Weakly Supervised Coupled Networks for Visual Sentiment Analysis

Jufeng Yang[†], Dongyu She[†], Yu-Kun Lai[‡], Paul L. Rosin[‡], Ming-Hsuan Yang[§]
[†]College of Computer and Control Engineering, Nankai University, Tianjin, China
[‡]School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[§] School of Engineering, University of California, Merced, USA

Abstract

Automatic assessment of sentiment from visual content has gained considerable attention with the increasing tendency of expressing opinions on-line. In this paper, we solve the problem of visual sentiment analysis using the high-level abstraction in the recognition process. Existing methods based on convolutional neural networks learn sentiment representations from the holistic image appearance. However, different image regions can have a different influence on the intended expression. This paper presents a weakly supervised coupled convolutional network with two branches to leverage the localized information. The first branch detects a sentiment specific soft map by training a fully convolutional network with the cross spatial pooling strategy, which only requires image-level labels, thereby significantly reducing the annotation burden. The second branch utilizes both the holistic and localized information by coupling the sentiment map with deep features for robust classification. We integrate the sentiment detection and classification branches into a unified deep framework and optimize the network in an end-to-end manner. Extensive experiments on six benchmark datasets demonstrate that the proposed method performs favorably against the state-of-the-art methods for visual sentiment analysis.

1. Introduction

Visual sentiment analysis from images has attracted significant attention with the increasing tendency of expressing opinions through posting images on social media like Flickr and Twitter. The automatic assessment of image sentiment has many applications, *e.g*. education, entertainment, advertisement, *etc*. Recently, with the advances of convolutional neural networks (CNNs), numerous deep approaches have been proposed to predict sentiment [20,31]. The effectiveness of machine learning based deep features have been demonstrated over hand-crafted features (*e.g*. color, texture, and composition) [17, 28, 34]) on visual sentiment prediction. However, several issues remain when using CNNs to address such an abstract task as follows.

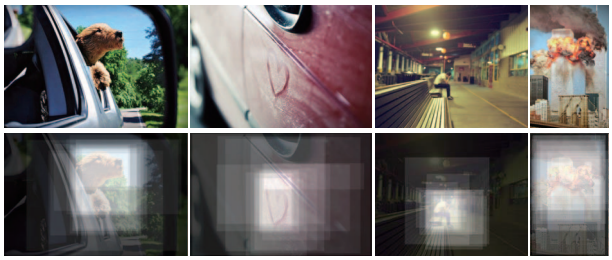

Figure 1. Examples from the EmotionROI dataset [21]. The normalized bounding boxes indicate the regions that influence the evoked sentiments annotated by 15 users. The first two examples are joy images, and the last two examples are sadness and fear images, respectively. As can be seen, the sentiments can be evoked by specific regions.

First, visual sentiment analysis is more challenging than conventional recognition tasks due to a higher level of subjectivity in the human recognition process [13]. It is necessary to take more cues into consideration for visual sentiment prediction. Figure 1 shows examples from the EmotionROI dataset [21], which provides the bounding box annotations that invoke sentiment from 15 users. As can be seen, humans' emotional responses to images are determined by local regions [29]. However, most existing methods employ CNNs to learn feature representations only from entire images [4, 30]. Second, providing more precise annotations (*e.g*. bounding boxes [11]) than image-level labeling for training generally leads to better performance for recognition tasks. However, there are two limitations for visual sentiment classification. On the one hand, the increased annotation cost prevents it from widespread use, especially for such a subjective task; on the other hand, different regions contribute differently to the viewer's evoked sentiment, while crisp proposal boxes only tend to find the foreground objects in an image.

To address these problems, we propose a weakly supervised coupled framework (WSCNet) for joint sentiment detection and classification with two branches. The first

branch is designed to generate region proposals evoking sentiment. Instead of extracting multiple crisp proposal boxes, we use a soft sentiment map to represent the probability of evoking the sentiment for each receptive field. In detail, we make use of a Fully Convolutional Network (FCN) followed by the proposed cross-spatial pooling strategy to preserve the spatial information of the convolutional feature maps. Based on this, the sentiment map is generated and utilized to highlight the regions of interest that are informative for classification. The second branch captures the localized representation by coupling the sentiment map with the deep features, which is then combined with the holistic representation to provide a more semantic vector. During the end-to-end training process, our approach only requires image-level sentiment labeling, which significantly reduces the annotation burden.

Our contributions are summarized as follows: First, we present a weakly supervised coupled network to integrate visual sentiment classification and detection into a unified CNN framework, which learns the discriminative representation for visual sentiment analysis in an end-to-end manner. Second, we exploit the sentiment map to provide image-specific localized information with only the image-level label, with which both holistic and localized representations are fused for robust sentiment classification. Our proposed framework performs favorably against the state-of-the-art methods and off-the-shelf CNN classifiers on six benchmark datasets for visual sentiment analysis.

2. Related Work

In this section, we review methods for image sentiment prediction [27, 33] and weakly supervised detection [37] that are closely related to our work.

2.1. Visual Sentiment Prediction

Most existing approaches to visual sentiment prediction are developed based on hand-engineered features [28] and deep learning frameworks [24]. In the early years, numerous methods have been used to design different groups of hand-crafted features inspired by psychology theory and principles of art. Machajdik *et al*. [17] define a combination of low-level features that represent the emotional content, *e.g*. color, texture, composition, while more robust features according to art principles are investigated in [34]. Zhao *et al*. [35, 36] further propose the multi-task hypergraph learning to predict personalized emotion perceptions and release the IESN dataset, which is the pioneering work towards the emotion subjectivity challenge. Different factors that may influence emotion perceptions are jointly considered, *i.e*. visual content, social context, temporal evolution and location influence. More recently, several approaches exploit deep models for learning sentiment representations. The DeepSentiBank [7] constructs a visual sentiment concept in terms of classification on adjective-noun pairs (ANP) for detecting sentiment depicted in images. Due to the expensive manual annotation of sentiment labels, the existing affective datasets mostly contain less than one thousand images as summarized in Table 1. To cope with limited training data, most approaches incorporate the CNN weights learned from a large-scale general dataset [9] and fine-tune the model for sentiment prediction [4, 5, 30]. To utilize sentiment ambiguity, Yang *et al*. [25] propose to learn a deep representation in a multi-task CNN, which jointly optimizes the classification and distribution learning.

Table 1. Statistics of the available affective datasets. Most datasets developed in this field contain no more than one thousand samples, mainly due to the subjective and labor intensive labeling process. As the last column shows, none of these datasets except EmotionROI provide ground truth regions that evoke sentiments.

Dataset	#Images	#Classes	Regions
IAPSa [17]	395	8	N
Abstract [17]	228	8	N
ArtPhoto [17]	806	8	N
Twitter I [30]	1,269	2	N
Twitter II [3]	603	2	N
EmotionROI [21]	1,980	6	Y
Flickr&Instagram [31]	23,308	8	N
Flickr [14]	60,745	2	N
Instagram [14]	42,856	2	N

While most CNN-based methods for sentiment classification extract deep features from the entire image, significantly less attention has been paid to utilize the local regions information for sentiment prediction. Li *et al*. [16] propose a context-aware classification model taking both the global and global-local context into account. Sun *et al*. [23, 26] discover affective regions based on an object proposal algorithm and combine deep features for classification. However, such methods are sub-optimal since the objectness algorithm is separate from the prediction method, and regions that are not object-like may be excluded at the very beginning. In [29], a method based on an attention model is developed in which local visual regions induced by sentiment related visual attributes are considered. Different from existing methods in the literature, we propose a joint model that trains two tasks simultaneously in an end-to-end network. We show that the proposed framework is able to learn a discriminative sentiment representation, and performs favorably against the state-of-the-art methods for visual sentiment analysis.

2.2. Weakly Supervised Detection

With the recent success of deep learning on large-scale object recognition [15], several weakly supervised CNNs have been proposed for the object detection task using mul-

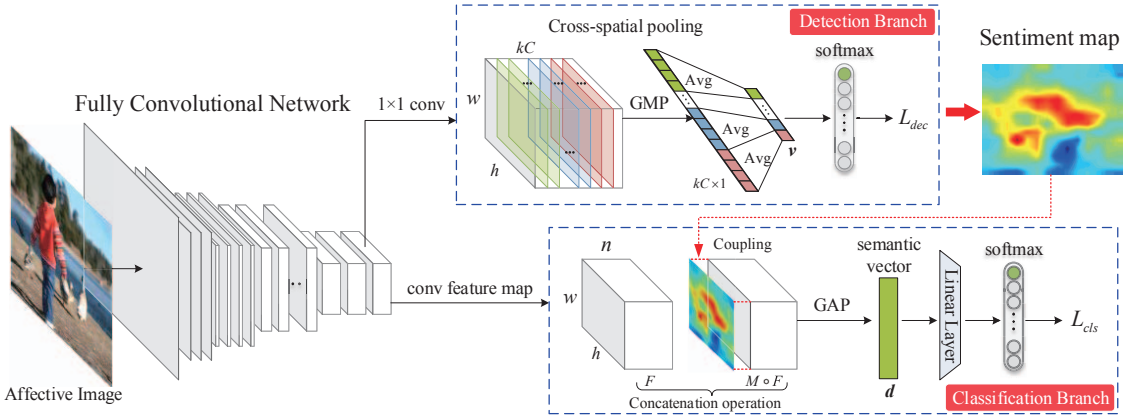

Figure 2. Illustration of the proposed WSCNet for visual sentiment analysis. The input image is first fed into the convolutional layers of FCN ResNet-101, and the response feature maps with good spatial resolution are then delivered into two branches. The detection branch employs the cross-spatial pooling strategy to summarize all the information contained in the feature maps for each class. The end-to-end training results in the sentiment map, which is then coupled with the conv feature maps in the classification branch capturing the localized information. Finally, both holistic and localized representations are fused as a semantic vector for sentiment classification.

tiple instance learning (MIL) algorithms [2]. MIL defines images as a bag of regions, and assumes that images labeled as positive contain at least one object instance of a certain category and images labeled as negative do not contain an object from the category of interest. One of the most common approaches [8] consists of generating object proposals and extracting features from the proposals in multiple stages, and employs MIL on the features to determine the box labels from the weak bag labels. However, since sentiment is more subjective, assuming that an instance only appears in a single category is suboptimal for sentiment detection. In addition, methods have also been proposed to use a unified network framework to perform both localization and classification, which takes the convolutional filters as detectors to activate locations on the deep feature maps [10, 37, 38]. Zhou *et al*. [37] utilize the global average pooling layer behind the top convolutional layer to aggregate class-specific activation, while Durand *et al*. [10] propose the WILDCAT method to learn multiple localized features related to different class modalities (*e.g*. object parts). Considering the object evidence, Zhu *et al*. [38] propose the soft proposal network (SPN) to generate soft proposals and aggregate image-specific patterns by coupling the proposal and feature maps, which tends to distinguish the foreground objects from the surroundings with a graph propagation algorithm.

To our knowledge, there is little work focusing on sentiment detection. Peng *et al*. [21] train a supervised network FCNEL to predict the emotion stimuli map (ESM) with manually labeled pixel-level ground truth, which would be extremely labor intensive if it were extended to large-scale datasets. We are the first to integrate sentiment-related proposals into CNNs for detection, and to jointly optimize the detection and classification tasks under weak supervision. Different from the existing weakly supervised methods, this work proposes to detect a unified sentiment map considering both the salient foreground as well as the sentiment-related areas, instead of using class-specific activation [10, 37] for each category. Moreover, the detected regions are utilized as localized information to boost the sentiment classification.

3. Weakly Supervised Coupled Network

Our weakly supervised coupled network is illustrated in Figure 2. The goal is to learn a discriminative model from images with regions that evoke sentiment where the only manual supervision required is image-level labels. Specifically, the proposed WSCNet learns both detection and classification tasks jointly with two network branches. We use the detection branch to generate a sentiment map providing the localized information, which is then fed into the classification branch to fuse the holistic as well as the localized representations together.

3.1. Sentiment Map Detection Branch

While attention and salience works aim to find salient objects in images, a sentiment image is defined as a person's disposition to respond to visual inputs, which may contain not only salient objects but other areas related to emotion [21]. As stated in Section 2.2, there are only a few end-to-end CNN frameworks for weakly supervised object detection that do not use additional localization informa-

tion. In order to infer the sentiment map directly in the CNN, the convolutional filters are viewed as the detector that produces the feature maps as the response. Our framework is based on the recently introduced FCN ResNet-101 [12] that naturally preserves spatial information throughout the network.

Cross-spatial pooling strategy. Let $\{(x_i, y_i)\}_{i=1}^{N}$ be a collection of N affective training examples, where x_i is an affective image, and $y_i \in \{1, \cdots, C\}$ is the corresponding sentiment label. For each instance, let $F \in \mathbb{R}^{w \times h \times n}$ be the feature maps of the conv5 in ResNet-101, where w and h are the spatial size (width and height) of the feature maps, respectively, and n is the number of channels. We first add a 1×1 convolutional layer to capture multiple information (*e.g.* views) for each sentiment category, which has high response to certain discriminative regions. Suppose k detectors are applied to each sentiment class, we obtain feature maps F' with the dimension of $w \times h \times kC$. We propose to summarize all the information as a single image-level score for each of the sentiment classes independently regardless of the input size, which is achieved by the cross-spatial pooling strategy:

$$v_c = \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{G}_{max}(f_{c,i}), c \in \{1, \cdots, C\}, \tag{1}$$

where $f_{c,i}$ represents the i-th feature map for the c-th label from F', and $\boldsymbol{G}_{max}(\cdot)$ denotes the Global Max Pooling (GMP). Here, GMP is employed to identify just one discriminative part for each feature map in the same sentiment class inspired by [37], which results in a $1 \times 1 \times kC$ vector. Then k responses for each label are unified with the average pooling operation, where the value can be maximized by finding all discriminative regions of the specific sentiment, as all low activations reduce the output of the particular map. The pooled vector $\mathbf{v} \in \mathbb{R}^C$ is then fed into a C-class softmax layer as the sentiment detection loss:

$$L_{dec} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbf{1}(y_i = c) \log v_c, \tag{2}$$

where $\mathbf{1}(s) = 1$ if the condition s is true, and 0 otherwise. Thus, the filter weights can be updated during the training process, which yields the discriminative location in the feature maps for each class. We use the cross-spatial pooling strategy to represent the GMP layer followed by a class-specific average pooling as a convenient term.

Generating Sentiment Map. Different from object locations [18] or 'class activation' maps [37], the activation feature maps for different sentiments are dependent due to the ambiguity existing in the sentiment labels [25]. Thus, this paper proposes to capture the regions evoking sentiment by considering all the class activation maps.

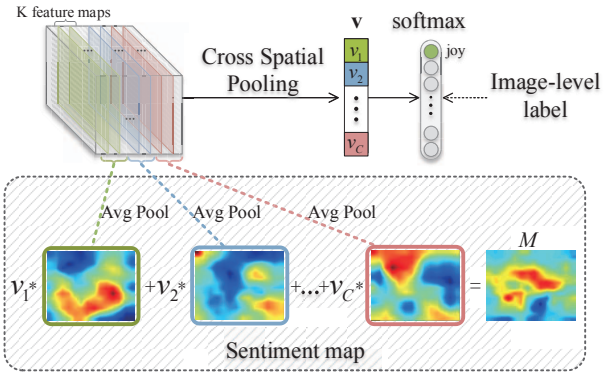

Figure 3. Overview of the sentiment map generation. The predicted class scores of the input image are mapped back to the classification branch to generate the sentiment map, which can highlight comprehensive sentiment regions.

We first obtain a single map from the k feature maps for each sentiment by the average pooling operation. We then combine all the C class-wise feature maps with corresponding weights to capture the comprehensive localized information, instead of using the feature maps with the largest response from a specific class. Thus, our sentiment map $M \in \mathbb{R}^{w \times h}$ is generated using v_c as the weight of the response map of class c:

$$M = \sum_{c=1}^{C} v_c \left(\frac{1}{k} \sum_{i=1}^{k} f_{c,i} \right). \tag{3}$$

Intuitively, based on prior methods [32], we expect that each unit is activated by some visual patterns within its receptive field. The sentiment map is a weighted linear sum of the presence of these visual patterns at different spatial locations. By simply up-sampling the activation map to the size of the input image, we can identify the regions most relevant to the evoked sentiment, as shown in Figure 3.

3.2. Coupled Sentiment Classification Branch

From the perspective of image representation, the sentiment map highlights the image-specific discriminative regions that are informative for image classification. The original convolutional feature F is viewed as the holistic representation, and the sentiment map is utilized to produce the local representation by coupling with the convolutional features. Inspired by [38], the Hadamard product is employed to couple each feature map from F with M. Thus, we obtain the coupled feature maps $U = [U_1, U_2, \cdots, U_n]$, where the element $U_i = M \circ F_i$, and $\circ$ denotes the element-wise multiplication. Then the coupled feature maps and the original feature maps can be encoded to form a more informative semantic feature $\mathbf{d} \in \mathbb{R}^{2n}$ by:

$$\mathbf{d} = \boldsymbol{G}_{avg}(F \uplus U), \tag{4}$$

where $\uplus$ denotes the concatenation of different convolutional features. In the above equation, $\boldsymbol{G}_{avg}(\cdot)$ is the global average pooling (GAP) operation, which outputs the average value of each feature map.

To classify an image, it is necessary to compute the predicted scores of the input image for different classes. We use those as features for a fully-connected layer and the sentiment scores $s(y_i = c|\mathbf{d}, \mathbf{w}_c)$ is defined as

$$s(y_i = c|\mathbf{d}, \mathbf{w}_c) = \frac{\exp(\mathbf{w}_c^\top \mathbf{d})}{\sum_{c'=1}^{C} \exp(\mathbf{w}_{c'}^\top \mathbf{d})}, \quad (5)$$

where $\mathbf{W} = \{\mathbf{w}_c\}_{c=1}^{C}$ is the set of model parameters. Thus, the classification is carried out by minimizing the following log likelihood function:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbf{1}(y_i = c) \log s(y_i = c|\mathbf{d}, \mathbf{w}_c). \quad (6)$$

In this network, the C-way classification layer is determined by the affective dataset with an arbitrary number of classes.

3.3. Joint Training Process

As shown in Figure 2 and referred to above, our WSCNet will produce two outputs for sentiment detection and sentiment classification tasks. Given the training set, we explicitly train the proposed deep model to optimize the joint loss function:

$$L = L_{dec}(x, y) + L_{cls}(x, y). \quad (7)$$

Since derivatives w.r.t. all the parameters can be derived, we can conduct an effective end-to-end representation learned using stochastic gradient descent (SGD) to minimize the joint loss function. With this scheme, we can detect the sentiment map using weakly supervised learning, and utilize the localized information for discriminative classification.

4. Experiments

In this section, we evaluate our method against the state-of-the-art algorithms to demonstrate the effectiveness of WSCNet for sentiment classification and detection.

4.1. Datasets

We evaluate our framework on six public datasets including the Flickr and Instagram (FI) [31], Flickr, Instagram [14], Twitter I [30], Twitter II [3] and EmotionROI [21] datasets. FI is collected by querying with eight sentiment categories (*i.e. anger, amusement, awe, contentment, disgust, excitement, fear, sadness*) as keywords from social websites. A group of 225 Amazon Mechanical Turk (AMT) participants was asked to label the images, producing 23,308 images receiving at least three agreements. The Flickr and Instagram datasets contain 60,745 and 42,856 images from Flickr and Instagram, and provide sentiment polarity (*i.e.* positive, negative) labels by crowd-sourcing based human annotation. We also evaluate the proposed method on three small-scale datasets. The Twitter I and Twitter II datasets are collected from the social websites and labeled with sentiment polarity categories by AMT participants, which consist of 1,269 and 603 images, respectively. The EmotionROI dataset is created for a sentiment prediction benchmark, which is assembled from Flickr resulting in 1980 images with six sentiment categories (*i.e. anger, disgust, fear, joy, sadness, surprise*). Besides, each image is also annotated with 15 regions that evoke sentiments, which are normalized to range between 0 and 1 as ESM.

4.2. Implementation Details

Our framework is based on the state-of-the-art CNN architecture ResNet-101 [12]. We first initialize our framework with the weights from the pre-trained model on the large-scale visual recognition [9]. In addition, we apply random horizontal flips and crop a random 448×448 patch as a form of data augmentation to reduce overfitting. We replace the last layers (global average pooling and fully connected layer) with the proposed two branches. We use a weight decay of 0.0005 with a momentum of 0.9, and fine-tune all layers with SGD. The learning rates of the convolutional layers and the last fully-connected layer on the classification branch are initialized as 0.001, 0.01 respectively, and drops by a factor of 10 every 10 epochs. The total number of iterations is 30 epochs. The FI datasets are split randomly into 80% training, 5% validation and 15% testing sets. For the Flickr and Instagram datasets, we randomly sample the same number of images for each class following the same configuration in [14], which are split randomly into 90% training, 10% testing sets. The small-scale datasets are split into 80% training and 20% testing sets randomly except those with specified training/testing splits [3, 21]. At test time, our prediction takes the output of the classification branch in the framework for classification evaluation. The sentiment map is extracted from the detection branch in Eq. 3 as the probability of regions evoking sentiment for detection evaluation. Our framework is implemented using PyTorch [19]. All of our experiments are performed on an NVIDIA GTX Titan X GPU with 32 GB on-board memory.

4.3. Evaluation Settings

To demonstrate the effectiveness of our framework for visual sentiment classification and detection, we evaluate the proposed WSCNet against the several baseline methods including methods using traditional features, CNN-based methods and weakly-supervised frameworks. For the traditional methods, we extract the principle-of-art features [34] from the affective images. We use a simplified version pro-

Table 2. Classification accuracy (%) on the testing set of FI, Flickr, Instagram, Twitter I, Twitter II, EmotionROI datasets. We evaluate the proposed WSCNet against several baseline methods including the traditional features, CNN-based methods and weakly-supervised frameworks. Note that Sun *et al*.'s method and Yang *et al*.'s method are proposed for binary classification and multi-class classification, respectively, and thus datasets with incompatible class numbers cannot be evaluated, denoted as '–'.

Method	FI	Flickr	Instagram	EmotionROI	Twitter I	Twitter II
Zhao *et al*. [34]	46.13	66.61	64.17	34.84	67.92	67.51
SentiBank [3]	49.23	69.26	66.53	35.24	66.63	65.93
DeepSentiBank [7]	51.54	70.16	67.13	42.53	71.25	70.23
ImageNet-AlexNet [15]	38.26	69.05	56.69	34.26	65.80	67.88
ImageNet-VGG16 [22]	41.22	69.88	63.44	37.26	67.49	68.79
ImageNet-Res101 [12]	50.01	72.26	67.28	40.79	72.55	70.42
Fine-tuned AlexNet	58.13	73.11	69.95	41.41	73.24	75.66
Fine-tuned VGG16	63.75	78.14	77.41	45.46	76.75	76.99
Fine-tuned Res101	66.16	80.03	79.33	51.60	78.13	78.23
Sun *et al*. [23]	–	79.85	78.67	–	81.06	80.84
Yang *et al*. [25]	66.79	–	–	52.40	–	–
WILDCAT [10]	67.03	80.67	80.31	55.05	79.53	78.81
SPN [38]	66.57	79.71	79.53	52.70	81.67	77.96
WSCNet	**70.07**	**81.36**	**81.81**	**58.25**	**84.25**	**81.35**

Table 3. Classification accuracy (%) of WSCNet using different numbers of feature maps on the test set of three large-scale datasets, *i.e*. FI, Flickr, Instagram.

Dataset	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$
FI	68.23	69.36	70.07	68.80	67.19
Flickr	81.46	81.87	81.36	81.15	81.98
Instagram	79.67	79.24	81.80	79.60	78.53

Table 4. Ablation study on the FI dataset. The baseline is the WSCNet ($k=1$) without the coupling operation, denoted as *Base*. Note that *SM* denotes using the sentiment map as the guidance, *Local* denotes that only the coupled feature map (with localized information) is used for classification, and *Coupling* denotes capturing both the holistic and localized information in Eq. 4.

Base	$k=4$	*SM*	*Local*	*Coupling*	FI
√					66.57
√	√				67.96
√		√	√		67.69
√		√		√	68.23
√	√	√		√	70.07

vided by the author to extract 27 dimension features and use LIBSVM [6] for classification. We use 1,200 dimensional mid-level representation with the ANP detector of SentiBank and apply the pre-trained DeepSentiBank to extract 2,089 dimensional features. For the basic CNN models, we report the results of using three classical deep learning methods pre-trained on ImageNet and fine-tuned on the affective datasets: AlexNet [15], VGGNet [22] with 16 layers and ResNet101 [12]. We also show the results of fully-connected features extracted from the ImageNet CNN with LIBSVM. We also report the results from three state-of-the-art deep methods for sentiment classification. For the binary datasets, we use Sun's method [23] to select top-1 regions and combine the holistic feature with the region feature from the fine-tuned VGGNet. For the multi-class datasets, we employ Yang's method [25] to transform the single label to a sentiment distribution and report the classification performance of ResNet. Moreover, we also evaluate our method against the state-of-the-art weakly supervised frameworks, *i.e*. the WILDCAT and SPN methods, which are based on ResNet-101 with the input size of 448×448 the same as our method.

4.4. Classification Performance

We first evaluate the classification performance on six affective datasets. We set the hyper-parameter $k = 4$ in the proposed WSCNet. Table 2 shows that the deep representations outperform the hand-crafted features, while the fine-tuned CNNs have the capability to recognize sentiment from images. Our proposed method consistently performs favorably against the state-of-the-art methods for sentiment classification, *e.g*. about 3.3% improvement on FI and 5.8% on EmotionROI datasets, which illustrates that WSCNet can learn more discriminative representation for this task. In addition, the weakly supervised frameworks improve the performance of Fine-tuned Res101 utilizing the regional information. Our WSCNet further improves the classification performance by 3% on Twitter I and II datasets, which shows the effectiveness of combining the sentiment-specific localized representation.

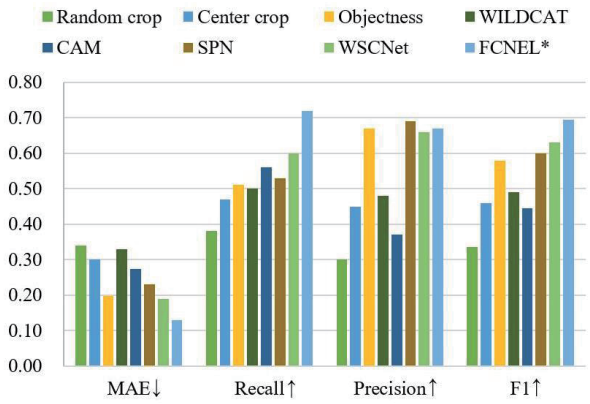

Figure 4. Sentiment detection performance on the test set of EmotionROI dataset by the baseline methods, objectness detection algorithm, weakly supervised frameworks and the supervised model. Note that '*' denotes that the method is supervised, using the bounding box annotation for training.

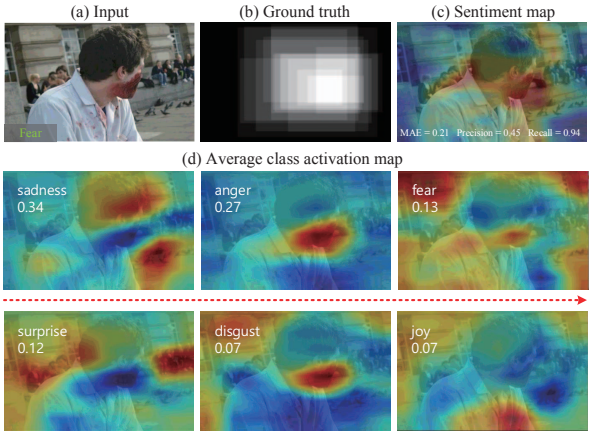

Figure 5. Detected sentiment map of the proposed WSCNet on the EmotionROI. Given the input (a) with ground truth (b), the detection result and the metrics are shown in (c). The class activation maps and the corresponding predicted scores are given in (d).

4.4.1 Hyperparameter Analysis

We now analyze the effect of the only hyper-parameter k of our framework in Eq. 1, which is the number of the response feature maps for each sentiment category. We report the classification accuracy of WSCNet with different k in the detection branch on three large-scale datasets, *i.e*. FI, Flickr, Instagram. Table 3 shows that with an increasing number of feature maps, our method is able to achieve better performance compared with the standard classification strategy in the CNN (*i.e*. $k = 1$), which captures multiple views for each sentiment category. However, over-amplifying the feature maps results in suboptimal performance mainly due to overfitting, which is similar to the finding reported in WILDCAT [10]. For the FI and Instagram datasets, our method achieves the best performance with $k = 4$, and for the Instagram dataset, the best performance is achieved with $k = 16$, although the performance is fairly stable with changing k. Therefore, we set $k = 4$ in our framework for a trade-off between efficiency and effectiveness.

4.4.2 Further Analysis

We perform an ablation study to illustrate the effect of each contribution. Our baseline is the WSCNet with $k = 1$ and without the coupling operation, where the classification branch is the original classification layer in the CNN (*i.e*. global pooling and fully connected layer). As reported in Table 4, we can draw the following conclusions: First, using both multiple feature maps ($k = 4$) and the sentiment map coupled representation improve classification accuracy by about 1% on FI, while combining the holistic and localized representations further improves the performance. Second, we achieve the best accuracy by utilizing the components to train our model in an end-to-end manner, which shows the complementarity of both these contributions.

4.5. Sentiment Detection

We evaluate the performance of sentiment map detection using the proposed WSCNet against different methods. Three baseline methods are employed to generate regions of interest for affective images. We crop the images randomly or from the center as the regions evoking sentiment, and also compare with the object regions from the objectness detection method [1]. For the weakly supervised methods, we directly extract CAM (class activation maps) from the fine-tuned ResNet-101 following [37], and the final feature maps from the WILDCAT and SPN methods are also compared. In addition, we test the supervised fully convolutional network with Euclidean Loss (FCNEL) [21] for predicting the ESM from the EmotionROI training images.

We employ the same evaluation metrics as [21], *i.e*. the mean absolute error (MAE), precision, recall, and F_1 score. All the detected regions/maps and ground truth are first normalized to 0 to 1. MAE corresponds to the mean absolute pixel-wise error between the predicted proposals and ground truth. Before computing precision and recall, we binarize each predicted map adaptively using Otsu thresholding. Thus, precision and recall represent the percentages of detected emotionally involved pixels out of all the pixels identified in the predicted region or the ground truth. F_1 score, defined as $\frac{2pr}{p+r}$, measures the harmonic mean of precision p and recall r.

4.5.1 Detection Results

Figure 4 shows that our WSCNet performs favorably against the baselines and weakly supervised methods, which

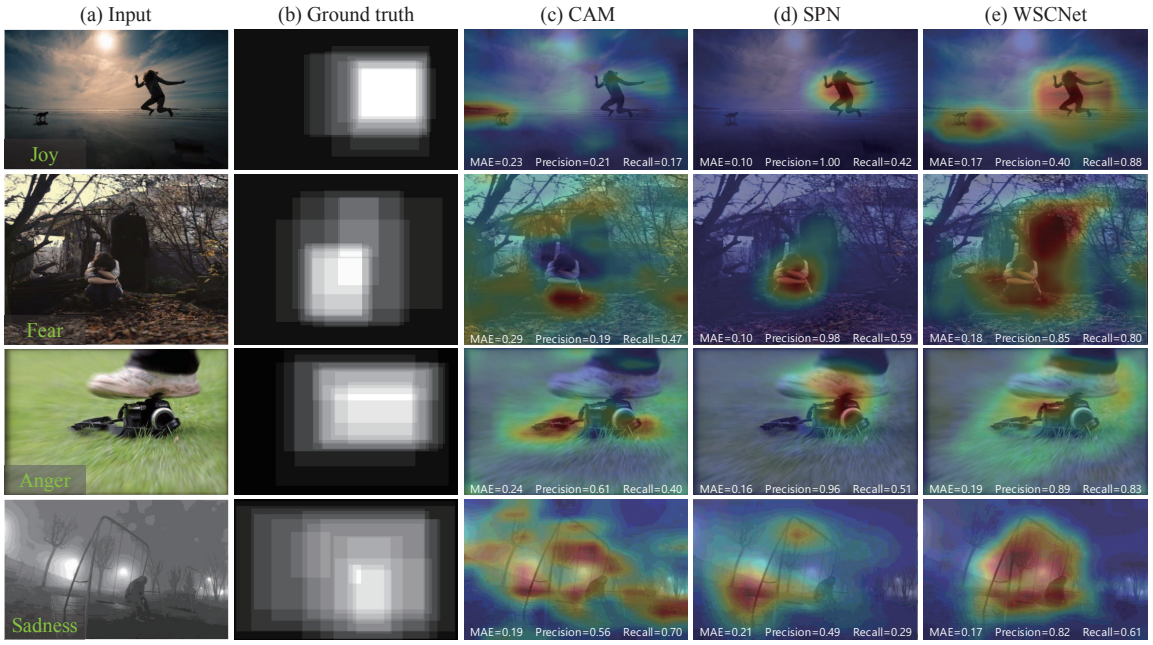

Figure 6. Weakly supervised detection results using different methods on the EmotionROI testing set. The input images and the ground truth are given in (a) and (b). The detected regions and metrics of weakly-supervised methods (*i.e*. CAM, SPN, ours) are shown in the last three columns. By activating the sentiment-related areas, our method is more accurate to the ground truth.

achieves comparable performance with the supervised FCNEL on most evaluation metrics. We notice that FCNEL benefits from supervised training with bounding box annotation, and has significantly better recall than other methods. The reason is that the regions evoking sentiments contain both the primary objects and additional contextual background, while Objectness [1] only focuses on the foreground objects and thus achieves a reasonable precision. Compared with the weakly supervised methods, our method improves the recall to 0.60, which illustrates the effectiveness of taking the sentiment characteristic into consideration for generating the sentiment map.

4.5.2 Visualization

Figure 5 shows the detected sentiment map for a fear image from the EmotionROI, and the activation map for each sentiment from the detection branch. Note that the sentiment scores are also from the detection branch corresponding with the pooled vector v illustrated in Section 3.1. Although the ground-truth class prediction (*i.e*. fear) is not always the highest, the high scores are from related classes (*e.g*. other negative sentiments like sadness and anger) providing the complementary information, which is reasonable since the detection branch achieves sub-optimal classification performance. Thus, the weighted combination is able to generate more reliable sentiment maps. In Figure 6, we show more detection results using different weakly supervised methods. Compared with the ground truth, the WSCNet is able to detect the relevant regions that influence the evoked sentiment, while the CAM and SPN may only focus on the salient objects leading to a reasonable precision. For example, on the third row, SPN only responds to the foreground objects, which leads to 0.96 precision but only 0.51 recall. In contrast, our detected sentiment map extends the object regions into the sentiment related background, which achieves the recall of 0.83.

5. Conclusions

This paper addresses the problem of visual sentiment analysis based on convolutional neural networks, where the sentiments are predicted using multiple affective cues. We present WSCNet, an end-to-end weakly supervised deep architecture, which consists of two branches for discriminative representations learning. The detection branch is designed to automatically exploit the sentiment map, which can provide the localized information of the affective images. Then the classification branch leveraging both holistic and localized representations can predict the sentiments. Experimental results show the effectiveness of our method against the state-of-the-art on six benchmark datasets.

Acknowledgments

This research was supported by NSFC (No. 61620106008, 61572264, 61633021, 61525306, 61301238, 61201424), NSF CAREER (No. 1149783), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[2] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

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

[4] V. Campos, B. Jou, and X. Giró i Nieto. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image Vision Comput.*, 65:15–22, 2017.

[5] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou. Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction. In *ACM ASM*, 2015.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[7] T. Chen, D. Borth, T. Darrell, and S. F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):189–203, 2017.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, 2017.

[11] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

[14] M. Katsurai and S. Satoh. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *ICASSP*, 2016.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[16] B. Li, W. Xiong, W. Hu, and X. Ding. Context-aware affective images classification based on bilayer sparse representation. In *ACM MM*, 2012.

[17] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[19] A. Paszke, S. Gross, S. Chintala, et al. Pytorch, 2017.

[20] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 2015.

[21] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *ICIP*, 2016.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[23] M. Sun, J. Yang, K. Wang, and H. Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *ICME*, 2016.

[24] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang. Retrieving and classifying affective images via deep metric learning. In *AAAI*, 2018.

[25] J. Yang, D. She, and M. Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017.

[26] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 2018.

[27] J. Yang, M. Sun, and X. Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI*, 2017.

[28] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek. Emotional valence categorization using holistic image features. In *ICIP*, 2008.

[29] Q. You, H. Jin, and J. Luo. Visual sentiment analysis by attending on local image regions. In *AAAI*, 2017.

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

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

[32] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[33] S. Zhao, G. Ding, Y. Gao, and J. Han. Approximating discrete probability distribution of image emotions by multi-modal features fusion. In *IJCAI*, 2017.

[34] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 2014.

[35] S. Zhao, H. Yao, Y. Gao, G. Ding, and T. S. Chua. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing*, 2018.

[36] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua. Predicting personalized emotion perceptions of social images. In *ACM MM*, 2016.

[37] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[38] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. In *ICCV*, 2017.