Beyond Object Recognition: Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks

Jingwen Wang[1*], **Jianlong Fu**[2], **Yong Xu**[1], **Tao Mei**[2]
[1]South China University of Technology, Guangzhou, China
[2]Microsoft Research, Beijing, China
w.jingwen@mail.scut.edu.cn, yxu@scut.edu.cn; {jianf, tmei}@microsoft.com

Abstract

Visual sentiment analysis aims to automatically recognize positive and negative emotions from images. There are three main challenges, including large intra-class variance, fine-grained image categories, and scalability. Most existing methods predominantly focus on one or two challenges, which has limited their performance. In this paper, we propose a novel visual sentiment analysis approach with deep coupled adjective and noun neural networks. Specifically, to reduce the large intra-class variance, we first learn a shared middle-level sentiment representation by jointly learning an adjective and a noun deep neural network with weak label supervision. Second, based on the learned sentiment representation, a prediction network is further optimized to deal with the subtle differences which often exist in the fine-grained image categories. The three networks are trained in an end-to-end manner, where the middle-level representations learned in previous two networks can guide the sentiment network to achieve high performance and fast convergence. Third, we generalize the training with mutual supervision between the learned adjective and noun networks by a Rectified Kullback-Leibler loss (*ReKL*), when the adjective and noun labels are not available. Extensive experiments on two widely-used datasets show that our method outperforms the state-of-the-art on SentiBank dataset with 10.2% accuracy gain and surpasses the previous best approach on Twitter dataset with clear margins.

1 Introduction

Recently, understanding the emotion and sentiment from visual content (e.g., image and video) has attracted great attention, since the sentiment conveyed from visual content can explain or even strengthen the sentiment conveyed from text. The capability of automatic visual sentiment analysis will promote visual understanding, and benefit a broad range of applications, such as affective computing [Datta *et al.*, 2008; Ko and Kim, 2015; Siersdorfer *et al.*, 2010], opinion mining [Morency *et al.*, 2011; Yuan *et al.*, 2013], image captioning [Mathews, 2015], etc.

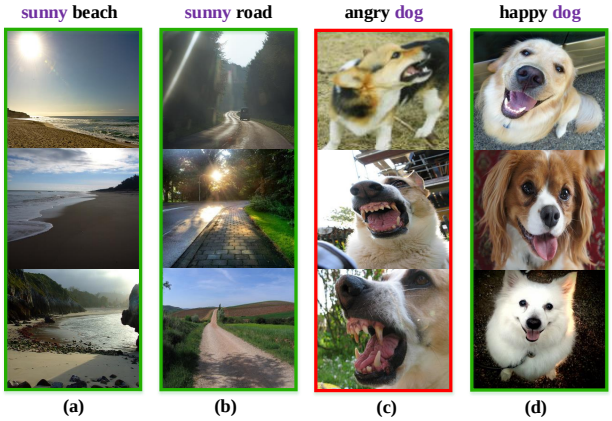

Figure 1: The three challenges for visual sentiment analysis. (1) Large intra-class variance. (a) and (b) depict different objects, but show the same positive sentiment. (2) Fine-grained image categories. Even the images in the same category, e.g., the "dogs" in (c) and (d), can even convey different sentiments. (3) Scalability. The labels, such as "sunny beach," are hard to obtain, which makes sentiment analysis hard to scale up. (Green: positive samples; Red: negative samples)

In this paper we consider visual sentiment analysis as a binary prediction problem which is to classify an image as positive or negative from its visual content [You *et al.*, 2015a; Yuan *et al.*, 2013; Campos *et al.*, 2015; You *et al.*, 2015b]. The difficulty is derived from the "affective gap" between low-level visual content and high-level semantics [Machajdik and Hanbury, 2010]. Significant progresses have been made by designing the sentiment-related visual sentiment ontology, which consists of more than 3,000 adjective noun pairs (denoted as "ANP") with associated training images from search engines [Borth *et al.*, 2013]. They consider each ANP as an affective concept, and thus sentiment prediction can be conducted by determining whether an image can be classified with the affective concepts. For example, images classified with the ANP of "beautiful sky" are positive, while images classified with "terrible accident" are negative.

[*] This work was performed when Jingwen Wang was visiting Microsoft Research as a research intern.

Although ANP-based representation has made sentiment analysis more visually detectable, general object recognition methods can only achieve limited performance. The challenges are three-folds. First, same positive/negative sentiment can be reflected in different objects, which results in large intra-class variance. For example, "beach" and "road" are visually dissimilar. However, images of both "sunny beach" and "sunny road" (Fig. 1 (a)(b)) show the same positive sentiment. Second, different sentiments can be inferred from the same object, and hence visual sentiment analysis needs to detect subtle differences even in the same object class. For example, images of "angry dog" and "happy dog" (Fig. 1 (c)(d)) are all dogs in terms of object class, but are obviously in different sentiment categories. Third, although object-level labels are widely available (e.g., ImageNet [Deng *et al.*, 2009], MS COCO [Chen *et al.*, 2015]), designing ANP and collecting clean ANP samples are expensive. To the best of our knowledge, only VSO [Borth *et al.*, 2013] and MVSO [Jou *et al.*, 2015] dataset provide ANP-level yet noisy training data, which makes sentiment analysis hard to scale up.

Despite a few recent works trying to formulate sentiment analysis as a fine-grained classification problem [Borth *et al.*, 2013; Chen *et al.*, 2014a] to solve the second challenge, few research has studied this topic by considering the above three challenges simultaneously. In this paper, we propose a novel visual sentiment analysis approach with Deep Coupled Adjective and Noun neural networks (DCAN), towards handling the above three challenges in a single framework. First, to overcome the large intra-class variance, we propose to jointly learn deep neural networks for both the adjectives and nouns of ANPs. The goal of designing such a network is to discover the shared features of the same adjective/noun. Second, since the ANP labels are not widely available, we propose to generalize the training with mutual supervision between the adjective and noun networks by a rectified Kullback-Leibler loss ($ReKL$) to other visual sentiment datasets. Third, once the middle-level sentiment representations of the adjectives and nouns are learned, the sentiment prediction network is further trained to detect subtle differences for the same object by weighting the prediction of both networks. Note that the adjective, noun and sentiment network are trained in an end-to-end manner. Moreover, the first two networks are considered as auxiliary classifiers with discount weights in the final loss function, which can guide the sentiment network to achieve high performance and fast convergence. The main contributions of this paper can be summarized as follows:

- We address the challenges of visual sentiment analysis by training a novel coupled deep adjective and noun neural network, which can learn middle-level sentiment representation and reduce intra-class variance.
- We generalize the proposed method to support sentiment analysis without ANP labels by mutual supervision and transfer learning scheme, which makes our method more scalable.
- We conduct experiments on two widely-used sentiment datasets (i.e., SentiBank [Borth *et al.*, 2013], Twitter [You *et al.*, 2015b]), and obtain superior performance over the state-of-the-art with 10.2% accuracy gain.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 introduces our proposed method. Section 4 provides evaluation and analysis, followed by conclusion in Section 5.

2 Related Work

The research on visual sentiment analysis proceeds along two dimensions, i.e., hand-crafted feature-based and deep learning feature-based approaches.

Traditional methods focus on designing hand-crafted features to represent images [Mei *et al.*, 2008; Xu *et al.*, 2009; Fu *et al.*, 2015a; Li *et al.*, 2015]. Previous literatures have been trying to design/apply important sentiment-related features including Wiccest features and Gabor features [Yanulevskaya *et al.*, 2008], global and local RGB histogram [Siersdorfer *et al.*, 2010], SIFT-based bag of features [Siersdorfer *et al.*, 2010; Chen *et al.*, 2014b], Gist features [Yuan *et al.*, 2013; Chen *et al.*, 2014b], and so on. Besides, some works have tried to utilize middle-level features to improve classification results. For example, 102 pre-defined scene attributes from SUN dataset [Xiao *et al.*, 2010] were used as features for sentiment in [Yuan *et al.*, 2013]. In [Borth *et al.*, 2013], 1200-dim features from 1,200 adjective-noun pairs were extracted as middle-level attributes to categorize sentiment based on Plutchiks psychological theorem [Plutchik, 1980].

With the growing of images and videos on the web, traditional methods found it hard to handle the scalability and generalization problem. In contrast, Convolutional Neural Networks (CNNs) [LeCun *et al.*, 1998] are capable of automatically learning robust features from a large number of images [Krizhevsky *et al.*, 2012; Fu *et al.*, 2015b] and videos [Karpathy *et al.*, 2014; Gan *et al.*, 2015], showing significant performances. Motivated by the great success of CNNs, some works have already made attempts to introduce CNNs to visual sentiment classification task [You *et al.*, 2015b][Chen *et al.*, 2014a][Campos *et al.*, 2015].

The most similar work to ours is Bilinear CNN [Lin *et al.*, 2015], which consists of two convolutional feature extractors for fine-grained categorization. Different from this method, our approach can simultaneously solve the above three challenges, i.e., large intra-class variance, fine-grained recognition and low-scalability.

3 Approach

In this section, we introduce the proposed DCAN for visual sentiment analysis. Images with sentiment labels (positive/negative) are first fed into two sub-networks (A-net and N-net) to extract sentiment representation, as shown in Fig. 2(a)-(c). The outputs of both A-net and N-net are further normalized and concatenated in Fig. 2(f), and finally mapped to sentiment in Fig. 2(g). To effectively guide the training process, a weak supervision is used in Fig. 2(d), if noisy ANP labels are available. Otherwise, a mutual supervision with a rectified Kullback-Leibler loss ($ReKL$) is used in Fig. 2(e), which makes the network more scalable.

3.1 Learning Deep Sentiment Representation

Unlike traditional object recognition task, where images of the same object class often share highly-similar visual pat-

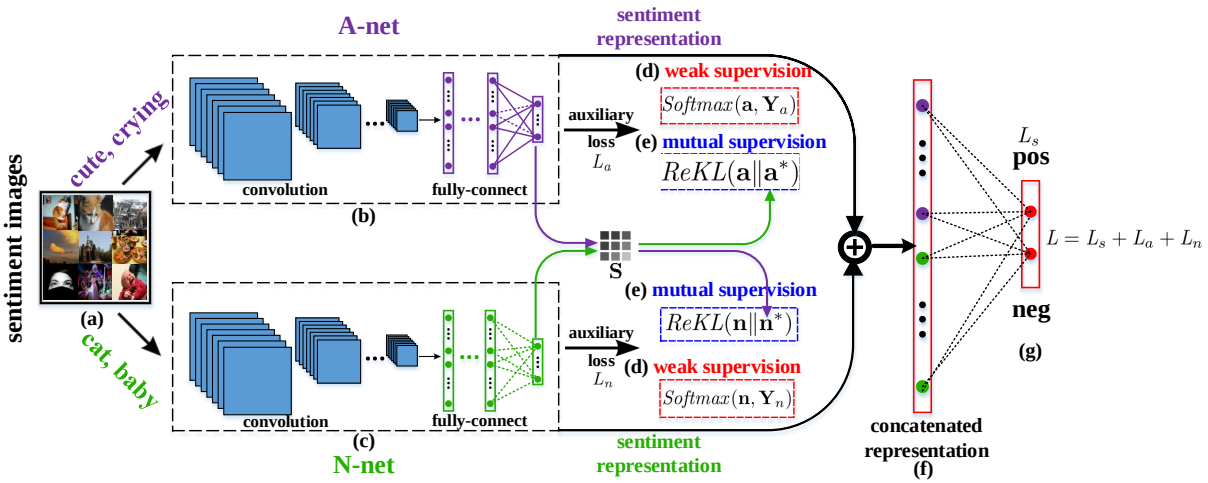

Figure 2: The deep coupled adjective and noun neural network (DCAN). Images in (a) are fed into two sub-nets to extract the descriptiveness features in (b) and the objectiveness features in (c), respectively. The learnt sentiment features are further concatenated in (f) and finally mapped to sentiment in (g). Weak supervision in (d) is applied when noisy labels $\mathbf{Y}_a$ and $\mathbf{Y}_n$ are supplied. Mutual supervision in (e) is applied by predicting an expected output of each sub-network with a transition matrix $\mathbf{S}$, when ANP is unavailable. The network is optimized by minimizing the sentiment loss L_s and the two auxiliary losses L_a, L_n.

terns such as contour and texture, image sentiment analysis are usually involved with great intra-class variance. Given the sentiment supervision (positive/negative), it is hard to learn a mapping function from low-level image pixels to high-level sentiment space, even for the powerful deep learning networks. Previous work has shown that a deeper network even leads to worse result if the appropriate guidance is missing [You *et al.*, 2015b].

To utilize middle-level features to guide the sentiment learning, we propose to learn robust sentiment representation with joint adjective/noun descriptions. Once the adjective/noun descriptions are learned, we consider them as middle-level representation to guide the learning of high-level sentiment. The network structure is shown in Fig. 2. We divide each ANP label into an adjective and a noun, and leverage the two types of labels as weak supervision, since the ANP labels are noisy. Sentiment representation of an image are extracted by two parallel sub-networks (i.e., A-net and N-net) with convolutional and fully-connected layers, under the supervision of adjectives (A) and nouns (N), respectively. The loss of each sub-network is measured by cross entropy:

$$L_{a/n} = L(z, t; \mathbf{I}, \mathbf{W}) = -\log P(z = t|\mathbf{I}, \mathbf{W}), \quad (1)$$

$$P(z = t|\mathbf{I}, \mathbf{W}) = softmax(\mathbf{z}_k) = \frac{\exp(\mathbf{z}_k)}{\sum_{i=1}^{\mathrm{K}} \exp(\mathbf{z}_i)}, \quad (2)$$

where $\mathbf{z}$ is a K-dim network output, $\mathbf{I}$ represents an input image, $\mathbf{W}$ denotes network parameters, z and t is the predicted and true label, respectively.

The advantages of jointly learning adjective network and noun network are two-folds. First, images with the same adjective/noun label usually share similar visual patterns. As shown in Fig. 1, images of both "sunny beach" and "sunny road" show similar visual patterns of "bright light," while the images of both "angry dog" and "happy dog" share common dog appearance. This strategy thus enables the sub-networks to effectively learn common representation for different images under the same adjective or noun. Second, from the perspective of sample distribution, compared to the images assigned by ANP labels, images under the same adjective/noun are much richer and thus benefit for sample expansion and balance. For example, "angry dog" contains less than 100 images in SentiBank [Borth *et al.*, 2013], while there are thousands of samples under "angry" or "dog."

3.2 Transfer Learning with an Rectified Kullback-Leibler Loss

As designing ANP and collecting clean image data with ANP labels are expensive, to further generalize our method to support the task where no ANP labels are provided, we propose a mutual supervision approach by the learnt sentiment representations and ANP transfer. Although the ANP labels are sometimes noisy to image content, this weak supervision can provide us a reliable correlation for constraining the adjective and noun pairs. It would be possible to generate unrelated adjective and noun pairs from A-net and N-net without the ANP supervision. For example, the adjective "cute" is reasonable to describe "dog" or "girl," but inappropriate to describe "bank" or "sky." In order to eliminate those unreasonable results and learn better sentiment representation, we introduce a rectified Kullback-Leibler loss term (denoted as $ReKL$) to mutually supervise the sub-network outputs of the adjective and noun.

Specifically, given the prediction of N-net, we calculate the expected output of A-net by a transition matrix from noun to adjective, and further to minimize the discrepancy between the predicted and expected distribution over different adjectives of the A-net. Since the N-net can supervise the training

of A-net (and vice versa), we call this training procedure as mutual supervision. Note that since the parameters of A-net and N-net can be first pre-trained in weakly-supervised scenarios with a relatively good discrimination ability, the A-net and N-net can be reenforced for each other on other datasets without ANP labels by the proposed mutual supervision. Formally, given the output of A-net, a rectified Kullback-Leibler loss ($ReKL$) of N-net is defined as:

$$\begin{aligned} ReKL(\mathbf{n}\|\mathbf{n}^*) &= \sum\nolimits_k \mathbf{n}_k \max\left(\log\frac{\mathbf{n}_k}{\mathbf{n}_k^*}, 0\right) \\ &= \sum\nolimits_k \mathbf{n}_k \max\left(\log\frac{\mathbf{n}_k}{\sum_j \mathbf{a}_j S_{jk}}, 0\right), \end{aligned} \quad (3)$$

where $\mathbf{n}$ is the predicted output of the N-net, and $\mathbf{n}^*$ is the expected output given the A-net output. k is the k^{th} dimension of the N-net output. S_{jk} is an element of the transition matrix $\mathbf{S}$, which has K_A columns and K_N rows. K_A and K_N indicates the number of adjectives and nouns, respectively. S_{jk} can be obtained by calculating the co-occurence times of the j^{th} adjective and the k^{th} noun from ANPs. The matrix $\mathbf{S}$ is further normalized by rows. Note that the term in the sum function is ignored if ($\mathbf{n}_k \leq \mathbf{n}_k^*$). The reason is that if $\mathbf{n}_k$ is larger enough compared to $\mathbf{n}_k^*$, the output probability of noun is irrational compared with the expectation and should be penalized. Since the expected output gives all the "possible" choices of the network output, we should avoid the penalty if ($\mathbf{n}_k \leq \mathbf{n}_k^*$). Similarly, given the output of the N-net, the loss of A-net is shown as:

$$ReKL(\mathbf{a}\|\mathbf{a}^*) = \sum\nolimits_k \mathbf{a}_k \max(\log\frac{\mathbf{a}_k}{\sum_j \mathbf{n}_j S_{jk}^T}, 0), \quad (4)$$

where $\mathbf{S}^T$ is obtained by transposing the matrix $\mathbf{S}$ and normalizing by columns. The derivatives of the proposed $ReKL$ loss are calculated as follows:

$$\frac{\partial ReKL(\mathbf{n}\|\mathbf{n}^*)}{\partial \mathbf{n}_k} = \begin{cases} \log\frac{\mathbf{n}_k}{\sum_j \mathbf{a}_j S_{jk}} + 1, & \mathbf{n}^* \leq \mathbf{n}_k, \\ 0, & otherwise, \end{cases} \quad (5)$$

$$\frac{\partial ReKL(\mathbf{a}\|\mathbf{a}^*)}{\partial \mathbf{a}_k} = \begin{cases} \log\frac{\mathbf{a}_k}{\sum_j \mathbf{n}_j S_{jk}^T} + 1, & \mathbf{a}^* \leq \mathbf{a}_k, \\ 0, & otherwise. \end{cases} \quad (6)$$

3.3 Sentiment Analysis

Since the visual sentiment analysis also needs to solve the fine-grained challenge, the learned representation of A-net and N-net are further concatenated and connected to an additional fully-connected layers with a softmax function for binary prediction. Adjectives are usually related to descriptiveness, while nouns represent objectiveness of an image. Therefore, combining the two kinds of representation could be reasonable and easy to learn a fine-grained and discriminative sentiment predictor as:

$$P_s = softmax(\mathbf{W}^{(M+1)} \cdot [\mathbf{a}; \mathbf{n}]), \quad (7)$$

$$L_s = -\log P_s, \quad (8)$$

where $\mathbf{W}^{(M+1)}$ is the parameter of the last fully-connected layer, L_s is the loss. To end-to-end train the whole neural network, we integrate the loss of A-net, N-net and sentiment by a linear combination in two forms:

Algorithm 1 Training the end-to-end DCAN

Input: Training images: $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$, sentiment labels $\mathbf{Y}_s$, noisy adjective and noun labels $\mathbf{Y}_a$, $\mathbf{Y}_n$, initial parameters $\mathbf{W} = [\mathbf{W}^{(1)}, ..., \mathbf{W}^{(M+1)}]$ ($M+1$ denotes the last fully-connected layer), rectified Linear Activation Function $f(\cdot)$.
Procedure:
Repeat:
Forward Propagation:
Last layer output $\mathbf{z}_s^{(M+1)} = \mathbf{W}^{(M+1)}[\mathbf{z}_a^{(M)}, \mathbf{z}_n^{(M)}]$
Other layers are similar as traditional CNN
Backward Propagation:
1. For $m = M+1$, calculate
$\frac{\partial L}{\partial \mathbf{W}^{(M+1)}} = \frac{\partial L_s}{\partial \mathbf{W}^{(M+1)}}$, $\delta_s^{(M+1)} = -w_s \frac{\partial L_s}{\partial \mathbf{z}_s^{(M+1)}}$
2. For $m = M$, calculate
$\frac{\partial L}{\partial \mathbf{W}^{(M)}} = \delta_s^{(M+1)}(f(\mathbf{z}_s^{(M+1)}))^T$
$\delta^{(M)} = [(\mathbf{W}^{(M)})^T \delta_s^{(M+1)}] \cdot f'(\mathbf{z}^{(M)})$
For **weak** supervision:
$\delta_a^{(M)} = \delta^{(M)} - w_a \frac{\partial L_a}{\partial \mathbf{z}_a^{(M)}}$, $\delta_n^{(M)} = \delta^{(M)} - w_n \frac{\partial L_n}{\partial \mathbf{z}_n^{(M)}}$
For **mutual** supervision:
$\delta_a^{(M)} = \delta^{(M)} - w_a \frac{\partial ReKL(\mathbf{a}\|\mathbf{a}^*)}{\partial \mathbf{z}_a^{(M)}}$,
$\delta_n^{(M)} = \delta^{(M)} - w_n \frac{\partial ReKL(\mathbf{n}\|\mathbf{n}^*)}{\partial \mathbf{z}_n^{(M)}}$
3. For $m = M-1$ to $m = 2$, calculate
$\frac{\partial L}{\partial \mathbf{W}^{(m)}} = \delta^{(m+1)}(f(\mathbf{z}^{(m)}))^T$
$\delta^{(m)} = [(\mathbf{W}^{(m)})^T \delta^{(m+1)}] \cdot f'(\mathbf{z}^{(m)})$
Until the max iteration
Output: $\mathbf{W} = [\mathbf{W}^{(1)}, ..., \mathbf{W}^{(M+1)}]$

- Weak Supervision

$$L = w_s L_s + w_a L_a + w_n L_n, \quad (9)$$

- Mutual Supervision

$$L = w_s L_s + w_a ReKL(\mathbf{a}\|\mathbf{a}^*) + w_n ReKL(\mathbf{n}\|\mathbf{n}^*), \quad (10)$$

where w_s, w_a, w_n are weights. We generally set w_s to be relatively larger for faster convergence and better result. In optimization, we adopt stochastic gradient descent method to train our model. Algorithm 1 summarizes the training procedure of the proposed DCAN.

3.4 Discussions

One of the challenges that hinders neural networks to go deeper is the notorious "vanishing gradient" problem. In the 22-layer GoogLeNet [Szegedy *et al.*, 2014], additional auxiliary classifiers are introduced at intermediate layers to relieve the vanishing gradient problem and improve the performance. In Deeply-Supervised Nets [Chen-Yu *et al.*, 2014], supervision on hidden neurons via companion objectives shows superior performance on ImageNet challenge. In our model, the adjective and noun classifier can also be viewed as auxiliary supervision on middle-level layers, which is expected to improve the accuracy of sentiment prediction. Different from GoogLeNet and Deeply-Supervised Nets, the learnt adjective and noun features have specific meaning and are further integrated into the sentiment analysis.

4 Experiments

We evaluate the proposed DCAN on two datasets and compare the performance of DCAN with the state-of-the-art.

Table 1: Results of CNNs with different depths on SentiBank.

Network Structure	Accuracy
6-layer CNN (2Conv+4FC)	0.718
8-layer CNN (AlexNet)	0.649
9-layer CNN (5Conv+4FC)	0.640

Table 2: The precision, recall, F1 and accuracy of different approaches on SentiBank testing set.

Methods	Prec.	Rec.	F1	Acc.
2Conv+4FC	0.714	0.729	0.722	0.718
PCNN	0.759	0.826	0.791	0.781
Bilinear CNN	0.669	0.666	0.668	0.667
DCAN (2Conv+4FC)	0.755	0.719	0.737	0.742
DCAN(ANP)	0.843	0.843	0.843	0.843
DCAN (Alex)	0.858	0.889	0.873	0.870
DCAN (share)	0.857	0.893	0.875	0.872
DCAN (pre-train)	**0.865**	**0.908**	**0.886**	**0.883**

4.1 Datasets

SentiBank [Borth *et al.*, 2013]. SentiBank is widely-used, which contains about one-half million images from Flickr with designed ANPs as queries. The sentiment label of each image is decided by sentiment polarity of the corresponding ANP. We use this dataset for weak supervision, as noisy ANP labels are provided. The training/testing split is 90% and 10%, respectively.

Twitter "five-agree" [You *et al.*, 2015b]. The dataset is more challenging than SentiBank, which contains 581 positive samples and 301 negative samples. Each sample is labeled by at least 5 AMT workers. The training/testing split is 80% and 20%, respectively. The network is pre-trained on SentiBank, and fine-tuned on Twitter. We use the mutual supervision as ANP labels are unavailable.

4.2 Compared Methods

We compare the proposed DCAN network with 8 baselines:

- **GCH/LCH/GCH+BoW/LCH+BoW:** 64-bin global color histogram features (GCH), 64-bin local color histogram (LCH) and SIFT-based Bag-of-words features, as defined in [Siersdorfer *et al.*, 2010].
- **SentiBank:** Using 1200-dim ANP representation with Linear SVM as classifier [Borth *et al.*, 2013].
- **Sentribute:** Using middle-level representation with 102 pre-defined scene attributes [Yuan *et al.*, 2013].
- **2Conv+4FC:** A CNN method with 2 convolutional layers and 4 fully-connected layers, as in [You *et al.*, 2015b].
- **PCNN:** Progressive CNN [You *et al.*, 2015b].
- **DeepSentiBank:** 2089-dim CNN features with linear SVM as classifier [Chen *et al.*, 2014a].
- **Fine-tuned CaffeNet:** An ImageNet pre-trained AlexNet [Campos *et al.*, 2015] followed by fine-tune.
- **Bilinear CNN:** A state-of-the-art fine-grained classification method [Lin *et al.*, 2015]. To be comparable, we use B-CNN[M,M]. The top 5 convolutional layers are pre-trained from ImageNet.

Table 3: Comparison results on Twitter "five-agree."

Methods	Accuracy
GCH	0.684
LCH	0.710
GCH+BoW	0.710
LCH+BoW	0.717
SentiBank	0.709
Sentribute	0.738
DeepSentiBank	0.774
2Conv+4FC	0.783
PCNN	0.773
Fine-tuned CaffeNet	0.830
DCAN (Alex)	0.823
DCAN (Alex)+ReKL	**0.838**

We also compare different variants of our DCAN:

- **DCAN (2Conv+4FC):** Using network structure from [You *et al.*, 2015b] as A-net and N-net.
- **DCAN (Alex):** Using AlexNet structure [Krizhevsky *et al.*, 2012] as A-net and N-net.
- **DCAN (ANP):** Similar as DCAN (Alex), except using each ANP as middle representation.
- **DCAN (share):** Similar as DCAN (Alex), except A-net and N-net sharing common convolutional layers.
- **DCAN (pre-train):** Similar as DCAN (Alex), except training sub-networks first before training the whole.

4.3 Evaluation of Network Depth

We first test the accuracy of neural networks by adding more layers. The result is shown in Tab. 1, which verifies the same observation in [You *et al.*, 2015b] that simply going deep can not help bridge the affective gap between low-level image pixels and high-level sentiment. It is important to learn effective middle-level features for visual sentiment analysis.

4.4 Experiments on SentiBank

In SentiBank, we separate the ANPs into about 180 adjectives and 300 nouns. The input images are first resized to 227×227 and mean-subtracted before propagating to the network. We train our model using a mini-batch of 256 and weight decay of 0.0005, as suggested by [Krizhevsky *et al.*, 2012]. The initial learning rate is 0.005 and divided by 10 every 75 epochs until convergence. We empirically set the weight of sentiment to 2 and the weight of sub networks to be 1 since this weighting gives the best result.

Tab. 2 summarizes the performances of our approach and other methods. Compared to the others, DCAN (Alex) gives the result of 87.0%, which is much better than PCNN [You *et al.*, 2015b]. DCAN (pre-train) further leads to the best performance, with over 10.0% accuracy gain compared to PCNN [You *et al.*, 2015b]. The performance of DCAN (2Conv+4FC) is unsatisfactory, because the two sub-networks have few convolutions and are inadequate to capture the adjective and noun features. However, the accuracy of DCAN (2Conv+4FC) is still higher than the pure CNN of 2Conv+4FC (74.2% *vs* 71.8%), which shows the effectiveness for building the parallel network. Similar observation has been

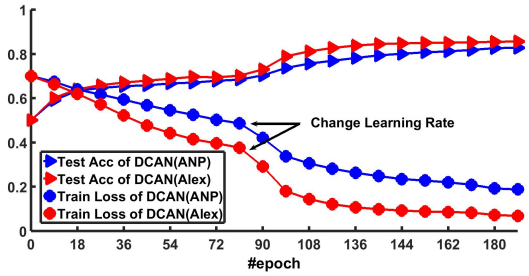

Figure 3: Accuracy (crop-center) and convergence of D-CAN(Alex) and DCAN (ANP).

found in Bilinear CNN [Lin *et al.*, 2015], which shows better results than the pure CNN (AlexNet). Our network D-CAN (Alex) shows better performance than Bilinear CNN with same number of convolutional layers, which verifies that recognizing sentiment cannot be simply formulated as a fine-grained problem. Instead, we should simultaneously consider the three challenges, i.e., large intra-class variance, fine-grained recognition and low-scalability. Besides, the proposed network achieves superior performance over D-CAN (ANP), which suggests that learning shared features for each adjective or noun can greatly promote the performances of neural network. We further visualize DCAN(Alex) and D-CAN (ANP) by comparing their testing accuracy and training loss through training stage. As shown in Fig. 3, DCAN(Alex) achieves better performance and faster convergence.

We are also interested in images with high response to some adjectives such as "crowded," "delicious," etc. We first divide an image into blocks, and feed them into the A-net to obtain a response score. We further visualize the response score by a heat map. The bright area indicates high response. As shown in Fig. 4, the proposed network is capable of learning common descriptive features among different objects. For example, we can capture the group of people as a visual clue to represent "crowded" in "crowded city," "crowded beach," etc, and also damaged windows/doors as a sign for "abandoned" in "abandoned building," "abandoned industry," and "abandoned factory."

4.5 Experiments on Twitter "five-agree"

On Twitter, we employ a 5-fold cross validation to test our model as the same as other baseline methods. We first separate the data into five partitions and in each time we use four partitions for fine-tuning and the rest one partition for testing. We initialize the model with learnt weights and ANP transition matrix from SentiBank, then fine-tune the whole model with a relatively small learning rate. We also implement our rectified KL to guide the learning of the affective concepts.

We report mean performance on Twitter "five-agree" in Tab. 3. As can be seen from the table, deep learning based approaches generally have better ability to describe and model sentiment information, compared to traditional features based methods in the first six rows. DCAN (Alex) achieves 5.0% accuracy gain compared to PCNN [You *et al.*, 2015b]. By using the proposed $ReKL$ loss, we can further improve the

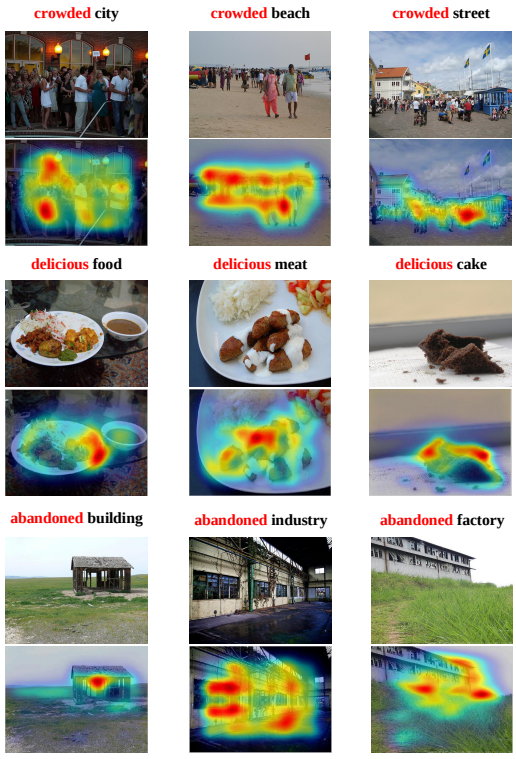

Figure 4: Region-based heat maps for different ANPs by using the prediction of A-net. The bright areas indicate high-response regions to the adjectives.

performance with clear margins, which supports the fact that the mutual supervision can help regularize the outputs of the two sub-networks and thus produces more accurate features. The method of fine-tuned CaffeNet also obtains good result. However, the method utilizes ImageNet dataset [Deng *et al.*, 2009] for pre-training, which contains over one million manually labeled images. In contrast, our approach can generate the best result with only weakly-labeled data.

5 Conclusion

In this paper, we propose DCAN which is a novel CNN structure with deep coupled adjective and noun networks for visual sentiment analysis. Our network can effectively learn middle-level sentiment features from noisy web images with ANP labels, and achieve the best result on both SentiBank and Twitter dataset to the best of our knowledge. Since the ANP labels are human-designed, we will focus on automatically discovering robust middle-level representation to guide the learning of sentiment in our future work.

Acknowledgments

This research was supported by National Nature Science Foundations of China (61273225, 61070091 and 61528204), Project of High Level Talents in Higher Institution of Guangdong Province (2013-2050205-47).

Reference

[Borth *et al.*, 2013] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia*, pages 223–232, 2013.

[Campos *et al.*, 2015] Victor Campos, Amaia Salvador, Xavier Giro-i Nieto, and Brendan Jou. Diving Deep into Sentiment: Understanding Fine-tuned CNNs for Visual Sentiment Prediction. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 57–62, 2015.

[Chen *et al.*, 2014a] 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.

[Chen *et al.*, 2014b] 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 *ACM Multimedia*, pages 367–376, 2014.

[Chen *et al.*, 2015] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[Chen-Yu *et al.*, 2014] L Chen-Yu, X Saining, G Patrick, Z Zhengyou, and T Zhuowen. Deeply-supervised nets. *CoRR, abs/1409.5185*, 3(4):93, 2014.

[Datta *et al.*, 2008] Ritendra Datta, Jia Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, pages 105–108, 2008.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Fu *et al.*, 2015a] Jianlong Fu, Jinqiao Wang, Yong Rui, Xin-Jing Wang, Tao Mei, and Hanqing Lu. Image tag refinement with view-dependent concept representations. *TCSVT*, pages 1409–1422, 2015.

[Fu *et al.*, 2015b] Jianlong Fu, Yue Wu, Tao Mei, Jinqiao Wang, Hanqing Lu, and Yong Rui. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *ICCV*, pages 1985–1993, 2015.

[Gan *et al.*, 2015] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alexander G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015.

[Jou *et al.*, 2015] 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 *ACM Multimedia*, pages 159–168, 2015.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[Ko and Kim, 2015] Eunjeong Ko and Eun Yi Kim. Recognizing the sentiments of web images using hand-designed features. In *ICCI & CC*, pages 156–161, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2015] Changsheng Li, Qingshan Liu, Jing Liu, and Hanqing Lu. Ordinal distance metric learning for image ranking. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1551–1559, 2015.

[Lin *et al.*, 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN Models for Fine-grained Visual Recognition. *arXiv preprint arXiv:1504.07889*, 2015.

[Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM Multimedia*, pages 83–92, 2010.

[Mathews, 2015] Alexander Patrick Mathews. Captioning Images Using Different Styles. In *ACM Multimedia*, pages 665–668, 2015.

[Mei *et al.*, 2008] Tao Mei, Yong Wang, Xian-Sheng Hua, Shaogang Gong, and Shipeng Li. Coherent image annotation by learning semantic distance. In *CVPR*, pages 1–8, 2008.

[Morency *et al.*, 2011] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 169–176, 2011.

[Plutchik, 1980] Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[Siersdorfer *et al.*, 2010] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *ACM Multimedia*, pages 715–718, 2010.

[Szegedy *et al.*, 2014] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[Xu *et al.*, 2009] Yong Xu, Hui Ji, and Cornelia Fermüller. Viewpoint invariant texture description using fractal analysis. *IJCV*, pages 85–100, 2009.

[Yanulevskaya *et al.*, 2008] Victoria Yanulevskaya, JC Van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *ICIP*, pages 101–104, 2008.

[You *et al.*, 2015a] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks. In *ACM Multimedia*, pages 1071–1074, 2015.

[You *et al.*, 2015b] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.

[Yuan *et al.*, 2013] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10, 2013.