediction, objectness estimation, affective region, CNNs

1. INTRODUCTION

Understanding the emotions of online visual content is of importance with the growing of users who prefer to use the images and videos to express their opinions. The precise prediction of visual sentiment is beneficial to a variety of applications, e.g. topic discovering, advertisement, social networks and online marketing.

Most existing work predict visual sentiment using the features extracted from the whole image [2], which are more or less inspired by psychology theory and principles of arts [3] [4]. While the emotion evoked by an image is not only from its global appearance but also interplays among local regions, to the best of our knowledge, few work have paid close attention to the usage of local features in sentiment analysis. Li et al. [5] proposed a context-aware classification model based on bilayer sparse representation. The main limits of [5] included that it depended heavily on the initial segmentation and took all local regions into account, among which most regions were neutral.

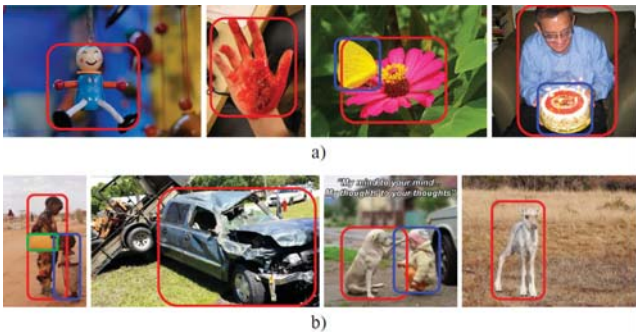

Fig. 1. Images from two popular datasets: a) Flickr [1], and b) Twitter [2]. People are affected by the local regions as well as the whole image apperance.

The recent trend in this area is to leverage deep neural networks to extract features and recognize emotions. You et al. [2] employed two convolutional layers and several fully connected layers for the prediction. They addressed the weakly labeled nature of the training image data by using a progressive training strategy. Chen et al. introduced DeepSentiBank [6], a visual sentiment concepts classification model which was trained under Caffe [7]. They found that initializing the model with the weights trained from ImageNet provided much better performance than training from visual sentiment dataset alone.

Aligned with the aforementioned trend, our approach also exploits deep neural networks for visual sentiment analysis. However, we introduce a notion named Affective Regions(ARs) into this area and extract the local features from ARs. An AR usually has two distinguishing characteristics: i) it's a salient region likely containing one or more objects, which could attract user's attention mostly; and ii) it car-

This work was supported by the National Natural Science Foundation of China(No.61301238, 61201424), China Scholarship Council(No.201506205024) and the Natural Science Foundation of Tianjin, China(No.14ZCDZGX00831).

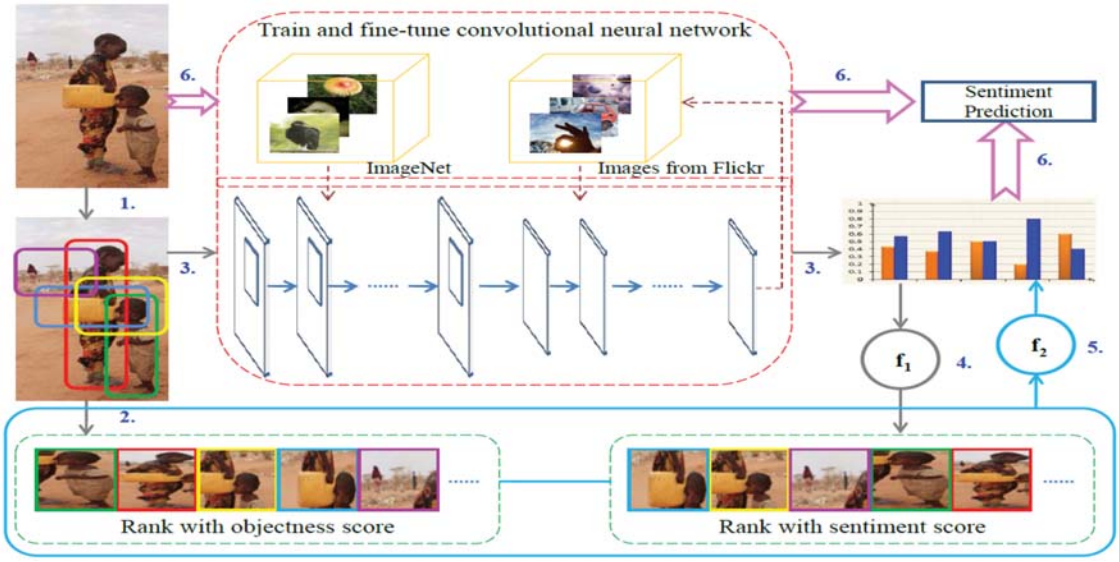

Fig. 2. Pipeline of the proposed approach. 1 & 2) First, we generate some object proposals and rank these candidates with objectness score. 3, 4 & 5) Then, sentiment score of each proposal is roughly computed using a pre-trained and fine-tuned deep convolutional neural network. We combine both objectness score and sentiment score to discover affective regions in low-level and high-level perspective, respectively. 6) Finally, the sentiment label is predicted with the local affective regions as well as the whole image. Details of f_1 and f_2 are described in Section 3.

ries massive emotions. Figure 1 shows some ARs in popular datasets [1] [2].

Our contributions are summarized as follows:

1) We develop an algorithm automatically discovering affective local regions which are likely containing objects and carrying massive emotions. Furthermore, the proposed algorithm is more general than literature because of its independence to object categories.

2) We build a novel visual sentiment prediction model in deep CNNs, which extracts features from both of the whole image and affective regions. Compared to previous work concerning contextual informations, it need no complicated preprocessing such as binarization and accurate object localization.

2. RELATED WORK

The general literature on visual sentiment classification is vast, in which the discussed work range from still images [1] [8] to videos [9]. Here, we focus on reviewing related work on affective image prediction, the use of Deep Convolutional Neural Networks, and objectness estimation.

Affective Image Prediction With a growing number of images being used to express opinions in social networks, image sentiment analysis has attracted more and more attentions. From the aspect of features used in this area, we can roughly divide prior work into low-level based [4] [10] and mid-level based methods [11] [12] [8]. Machajdik and Hanbury [4] developed methods to extract low-level features, such as colors and textures to represent the emotional content of an image. To bridge the affective gap between low-level features and high-level sentiment, Borth et al. [1] modeled a mid-level concept, Adjective Noun Pairs(ANPs). Yuan et al. [11] proposed Sentribute, an image-sentiment analysis algorithm based on 102 mid-level attributes, of which results were easier to interpret and ready-to-use for high-level understanding.

[8] is closest to our work. They built object detection models to recognize six frequent objects including car, dog, dress, face, flower and food, and proposed a unique classification model to handle attributive and proportional similarity between visual sentiment concepts. In contrast, our algorithm concentrates only on whether a selected region contains objects or not. It's generic over categories and we don't recognize concrete objects, which shows robustness in real applications.

Convolutional Neural Networks Several recent work have exploited deep convolutional neural networks for image sentiment prediction. Based on previous work [1], Chen et al. trained a deep CNN model on Caffe [7] and named it DeepSentiBank [6]. In [13] [14], two types of activations were used as image-level features, namely the 4096-dimension output from fc7 and the 1000-dimension output from fc8. You et al. [2] proposed PCNN, in which they obtained half million training samples by using a baseline sentiment algorithm [1] to label Flickr images. To make use of such noisy machine labeled data, they employed a progressive

strategy to fine-tune the deep network. Furthermore, [15] [16] combined visual and textual features in CNNs for multimodal affective analysis and retrieval.

Different from pioneer work [6] [2] using deep model in this area, which concentrated on constructing discriminative whole-image representations, we propose a novel framework which combines global and local visual features together. In this framework, the pre-trained CNN model is used to extract global features, as well as to select affective local regions. Then, most irrelevant and noisy regions carrying few emotions are dropped.

Objectness Estimation Objectness is usually represented as a value reflecting how likely an image window covers an object of any category. Based on the human reaction time that is observed and the biological signal transmission time that is estimated, human attention theories hypothesize that the human vision system processes only parts of an image in detail, while leaving others nearly unprocessed. This further suggests that before identifying objects, there are simple mechanisms in the human vision system to select possible object locations [17].

In this paper, we employ an off-the-shelf tool [18] to generate N proposals as candidate affective regions. Compared to the work requiring accurate segmentation [5] or concrete category information [8], it's much easier to acquire object proposals in the pre-processing stage, which has shown its generalization ability in a variety of datasets.

3. METHODOLOGY

In this section, we aim to discover affective regions. Given an image, we find object proposals at first. Inspired by R-CNN [19], we examine whether each proposal carries sentiment and select the most valuable regions. Finally, we predict image sentiment labels using both the affective regions and global image. Figure 2 shows the pipeline of our proposed approach.

Generating object proposals Previous work has proved that associating adjectives with concrete physical objects could make the combined visual concepts more detectable and tractable for image sentiment analysis [8]. Inspired by that, we argue that object proposals can be used as the potential affective regions. In fact, local regions covering objects are inclined to attract more attentions and evoke people's feelings. As studied in visual attention and objectness estimation [17], the potential object position could be generated based on some hypotheses, such as a different appearance from their surroundings.

In this paper, we use [18] to generate object proposals. To achieve the high recall, [18] employs a bottom-up strategy which generates hundreds of redundant proposals. Therefore, it's necessary for us to filter out the noisy regions, which are defined as containing little sentiment, at the initial stage of the work. For each image I, we ultimately get N proposals.

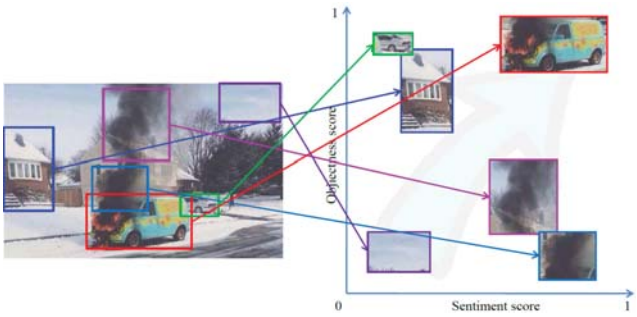

Fig. 3. Discovering affective regions which have two distinguishing characteristics: high objectness score and high sentiment score.

We use $Obj\_score_j^I$, provided by [18], to represent the objectness score, that is, the possibility of the j-th region containing an object of any category, where $j = 1, 2, ..., N$. Due to the strong co-occurrence relationship between sentiment and objects, the $Obj\_score_j^I$ is viewed as cue indicating the j-th region carrying sentiment in a low-level perspective.

Discovering affective regions Since we have got $Obj\_score_j^I$ in previous steps, in the following we focus on computing $Senti\_score_j^I$, which indicates the probability of the j-th region carrying sentiment in a high-level perspective. First, we build a rough global sentiment model by fine-tuning the VGG network [20] with the help of transfer learning. In details, we adopt the pre-trained deep architecture for object recognition [20], change the 1000-classes way into 2-classes, and fine-tune the model with a large-scale sentiment dataset [1]. The fine-tuned global model fills the semantic gap better than low-level representations. Then, we take each object proposal as the input of the deep model, and output the features from the last layer. As is shown in Equation (1), features from the last layer are represented as sentiment class distribution, which can be regarded as the sentiment probability of each proposal. Since our aim is to describe how much sentiment each proposal contains, we use the following probabilistic sampling function(f_1 in Figure 2) to evaluate the sentiment score of the j-th region.

$$Senti\_score_j^I = 1 + \sum_{i=1}^{M} s_{ij} \log s_{ij} \tag{1}$$

where M is the sentiment classes number in current system. Since we're distinguishing positive and negative affects in images, M is set to 2 in this paper. s_{ij} indicates the probability of the j-th proposal carrying the i-th emotion, where $i = 1, 2, ..., M$. Different from PCNN [2] which removes only a little portion of total images, 30% theoretically and actually less than 10%, we are selecting the most valuable one or several local regions for prediction. The intuition is that we want to keep proposals with distinct sentiment probabilities between the two classes with a high $Senti\_score_j^I$. In

Algorithm 1 Prediction using both affective regions and global image

Input:

Image: I

Number of affective regions: K

Output:

Image sentiment label : $label$

1: Generate N object proposals using [18] and get objectness score $Obj\_score_j^I$ for the j-th region.
2: Let $\vec{Y}_{Global}$ be the prediction response of the whole image using a pre-trained and fine-tuned model.
3: Let s_{ij} be the prediction response of the j-th proposal on the i-th emotion using the same model as Step 2.
4: **for** j $\in N$ **do**
5: Achieve sentiment score $Senti\_score_j^I$ for the j-th region. (Equation (1))
6: Using AR to evaluate the sentiment quality of each proposal in both of low-level and affective-level perspectives. (Equation (2))
7: **end for**
8: Rank proposals with AR and select top K affective regions.
9: Predict the $label$ using $\vec{Y}$. (Equation (3))
10: **return** $label$

contrast, s_{1j} and s_{2j} having similar prediction values usually indicates that it's difficult for people to summarize the emotions evoked by the j-th proposal.

Figure 3 shows that an affective region should have both high $Obj\_score_j^I$ and $Senti\_score_j^I$, which is summarized as the following equation(f_2 in Figure 2).

$$AR = \sqrt{Obj\_score_j^{I\,2} + \alpha Senti\_score_j^{I\,2}} \qquad (2)$$

where AR is the sentiment quality of each region and α is the tradeoff between low-level and affective-level perspectives. For a proposal with high AR, we suppose it's an affective region which is valuable for sentiment predication. Otherwise, we remove proposals with small AR from our candidate list. In our experiment, we set α = 1.

Predicting image sentiment In general, there are two kinds of strategies for the combination of global and local information. One is combining features together and the other is combining the prediction responses. In Section 4, our experiment shows that later fusion outperforms former fusion on all datasets. Therefore, we use the following function to predict image's sentiment distribution $\vec{Y}$.

$$\vec{Y} = \vec{Y}_{Global} + \frac{\beta}{K}\sum_{j=1}^{K}\vec{Y}_{AR_j} \qquad (3)$$

where $\vec{Y}_{Global}$ represents the prediction using the whole image and $\vec{Y}_{AR_j}$ represents the prediction using the j-th affective region. $\vec{Y}$, $\vec{Y}_{Global}$ and $\vec{Y}_{AR_j}$ share the similar vector structure of (P_{Pos}, P_{Neg}), where P_{Pos} and P_{Neg} indicate the predicted probability of positive and negative emotions, respectively. We select top K affective regions based on Equation (2). β is the tradeoff between global and local prediction. In our experiment, we set β = 0.5.

4. EXPERIMENTAL RESULTS

4.1. Experiment setup

Datasets We evaluate our proposed method on three widely used datasets. The datasets are from Twitter [2] [1] and Flickr [1], respectively. Twitter I dataset contains 1,269 images in total. For each image, the sentiment label is generated by five Amazon Mechanical Turk(AMT) workers. We test our method on all of the three subsets of Twitter I, including "Five agree", "At least four agree" and "At least three agree", in a similar fashion to [2]. Twitter II dataset [1] contains 603 images. Ground truths of sentiment values are obtained by AMT annotation too, resulting in 470 positive and 133 negative labels. Flickr dataset [1] contains 484,258 images, each of which is weakly labeled and annotated with an ANP. We randomly choose 90% images of this dataset to fine-tune our deep model and use the rest 10% images for testing.

Training and fine-tuning CNNs Convolutional neural networks have the capability to incorporate model weights learned from more general dataset, which can be applied to our case by transferring the model learned over ImageNet to the specialized sentiment dataset. Following previous work on visual sentiment analysis with deep networks from, e.g. [6][13][14], we initialize our model with the weights trained from ImageNet [20]. Then, we fine-tune the pre-trained model using half million Flickr images [1] and run a total of 100,000 iterations to extract more discriminative features. VGGNet [20] is employed for sentiment prediction in this paper. Due to lack of training images on both Twitter datasets, we also use the deep model fine-tuned on Flickr dataset to predict Twitter images with the help of transfer learning.

4.2. Baseline

In this section, we compare our experimental results with other state-of-the-art algorithms for image sentiment prediction, including hand-crafted features based and deep features based methods. In addition, we also show the results of our proposed method on different configurations, especially with various components and different fusion strategies.

Sentibank Borth et al. [1] proposed a systematic, data-driven methodology to construct a large-scale sentiment ontology built upon psychology and web crawled folksonomies. SentiBank is a concept detection library based on the constructed ontology to establish a mid-level representation for bridging the affective gap. In this paper, we exploit the pre-trained 1,200 ANP detectors of SentiBank to extract 1,200

Table 1. The prediction accuracy on Twitter I, Twitter II and Flickr dataset (%). ft represents fine-tuned model. obj and senti indicate we select affective regions using objectness score and sentiment score, respectively. ff is former fusion and lf is later fusion. Note that, for fair comparision, we implement PCNN based on VGGNet and show its results in the third line.

	Twitter I				
Method	Five agree	At least four agree	At least three agree	Twitter II	Flickr
SentiBank [1]	71.02	67.68	65.93	67.56	65.89
PCNN + CaffeNet [2]	77.30	70.09	68.01	70.80	66.69
PCNN + VGGNet [2]	86.06	81.81	78.72	75.23	69.95
VGGNet	83.51	78.89	75.36	72.82	60.88
VGGNet + ft	86.56	80.42	76.93	72.30	69.48
VGGNet + ft + obj + lf ($K = 1$)	85.22	78.57	74.48	75.55	69.38
VGGNet + ft + senti + lf ($K = 1$)	85.52	81.98	76.06	77.74	69.47
VGGNet + ft + obj + senti + lf ($K = 1$)	86.72	82.22	76.80	78.59	69.58
VGGNet + ft + obj + senti + ff ($K = 1$)	85.14	81.25	76.03	73.85	63.83
VGGNet + ft + obj + senti + lf (K)	**88.94** (11)	**84.90** (13)	**80.33** (11)	**78.97** (7)	**70.24** (12)

dimensional features. Previous work have proved these features outperform other low-level and mid-level hand-crafted features in visual sentiment analysis.

PCNN You et al. [2] proposed PCNN, a progressive convolutional neural network architecture. While they fine-tuned the deep network based on CaffeNet [7], we utilize a more powerful framework, VGGNet [20], to train our deep model. For a fair comparision, we also implement PCNN based on VGGNet. Table 1 shows the results with both PCNNs, where "PCNN + CaffeNet" is the former PCNN implemented in [2] and "PCNN + VGGNet" indicates the latter.

4.3. Results and analysis

Table 1 shows the sentiment prediction accuracy on three famous datasets. "ft" means that the CNN models are fine-tuned using half million Flickr images. Note that, both of the two PCNNs are fine-tuned in the same fashion to [2], even there's no explicit "ft" signs in their names. "obj" means that we only regard the proposals with high objectness score as affective regions, while "senti" refers to the proposals having high sentiment score. In a "obj + senti" method, we use Equation (2) to rank proposals and select affective regions, where both objectness score and sentiment score are considered. "ff" and "lf" are two fusion strategies. "ff" indicates the former fusion which combines deep features, while "lf" indicates the later fusion combining the prediction responses.

We show our experimental results on different configurations and compare to several state-of-the-art work. Our proposed method employing affective regions outperforms both hand-crafted features based [1] and deep features based approaches [2]. In details, compared to the SentiBank [1], our method outperforms by a large margin. Furthermore, our method also shows an advantage over other deep architectures [2], especially on both Twitter datasets having high-quality labels. PCNN was mainly designed to overcome the noisy labels of large-scale images, even that, our affective regions based method earns a better result on the weakly-labeled Flickr dataset.

When selecting and combining affective regions into deep model, we have different choices: we can use objectness score, sentiment score or both of them. We roughly regard the objectness score as a low-level cue and sentiment score as a high-level cue. The experimental results show that sentiment score is more reliable than objectness score. When both scores are combined into deep model, we can achieve the most valuable affective regions and reach the best performance.

We also compare different fusion methods over the whole image and affective regions. When former fusion is applied, we concatenate local features to global features. In contrast, when later fusion is applied, each kind of features are used to generate a prediction response separately. Then, we combine the responses with Equation (3), which outperforms the former fusion on all datasets.

Finally, we examine the impact of different "K" on the prediction. While [18] usually generates hundreds of proposals, we achieve the best performance using 10$\sim$13 affective regions. Figure 4 shows the results on three subsets of Twitter I. In fact, an image usually has very finite affective regions. When we increase the regions number, many of them have little related to image emotions and even weaken the prediction. This shows another advantage of our method, that is, we only need a few local regions involved in the deep model, indicating an acceptable increasing of overheads.

5. CONCLUSION

In this paper, we address the problem of automatically recognizing emotions in images. Inspired by the observation that both global appearance and salient objects carry massive sen-

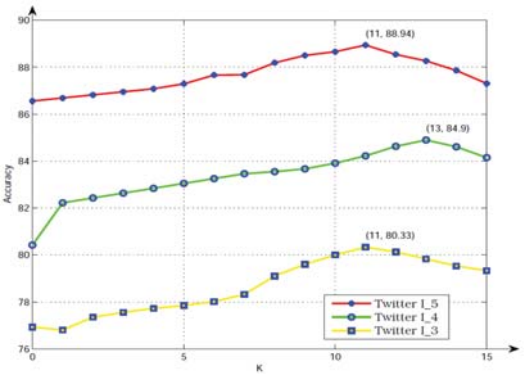

Fig. 4. Impact of different K on Twitter I, where Twitter I_5, Twitter I_4 and Twitter I_3 indicate "Five agree", "At least four agree" and "At least three agree", respectively.

timents, we propose an algorithm to discover affective regions and combine these local informations into a deep convolutional neural network. We employ later fusion strategy and implement the proposed model on VGGNet [20], the expremental results show that our method outperforms state-of-the-art on several popular datasets.

6. REFERENCES

[1] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 223–232.

[2] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[3] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 47–56.

[4] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 83–92.

[5] Bing Li, Weihua Xiong, Weiming Hu, and Xinmiao Ding, "Context-aware affective images classification based on bi-layer sparse representation," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 721–724.

[6] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[7] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[8] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang, "Object-based visual sentiment concept analysis and application," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 367–376.

[9] Sergio Benini, Luca Canini, and Riccardo Leonardi, "A connotative space for supporting movie affective recommendation," *Multimedia, IEEE Transactions on*, vol. 13, no. 6, pp. 1356–1370, 2011.

[10] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang, "Can we understand van gogh's mood?: learning to infer affects from images in social networks," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 857–860.

[11] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo 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*. ACM, 2013.

[12] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 2015, pp. 159–168.

[13] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li, "Visual sentiment prediction with deep convolutional neural networks," *arXiv preprint arXiv:1411.5731*, 2014.

[14] Victor Campos, Amaia Salvador, Brendan Jou, and Xavier Giró-i Nieto, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," *arXiv preprint arXiv:1508.05056*, 2015.

[15] Lei Pang, Shiai Zhu, and Chong-Wah Ngo, "Deep multimodal learning for affective analysis and retrieval," *Multimedia, IEEE Transactions on*, vol. 17, no. 11, pp. 2008–2020, 2015.

[16] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 2015, pp. 1071–1074.

[17] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3286–3293.

[18] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, "What is an object?," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 73–80.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.