Learning Discriminative Sentiment Representation from Strongly- and Weakly-Supervised CNNs

DONGYU SHE, Nankai University, China
MING SUN, SenseTime, China
JUFENG YANG, Nankai University, China

Visual sentiment analysis is attracting increasing attention with the rapidly growing amount of images uploaded to social network. Learning rich visual representations often requires training deep convolutional neural networks on massive manually labeled data, which is expensive or scarce especially for a subjective task like visual sentiment analysis. Meanwhile, a large quantity of social images is quite available yet noisy by querying social network using the sentiment categories as keywords, where a various type of images related to the specific sentiment can be easily collected. In this paper, we propose a multiple kernel network (MKN) for visual sentiment recognition, which learns representation from strongly- and weakly- supervised CNNs. Specifically, the weakly-supervised deep model is trained using the large-scale data from social images, while the strongly-supervised deep model is fine-tuned on the affecitve datasets with manual annotation. We employ the multiple kernel scheme on the multiple layers of CNNs, which can automatically select the discriminative representation by learning a linear combination from a set of pre-defined kernels. In addition, we introduce a large-scale dataset collected from popular comics of various countries, *e.g.*, America, Japan, China and France, which consists of 11,821 images with various artistic styles. Experimental results show that MKN achieves consistent improvements over the state-of-the-art methods on the public affective datasets as well as the newly established Comics dataset. **The Comics dataset can be found on http://cv.nankai.edu.cn/projects/Comic.**

CCS Concepts: • **Computing methodologies** → **Computer vision tasks**; *Object recognition*; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Social and professional topics**;

Additional Key Words and Phrases: Visual sentiment analysis, convolutional neural network, multiple kernel learning

ACM Reference Format:
Dongyu She, Ming Sun, and Jufeng Yang. 2019. Learning Discriminative Sentiment Representation from Strongly- and Weakly-Supervised CNNs. *ACM Trans. Multimedia Comput. Commun. Appl.* 00, 0, Article 000 (2019), 18 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

With the increasing popularity of online social network, more and more users tend to express their opinions through visual contents, *e.g.*, images and videos. How to make machine perceive human beings' understanding towards the affecitve contents is of significance due to its broad applications, *e.g.*, opinion mining [50, 68], affective computing [1, 51], entertainment [10, 22], *etc.* In the last few years, convolutional neural network (CNN) has enabled robust and accurate representation

Authors' addresses: Dongyu She, Nankai University, NO. 38 Tongyan Road, Tianjin, 300350, China, sherry6656@163.com; Ming Sun, SenseTime, NO. 1 East Zhongguancun Road, Beijin, 100084, China, m_sunming@163.com; Jufeng Yang, Nankai University, NO. 38 Tongyan Road, Tianjin, 300350, China, yangjufeng@nankai.edu.cn.

1551-6857/2019/0-ART000 $15.00
https://doi.org/0000001.0000001

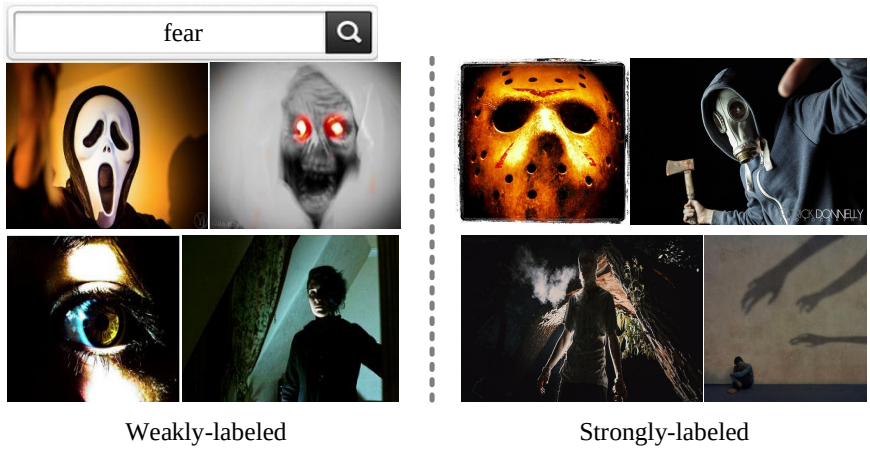

Fig. 1. Samples with different annotation for the same sentiment. (a) The weakly labeled affective images are queried from social websites using the sentiment category (*i.e.*, fear) as the keyword. (b) The well labeled images are selected from the same category of the Flickr and Instagram dataset [60]. The availability of the large-scale weakly labeled images encourages us to exploit the web data to improve the performance of visual sentiment prediction.

learning given massive manually labeled data [26], which in turn produces the state-of-the-art performance on various computer vision related tasks, *e.g.*, object recognition [19, 64], object detection [41], semantic segmentation [43], *etc.* The traditional CNN framework extracts features from the topmost layer which has the most discriminative ability for the concrete classification tasks [62]. However, recognizing the affective content of images [29, 60] is more challenging due to the following reasons.

First, the performance of CNN highly depends on the amount of training data, and satisfactory performance can be achieved only when adequate samples are provided. While most of the well labeled datasets introduced in this domain rarely contain more than thousands of samples. An exception is the Flickr and Instagram dataset [60] consisting of 23,308 images, which is collected from 3 million web images and 225 workers are employed for labeling. It is too labor- and time-consuming to obtain a well labeled dataset which is large enough for training robust CNN, especially for the abstract and artistic images where the expert knowledge is required. Second, as visual sentiment analysis involves a higher level of subjectivity in the human recognition process [25], it's insufficient to use the topmost features to represent emotions. In fact, some studies reveal that texture information is one of the most important elements related to visual emotion [29, 40], which can be reflected in the non-topmost layers.

We address the above problem via a multiple kernel network (MKN), which can learn representation leveraging the large-scale web data. Instead of collecting massive manually labeled images, one solution is to query images from the social network using the sentiment categories as keywords. As shown in Figure 1, though web images may be captioned or labeled imprecisely while uploading, the large sources of visual sentiment content can enhance the generalizability of the learned model. In specific, we train a weakly-supervised sentiment prediction model using the large-scale web images, and finetune a strongly-supervised model based on the affecitve datasets that are manually labeled. We employ the multiple kernel scheme on the non-topmost layers of both weakly- and strongly-

model, making full use of the spatial and semantic information for visual sentiment prediction. We learn the essential information from different layers of CNNs by a linear combination of a set of pre-defined kernels. To validate the generalizability of weakly-supervised model, we also introduce a new dataset collected from a different domain of comics. We collect images from popular comics of various countries, *e.g.*, America, Japan, China and France, which finally results in 11,821 images with various artistic style. Then ten participants are employed to label the comic images with eight sentiment categories in Mikels' theory [32]. Exploring the relationship between sentiment and such abstract scene can provide higher-level information for bridging the semantic gap, which is also useful since it can bring convenience to the automatic classification of digital comics.

Our contributions are summarized as follows. First, this paper is the first to consider the multiple-kernel scheme in the CNN framework for visual sentiment analysis. Instead of simply and directly combining the multi-kernels and multiple features, we propose to automatically learn the kernel coefficients of features from the weakly-and strongly- supervised models for robust representation. Second, this paper collects a large-scale Comics dataset from popular comics of various countries, *e.g.*, America, Japan, China and France, which finally results in 11,821 images with various artistic style. Both quantitative and qualitative experimental evaluations demonstrate that the proposed algorithm performs favorably against the state-of-the-art classification methods on the public affective datasets as well as the newly established Comics dataset.

The rest of this paper is organized as follows. Section 2 summarizes the related work on visual sentiment analysis and the related CNN-based methods. Section 3 introduces the proposed MKN for visual sentiment prediction. Section 4 analyzes the experimental results on the popular benchmark datasets as well as the collected dataset. Finally, Section 5 concludes this paper.

2 RELATED WORK

2.1 Visual Sentiment Analysis

The literature on visual sentiment analysis can be roughly divided into dimensional space approaches [65] and categorical approaches [72]. The dimensional space approaches represent sentiment in a two or three-dimensional space, *e.g.*, valance-arousal (VA) [2, 70], and valance-arousal-dominance (VAD) [17]. The valence denotes the pleasantness ranging from happy to unhappy, and the arousal denotes the intensity of sentiment ranging from excited to calm, while the dominance denotes the degree of control ranging from controlled to in control. The categorical approaches represent sentiments with a set of basic categories, such as Ekman's six basic emotions [12, 13] (*i.e.*, happiness, sadness, disgust, anger, fear, and surprise) and Mikels' eight categories [32] (*i.e.*, amusement, contentment, awesome, excitement, sadness, disgust, anger and fear). The categorical perspective has been widely applied in recent studies due to the intuitive appeal [7, 28, 54].

In the early stage, several methods propose to design different hand-crafted features [29, 66] for predicting visual sentiment. For example, Machajdik *et al.* [29] define a combination of rich low-level features based on art and psychology theory, *e.g.*, color, texture, composition, and content, *etc.* Zhao *et al.* [69] introduce more robust and invariant visual features designed according to art principles, while Sartori *et al.* [42] exploit art theory concepts to infer the sentiments elicited by abstract paintings based on the sparse group lasso approach. In [27], shape features are also proved to be of significance for sentiment prediction. As low-level information is limited in bridging the semantic gap between pixels and sentiment, both [61] and [3] propose mid-level representations for sentiment prediction. The former work proposes Sentribute, an image-sentiment analysis algorithm based on 102 mid-level attributes, while the latter designs a large scale visual sentiment ontology based on adjective noun pairs (ANP) and trains 1,200 visual detectors, namely SentiBank. Moreover,

Table 1. Statistics of the available affective datasets, which are divided according to different labeling supervision. The dataset size and annotation voters number are also shown. Here, '*' denotes the total voter number. Note that ArtPhoto is labeled by the artist who uploads the image to the art-sharing website. And the weakly labeled Flickr dataset is collected via the search engine.

Annotation Type	Dataset	Class Number	Images Number	Voter Number
Strongly-supervised	IAPSa [29]	8	395	60*
	ArtPhoto [29]	8	806	1
	Abstract Paintings [29]	8	228	230*
	Twitter I [59]	2	1,269	5
	Twitter II [3]	2	596	3
	Emotion6 [37]	6	1,980	5
	Flickr and Instagram [60]	8	23,308	225*
Weakly-supervised	Flickr [3]	2	484,222	0

Chen *et al.* [9] model sentiment concepts based on six frequent objects (car, dog, dress, flower, face, food), proving that objects play important roles in the affective presentation. Zhao *et al.* [71] combine both low-level and mid-level features into a multi-graph learning framework for affective images retrieval.

More recently, CNN-based approaches have also been applied to recognize sentiments, and have achieved significant advances. The existing CNN frameworks can be viewed as classification [4, 5, 34] or regression [37, 55, 57] models, which employ a softmax loss to maximize the probability of the correct class or Euclidean loss to minimize the difference of the squares between the prediction and the ground truth. For the classification task, Chen *et al.* [8] train a deep convolutional neural network model named DeepSentiBank on Caffe [23], which improves the results on both annotation accuracy and retrieval performance. You *et al.*[60] further train deep models on a well-labeled dataset, which is collected from 3 millions of weakly labeled Flickr and Instagram images. In [40], a multi-level deep network (MldeNet) is proposed to unify both low-level and high-level information in well labeled images. Zhu *et al.*[74] further propose a unified CNN-RNN framework that learns different levels of features and integrates them by exploring the dependencies. He and Zhang [21] introduce an assisted learning strategy in the CNN training to boost the recognition performance, which composes of a binary positive-or-negative emotion network and a deep network to recognize the specific emotion of an image. In addition, different from improving whole image representations using a CNN, there are several methods that prove sentiments of some images are much related to local salient regions [14, 45, 73]. You *et al.*[58] present the attention model to learn the correspondence between local image regions and the sentimental visual attributes. Limited by the visual attribute detector, such local features on visual sentiment analysis is not significant. Sun *et al.* [48] and Yang *et al.* [56] propose a framework to discover the affective regions and combine the local information with the global representation for sentiment analysis. Furthermore, an end-to-end framework is proposed in [53] to detect the sentiment regions in a weakly-supervised manner, which is then coupled with the global feature maps for sentiment recognition. General surveys are provided in [63, 67].

However, the existing CNN methods are trained on the well-labeled affective datasets with limited scale. Table 1 shows the statistics of the popular affective datasets, which are divided according to different labeling way. As can be seen, most datasets includes limited data compared to the Flickr dataset [3], including IAPSa, ArtPhoto, Abstract Paintings [29], Twitter I [59], and Twitter II [3], and Emotion6 [37]. The IAPSa dataset includes 246 images selected from IAPS, which are labeled by 60 participants. The ArtPhoto dataset includes 806 artistic photographs from a photo sharing

site, the labels of which are determined by the artist who uploaded the photo. And the Abstract dataset consists of 230 peers that vote for the abstract paintings without contextual content. 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 Emotion6 dataset is created for a sentiment prediction benchmark, which is assembled from Flickr resulting in 1,980 images labeled with six sentiment categories. The Flickr and Instagram dataset (FI) contains about 90,000 noisy images by querying the emotional keywords from the social platform, *e.g.*, Flickr and Instagram. And 225 Amazon Mechanical Turk (AMT) workers, selected through a qualification test, are employed to vote for the keyword representing images, resulting in 23,308 images receiving at least three agreements. The number of images in each class is larger than 1,000. The Flickr dataset is constructed by retrieving the Flickr creative common images for the 3,000 ANPs after excluding the images that do not contain the ANP string in the title, tag or description, resulting in about 500k images for 1,553 ANPs.

Due to lack of data, several methods are proposed to utilize web images by filtering the noisy data [52, 59]. You *et al.*[59] propose a novel progressive CNN architecture to make use of noisy machine labeled Flickr dataset, and they suggest that leveraging larger weakly labeled dataset can rise the generalizability of the deep model. Wu *et al.* [52] propose to refine the dataset from the social network based on the sentiments of ANPs and tags. Different from the previous work, this paper incorporates the MKL scheme for learning the discriminative representation from the weakly-supervised model trained on the web data.

2.2 Combining the CNNs and Kernel Methods

The existing works for sentiment prediction use either pixel-level features or utilize CNNs as a generic feature extractor, which does not take advantage of the influence of different layers. Several works have investigated using CNNs related features from multiple layers in computer vision tasks. In these work, the activations from certain layers are used as descriptors by concatenating [35] or pooling [16, 18]. Peng *et al.*[36] utilize multi-scale convolutional neural networks to extract the features with different scales for the task of interest that is label-inheritable. Gao *et al.*[15] propose to combine features from different models, where the pooling operation contributes to dimensional reduction, which leads to less memory-consumption.

Several attempts have been made in the past to unify the kernel methods and the potentially attractive features from deep networks. Cho *et al.*[11] first introduce the concept with the arc-cosine kernel, which admits an integral representation that can be interpreted as a one-layer neural network with random weights and an infinite number of rectified linear units. Then, hierarchical kernel descriptors and convolutional kernel networks extend a similar idea in the context of images leading to unsupervised representations [30]. The backpropagation algorithm for the Fisher kernel introduced in [49] learns the parameters of a Gaussian mixture model with supervision, which requires a probabilistic model and learns parameters at several layers. Multiple kernel learning (MKL) [46, 75] provides techniques to select a combination of kernels from a pre-defined collection, and reflects the fact that typical learning problems often involve multiple, heterogeneous data sources. Poria *et al.*[38] propose a multi-modal affective data analysis framework, and employ MKL to combine different modalities. In addition, Jiu *et al.* [24] propose a deep kernel framework for classification, focusing on selecting an appropriate kernel by a multi-layered linear combination of activation functions. However, such method is based on traditional visual features (*e.g.*, SIFT, GIST, LBP) as the fixed input data, which is sub-optimal due to that the extracted features do not consider sentiment semantics during learning.

Different from the previous work, we leverage different models trained on well labeled dataset as well as large scale web data. We propose to joint supervised neural units with the weakly supervised

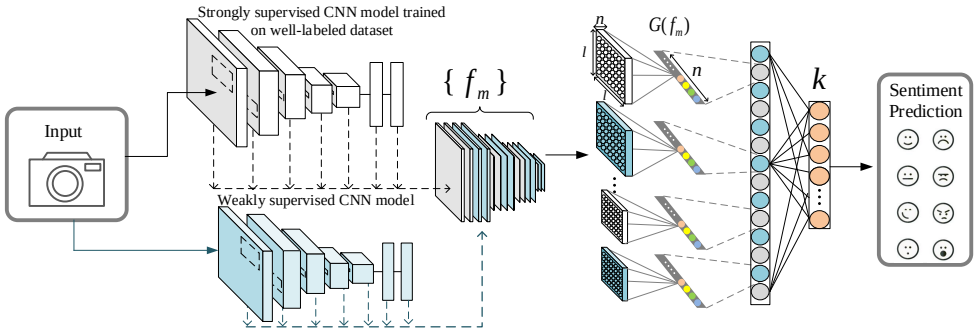

Fig. 2. Overview of the proposed MKN. We first train the strongly- and weakly-supervised models based on the manually labeled affective dataset, and a large scale web dataset from the social network, respectively. Both models adopt the VGGNet architecture. Then, features from the multiple layers of both models are organized as the alternatives $\{f_m\}$. Three pooling strategies are employed to approximately transform f_m to $G(f_m)$, avoiding calculating the massive w_m in Equation 5. Thus, the sentiment representation can be obtained adopting the multi-kernel scheme on the alternatives for visual sentiment prediction.

representation. A simple way to combine different representation is concatenating features, which preserves information captured by different models to some extent. However, redundancy may occur at the same time. Inspired by the kernel methods, we develop the algorithm to select the specific neural units crossing layers, which avoids redundancy and is able to conserve useful information hidden in various features.

3 METHODOLOGY

Figure 2 illustrates the framework of the proposed MKN for learning sentiment representation from strongly- and weakly- supervised CNNs. Specifically, we employ the CNN architecture as the basic models for both strongly- and weakly- supervised deep model. Then, the stronly-supervised CNN model is fine-tuned with the manually labeled affective dataset, while the weakly- supervised CNN model is trained using the large-scale social images with noisy labels. Our proposed MKN takes multiple layers from both strongly- and weakly- supervised models as an alternative, and then automatically select the discriminative representation by learning a linear combination of a set of pre-defined kernels.

3.1 Traditional Multiple Kernel Learning

In computer vision the problem of learning a multi-class classifier from training data is often addressed by means of *kernel methods*. Kernel methods utilize kernel function defining a measure of similarity between pairs of instances. For a kernel function $K(x, x')$ between real vectors, it actually plays several roles: it defines the similarity between two example x and x', while defining an appropriate regularization term for the learning problem. A common approach is to consider that the kernel $K(\cdot, \cdot)$ is actually a convex linear combination of other basis kernels:

$$K(x, x') = \sum_{m=1} d_m k_m(x, x'). \tag{1}$$

Algorithm 1 Approximate Solution for MKN

Input: $X = \{x_1, x_2, ...x_N\}$ is a set of images in size 227×227
$Y = \{y_1, y_2, ...y_N\}$ denotes the sentiment labels of X
Model is a fine-tuned CNN model
Pass x_i through the *Model* from the second layer to the last layer
Organize features from different selected layers and models as set$\{f_m\}, m = 1, \cdots, F$.
for m from 1 to F **do**
  Kernel selection: $\{Gaussian, Poly, Linear, Sigmoid\}$
  Let $k_m(x, x') = k(G(f_m(x)), G(f_m(x'))$
  Calculate β_m in Equation 4 using SILP and then get $K(x, x')$.
end for
Training multi-kernel SVM with $K(x, x')$.
$\hat{y'}_j \leftarrow$the predicted label corresponding to the j-th image x'_j in testing set X';
$\hat{Y'} \leftarrow$ the predicted labels corresponding to X'.

For kernel algorithms, the solution of the learning problem is of the form:

$$f(x) = \sum_{i=1}^{l} \alpha_i^* y_i K(x, x_i) + b^*, \tag{2}$$

where $\{x_i, y_i\}_{i=1}^N$ are the training examples, and N denotes the number of the learning examples. Here, $K(\cdot, \cdot)$ is a chosen positive definite kernel and $\{a_i^*\}$,b^* are some coefficients to be learned from examples. Thus, learning both the coefficients α_i and the weighted d_k in a single optimization problem is known as the multiple kernel learning problem.

For the learning algorithms, we use the MKL-SILP [47] implementation of the original authors available at http://www.shogun-toolbox.org.

3.2 Problem formulation

For each instance, different layers of weakly- and strongly-supervised CNNs are used to generate distinct representations. In this paper, we choose features from both pooling and fully connected layers, called $\{f_m\}_{m=1,...,F}$, where F is the number of the selected layers. Then, we define $d_m = l_m \times l_m \times n_m$ to represent the dimensionality of f_m, where l_m is the size of the feature map and n_m denotes the number of feature maps in f_m. To make full use of different aspects of different features, we use kernel method for feature combination.

For the m-th feature, the relationship between two samples is defined by

$$k_m(x, x') = k(f_m(x), f_m(x')), \tag{3}$$

where k corresponds to a specific function, such as Gaussian function, kernel $k_m : \chi \times \chi \rightarrow \mathbb{R}$ calculates the similarity with respect to the m-th feature. Therefore, we can combine several kernels into a single model $K(x, x')$ for feature set $\{f_m\}$:

$$K(x, x') = \sum_{m=1}^{F} \beta_m k_m(G(f_m(x)), G(f_m(x'))), \tag{4}$$

where coefficients β_m satisfy the constraint $\sum_{m=1}^F \beta_m = 1$. $G : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{d'_m}$ represents a non-linear map function that maps f_m to another space $\mathbb{R}^{d'_m}$, where $d'_m \leqslant d_m$. G is defined by

$$G(f_m(x)) = \langle f_m(x), w_m \rangle, \tag{5}$$

where w_m denotes weights. The features extracted from different layers of CNNs are usually with high dimensions and sparsity. In this work, G help transform the deep features, making them clearer and less redundant. As a result, the feature space is condensed.

Given a training set $\{(x_i, y_i)\}_{i=1,\ldots,N}$ of N instances consisting of an image $x_i \in \chi$ and a sentiment label $y_i \in \{1, \ldots, C\}$, our aim is to learn a discriminative representation for sentiment classification.

3.3 Approximate Solution

However, the w_m in map function G rely on the characteristics of the m-th layer, and the joint learning of β_m and w_m in Equation 4 is difficult [46]. To deal with this problem, we replace the calculation of w_m with relaxed methods and then optimize β_m. Since the feature maps of convolutional layers contain many 0 elements, the relaxed methods are used to pick out the useful neural units. In this paper, three kinds of pooling strategies, *i.e.*, random pooling, sum pooling, and max pooling, are employed.

Random pooling G_R means that we randomly select one element from each feature map as its representation. The network employing random pooling is named as MKN_R in our work. G_R is defined as

$$G_R(f_m(x)) = [r f_m^q(x) r^T]_{q=1,2,\ldots,n_m}, \tag{6}$$

where $f_m^q(x)$ is the q-th feature map of the m-th selected layer. q ranges from 1 to n_m, $r \in \mathbb{R}^{l_m}$ is a vector whose elements are defined as

$$r_i = \begin{cases} 1, & i = random(l_m) \\ 0, & otherwise \end{cases} \tag{7}$$

where $random(l_m)$ randomly selects a location i in the vector r, and sets the corresponding element r_i as 1. Other elements in r are set to 0. Finally, G_R generates a new n_m-dimensional feature vector to represent $f_m(x)$.

Similarly, the network employing sum pooling is named as MKN_M, where sum pooling G_S indicates each feature map is represented by the sum of all its units values. G_S is defined as

$$G_S(f_m(x)) = [\sum_{i=1}^{l_m} \sum_{j=1}^{l_m} (f_m^q)_{ij}]_{q=1,2,\ldots,n_m}. \tag{8}$$

The network employing max pooling is named as MKN_M, where max pooling G_M indicates each feature map is represented by the maximum value of its units. G_M is defined as

$$G_M(f_m(x)) = [\max_{1 \leqslant i,j \leqslant l_m} (f_m^q)_{ij}]_{q=1,2,\ldots,n_m}. \tag{9}$$

The three "degradation" strategies, *i.e.*, random, sum, and max operation, can simplify Equation 4 into an easily solved solution. Here, β_m can be computable by existing kernel method, *e.g.*, SimpleMKL [39]. In this paper, SILP [46] is employed to calculate β_m. To validate the effectiveness of our methods, we also use several combinations of classical kernels to dynamically select sensitive units from the feature maps, *e.g.*, Polynomial, Gaussian, Linear, and Sigmoid. Our proposed algorithm is summarized in Algorithm 1. We have compared the performance of the above three strategies in the experiments section. Despite that our method is an approximate solution, it can dynamically find important neural units for sentiment analysis from the multi-scale feature maps.

3.4 Computational Complexity

In traditional multiple kernel learning algorithms, the complexity is tied to the one of single kernel SVM algorithm. Thus, given the dataset with N samples, the computational complexity of the proposed Algorithm 1 is of the order of $F \cdot n_{SV}^2$, where $n_{SV} < N$ denotes the number of support

Table 2. Performance of our methods on the FI dataset. Here, "ft" refers to finetuning the model.

Methods	Acc. (%)
CaffeNet [23]	46.28
CaffeNet + ft	56.79
VGGNet [44]	54.10
VGGNet + ft	59.61
ResNet [20]	50.01
ResNet + ft	61.82
DeepSentiBank [8]	51.54
PCNN (CaffeNet) [59]	46.09
PCNN (VGGNet) [59]	55.24
MKN_R	**61.75**
MKN_M	**63.92**
MKN_S	**63.76**

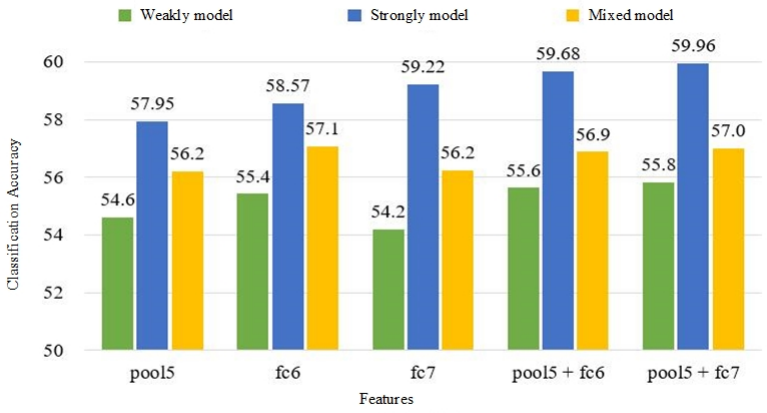

Fig. 3. Performance comparison of using different layers (pool5, fc6, fc7) for visual sentiment classification. "pool5+fc6" and "pool5+fc7" mean concatenating the pool5 with the fully connected layers. The mixed model refers to the model trained on the combination of well labeled and weakly labeled data. The results of the mixed model underperforms the strong model indicates the combination of data does not work well in this task and kernel methods are needed.

vectors, and F is the number of features. Here, since we have $F << N$, the complexity of Algorithm 1 is $O(N^2)$.

4 EXPERIMENTS

In this section, we present the experiments and evaluate our method against the state-of-the-art approaches to validate the effectiveness of MKN for sentiment analysis.

4.1 Experiment Setup

We evaluate our proposed method with two datasets, including the Flickr dataset [3] and the Flickr and Instagram dataset [60]. For convenience, we use the abbreviation "FI" to denote the Flickr and Instagram dataset. Following [60], we randomly split FI into 80% training, 5% validation and 15% testing set. In this work, the deep CNNs are pre-trained in a fully-supervised way on the large scale annotated dataset for image classification [44]. Then we employ the same strategies to transfer the

Table 3. Classification accuracy of multi-scale features using a supervised model ('S') as well as weakly-supervised model ('W') on the 15% randomly chosen testing data. "MKN_R", "MKN_M", "MKN_S" indicate using the random pooling, max pooling, sum pooling to select sensitive neural units from feature maps, respectively.

Pool1		Pool2		Pool3		Pool4		Pool5		FC7		MKN_R	MKN_M	MKN_S
S	W	S	W	S	W	S	W	S	W	S	W			
√										√		60.52	61.36	61.52
		√								√		61.11	61.55	61.40
				√						√		61.60	61.64	61.81
						√				√		61.73	61.87	62.14
								√		√		61.58	62.14	62.17
√		√		√						√		60.82	61.73	61.87
√				√				√		√		61.55	61.99	62.19
				√		√		√		√		**61.68**	**62.29**	**62.44**
								√	√	√		61.67	62.49	62.28
								√	√	√	√	61.69	62.50	62.27
√	√			√	√			√	√	√		61.65	62.89	62.56
				√	√	√	√	√	√	√		61.72	63.05	62.97
√	√	√	√	√	√	√	√	√	√	√		**61.75**	**63.92**	**63.76**

Table 4. Performance of three "degradation" strategies on the 15% randomly chosen testing set on the FI dataset. Note that 'RP', 'MP' and 'SP' denote the random pooling, max pooling, and sum pooling, respectively.

Layer	Dimension	RP	MP	SP
pool1	64	27.03	36.48	35.87
pool2	128	27.73	41.06	42.74
pool3	256	32.67	46.52	48.61
pool4	512	38.83	53.53	55.74
pool5	512	45.08	58.53	60.19
pool1 + fc7	564	**59.79**	60.99	60.87
pool2 + fc7	628	59.35	60.89	60.98
pool3 + fc7	756	58.94	60.82	**60.99**
pool4 + fc7	1012	59.32	60.76	60.87
pool5 + fc7	1012	59.44	**61.11**	60.64

learned parameters to the task of sentiment analysis. We first fine-tune the VGGNet [44] on half a million web Flickr images in a weakly-supervised way. The 1000-way fc8 classification layer is replaced to 2-way due to the binary classification task. This model addresses the weakly labeled nature of the dataset, using which we can capture more kinds of interests from emotional images. Meanwhile, we fine-tune VGGNet on the FI dataset as our strongly-supervised model, where the last layer is changed to 8-way since the dataset is labeled into eight emotional categories. Both models run 100,000 iterations to update the parameters. Starting from VGGNet, we resize the images to 224×224 required by the network and pass it through different CNNs models. To evaluate the performance of the none-topmost as well as the topmost layers for sentiment classification, we also train classifiers based on the features extracted from the single model using LIBSVM [6]. For each layer, we simply consider the output as a compact feature vector and employ the *one v.s. all* strategy following the same routine of previous work [29].

Table 5. Classification accuracy of multi-scale features using the mixed model on the 15% randomly chosen FI testing data. "MKN_R", "MKN_M", "MKN_S" indicate using the random pooling, max pooling, sum pooling to select sensitive neural units from feature maps. Here, the mixed model denotes training with all the data.

Model	Pool1	Pool2	Pool3	Pool4	Pool5	FC7	MKN_R	MKN_M	MKN_S
Mixed Model					√	√	58.75	59.87	59.98
	√		√		√	√	58.81	59.15	59.32
			√	√	√	√	58.88	59.97	60.26
	√	√	√	√	√	√	58.97	61.53	61.87

Table 6. Statistics of the traditional affective datasets (*i.e.*, IAPSa, ArtPhoto, Abstract Paintings) and the collected Comics dataset. The number of images per emotional category of the Comic and Manga subsets are also given. Compared with the others, our collected Comics dataset is multiple times larger.

Datasets	Amusement	Awe	Contentment	Excitement	Anger	Disgust	Fear	Sadness	Sum
IAPSa	37	54	63	55	8	74	42	62	395
ArtPhoto	101	102	70	105	77	70	115	166	806
Abstract Paintings	25	15	64	36	3	18	36	32	229
Comics	2005	449	2446	900	1776	829	2137	1279	11,821
Comic set	308	481	107	850	368	458	917	272	3761
Manga set	1697	1295	342	1596	461	442	1220	1007	8060

4.2 Results on the FI Dataset

Table 2 shows the performance of the current state-of-the-art methods on the FI dataset. The ft refers to finetuning the pre-trained model on the FI dataset.

It is obvious that the VGGNet performs better than the CaffeNet with the help of deeper layers and more parameters. Besides, the ImageNet is useful for the boosting of performance and consistent with the expectation. Compared to the method proposed by [8], our method is able to generate the robust representation of the image sentiment and describe the sentiment distribution. Apart from it, You *et al.*[59] propose the promising way of using lots of web images and train the robust model. However, our method can dynamically select the neural unit of the feature map and get rid of the noisy web images. The proposed method can mine the useful features from both models and boost the results on. To further analyze the effectiveness of our method, we discuss the different aspects containing the different layers of both models and different combination ways in the following subsections. To mine the information from the weakly model, we discuss how to make full use of the weakly labeled data in the next subsection.

4.3 Different Layers and Models

Figure 3 summarizes the performance of different layers output on the 15% testing set of FI, including the fifth pooling layer and the following two fully connected layers, called *pool5, fc6, fc7*, respectively. In addition, we also compare the results of simply concatenating the earlier layer and the fully connected layer based on different models. For example, pool5 + fc7 represents concatenating the features of the fifth pooling layer and the fully connected layer.

Different layers For the weakly- and strongly-supervised models, the fc7 features perform better than the earlier layers, and the overall accuracy of the strong model is about 59%. However, the concatenating of pooling and fully connected layers is better than using a single layer only,

Fig. 4. Examples of the proposed Comics dataset. The first group (top rows) shows images collected from Asian comics, consisting of abundant of lines. The second group (bottom rows) shows images from European and America comics, which are more realism and exquisite.

which demonstrates that the earlier layer is necessary for the image sentiment analysis. The features from the earlier layers are related to the lower level information while the latter layers correspond to the specific task.

Different models As shown in Figure 3, the weakly-supervised model individually has limited ability for sentiment prediction compared to the strong model. It is natural because the weakly-supervised model is trained based on the large scale web images which contain a lot of noises. However, as is shown in Table 3, combining the output from the earlier layer with the fully connected layer brings an improvement of the accuracy on both models. It shows that the weakly-supervised model can be useful for the sentiment classification task and we will discuss how to combine it with a strong model in the next section. Since that fc6 layer and fc7 layer have similar impacts on the combination, the following experiments will only use the fc7 to represent the fully connected layer in default.

Table 7. Performance comparison with the state-of-the-art algorithms on the proposed Comics dataset.

Method	Amuse	Awe	Content	Excite	Anger	Disgust	Fear	Sad	All
GIST [33]	23.03	11.35	38.29	6.75	10.03	20.98	50.21	9.53	27.75
SentiBank [3]	38.40	42.70	30.88	20.56	30.90	26.51	29.44	21.18	30.20
CaffeNet [44]	36.00	27.94	33.88	13.33	27.44	19.35	33.71	28.08	30.51
CaffeNet + ft	46.88	57.78	32.24	22.22	37.18	35.54	30.61	21.09	34.40
VGGNet [44]	47.56	58.21	38.32	22.66	39.01	35.86	32.19	20.75	36.37
VGGNet + ft	50.23	59.31	45.16	20.95	43.87	38.94	32.72	26.19	39.78
DeepSentiBank [8]	55.00	35.29	51.09	11.85	15.41	3.23	55.45	8.85	35.67
PCNN(CaffeNet) [59]	36.41	38.33	36.62	17.47	29.86	23.02	40.60	23.27	32.50
PCNN(VGGNet) [59]	44.33	50.00	42.62	15.56	31.58	22.58	38.94	23.96	35.38
MKN_R (ours)	56.67	48.53	46.72	20.00	43.23	20.97	50.78	29.69	**43.02**
MKN_M (ours)	53.47	60.87	49.59	17.78	44.32	38.10	51.40	23.81	**44.26**
MKN_S (ours)	59.62	43.34	50.54	18.82	42.41	18.85	53.82	31.23	**44.45**

4.4 Different Combination Ways

In this section, our aim is to investigate how to combine both the models to generate robust sentiment representations.

We employ the proposed Algorithm 1 to aggregate the feature maps across all locations for the supervised model. We first choose the suitable kernel for each layer, it turns out that the Gaussian kernel is most beneficial compared to others. Then three pooling strategies, *i.e.*, MKL_R, MKL_S, MKL_M, are employed for comparison. Moreover, the multi-scale feature maps extracted from different convolutional layers are combined in different strategies, and the results are shown in the first eight rows of Table 3.

The first observation from Table 3 is a multi-kernel combination of non-topmost layers with the fc7 layer brings improvement for sentiment prediction, which also performs better than the concatenating way. As for the non-topmost layer, deeper neural units of the pooling layer are more discriminative while combined with fc7 for classification. The combination of pool3, pool4, pool5, fc7 achieves 62.29% on MKN_M and 62.44% on MKN_S, respectively. It proves that our method can extract more representative features for sentiment analysis when only the strong model is applied.

Then, both the weakly- and strongly-supervised neural networks are employed in our proposed multi-kernel method, and the results are summarized in the last five rows in Table 3. It's interesting to find that with the help of the model trained on the weakly labeled dataset, our method (63.92%) outperforms employing the strong model individually (59.61%) by a wide margin. Compared to the algorithm using strong model individually, the fc7 from the weakly-supervised model has fewer effects on the accuracy. Therefore, we propose to take the pooling layers from the weakly-supervised model into consideration. The overall accuracy of MKN_M and MKN_S outperform MKN_R, which proves that choosing the neural units with specific filters is better for sentiment analysis.

To validate the effectiveness of the three strategies for picking the efficient emotion, we also show the results of directly using the LIBSVM trained on the selected feature maps instead of employing our MKN method. Table 4 summarizes the results. For using the single layer from the strong model trained on the FI dataset, the random pooling provides a baseline for the comparison, the max and sum pooling achieve higher accuracy with the pooling features of deeper layers. When combined with the last fully connected layer, the performance of all the three strategies is better than using the non-topmost layer only.

Table 8. Performance of different layers from different models on the Comics dataset. Here, the strong model is trained on the Comics dataset, and weakly-supervised model is trained on the machine generated Flickr dataset. The mixed model denotes training with all the data.

Model	Layer	Amuse	Awe	Content	Excite	Anger	Disgust	Fear	Sad	All
Strongly Model	pool5	46.10	42.65	40.98	12.59	38.72	16.94	41.74	28.65	36.51
	fc6	47.67	51.47	43.44	15.56	36.09	26.61	41.74	26.04	37.87
	fc7	50.23	59.31	45.16	20.95	43.87	38.94	32.72	26.19	39.78
	pool5 + fc6	49.67	51.47	43.17	14.81	35.71	25.00	45.17	31.77	39.16
	pool5 + fc7	50.33	48.53	43.44	14.81	38.35	25.81	43.30	31.25	39.28
Weakly Model	pool5	43.07	44.12	39.89	20.03	30.45	12.90	46.42	25.10	35.33
	fc6	47.33	51.47	46.45	14.81	35.34	14.52	47.04	25.03	38.26
	fc7	44.33	45.59	44.54	14.81	31.20	12.10	42.68	22.40	35.27
	pool5 + fc6	48.15	48.53	41.53	17.78	33.83	17.74	42.68	21.88	36.34
	pool5 + fc7	49.52	52.94	45.36	14.07	33.46	16.13	42.37	22.92	37.08
Mixed Model	pool5	43.07	44.12	39.89	21.13	30.45	12.90	46.42	25.64	35.33
	fc6	49.67	47.06	40.44	20.74	31.95	16.13	43.30	25.52	36.68
	fc7	50.33	44.12	45.08	11.85	27.07	7.26	46.11	19.58	35.93
	pool5 + fc6	44.67	42.65	42.90	19.26	33.46	14.52	45.79	25.07	36.57
	pool5 + fc7	50.33	48.53	40.71	20.03	32.71	19.35	41.43	24.48	36.74

In addition, we train a mixed model based on both the well labeled and weakly labeled samples. Table 5 shows it outperforms the weakly-supervised model while underperforms the strong model, which demonstrates it is less efficient to directly combine samples from different sources together.

4.5 Results on the Comics Dataset

In contrast to the FI dataset in which the images are usually uploaded by web users and captured for the natural scene and people, we build a new Comics dataset to demonstrate our proposed method also works well on abstract and artistic images.

Following the emotional definition system derived from psychological study [32], we construct a well labeled Comics dataset, using *Amusement, Awe, Contentment, Excitement* as positive emotions, and *Anger, Disgust, Fear, Sadness* as negative emotions. These categories also correspond to the list of basic emotions that can be visualized in comics [31]. Started from popular comics of various countries such as America, Japan, China and France, about seventy comics are selected as our candidates, *e.g.*, Sponge Bob, Spiderman, The Avengers, One Piece, Slam Dunk, *etc.* Next, ten workers are chosen as the participants (5 females and 5 males; mean age=20.3), who view comics and then cut out the panels corresponding to one of the emotional categories. Finally, total 11,821 comic images are selected and roughly divided into Comic subset and Manga subset. The statistics of the Comics datasets are shown in Table 6. And Figure 4 shows the samples from the collected dataset. The former is comprised of European and America comics that are drawn in the realism style, while the latter is from Asian comics mostly consisting of abstract lines. These images vary in appearance and style, but all convey obvious emotions.

The Comics dataset is randomly split into subsets for training (80%), testing (15%) and validating (5%), in which the training set is used to fine-tune the strongly-supervised model. Similar to the experiments on FI, the Flickr dataset is used to train the weakly-supervised model. We evaluate the performance of the state-of-the-art algorithms as well as the proposed method on the testing set. The results are summarized in Table 7. Compared to the low-level features, the deep CNNs show a great advantage. The traditional simple color and texture features are not enough for predicting

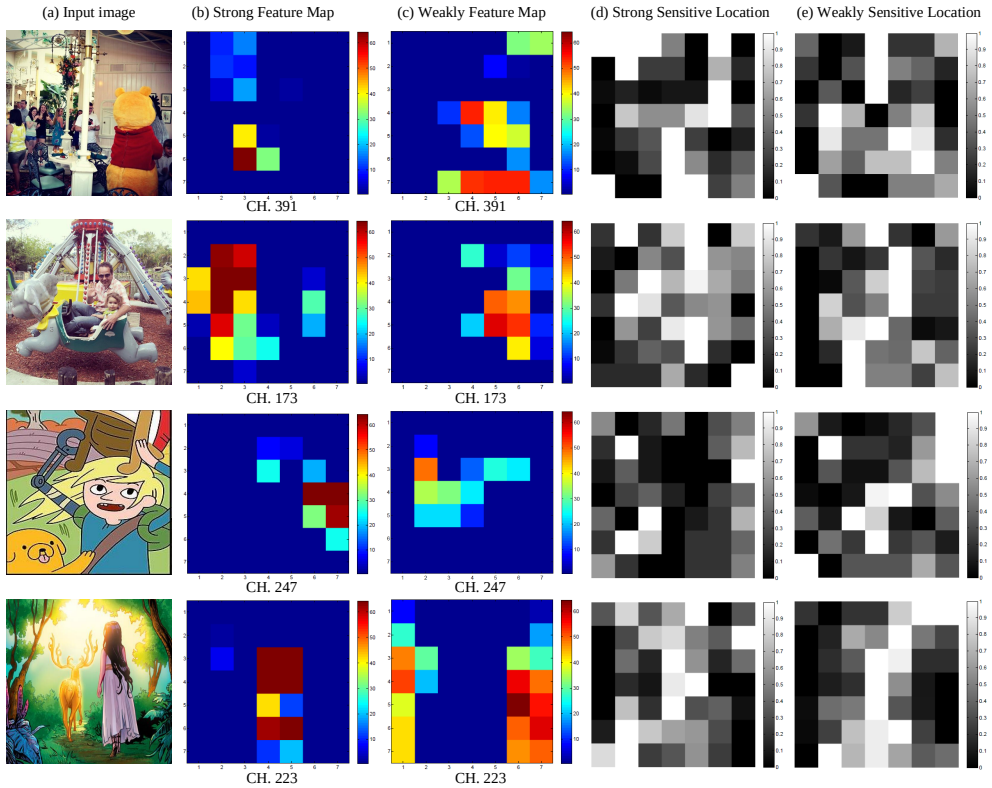

Fig. 5. Examples show the responses from the same layer (pool5) are complementary in different models.

emotions in comic images. After fine-tuning, the performance of CNN model is improved to 39.78% on VGGNet. Since our method can generate more representative sentiment features for images, it achieves the best result 44.45%. We also train a mixed model on all the weakly labeled and well labeled data. As shown in Table 8, the combination of training data does not perform as well as the combination of strongly- and weakly-supervised models.

4.6 Visualization

Figure 5 shows the examples from the FI and Comics datasets and the average feature maps from the corresponding strong model as well as the weakly-supervised model. The response locations are also shown in the last two columns.

For different images, the neurons corresponding to strongly- and weakly-supervised models focus on different regions. It is natural that the models are trained based on different samples. Furthermore, the weakly-supervised model can discover some sentiment regions which strong model may ignore. The proposed method captures these different responses from both models and further extracts discriminative representations for sentiment analysis.

5 CONCLUSION

In this paper, we introduce the challenging problem of visual sentiment prediction. Since such the abstract task adopted in the CNN framework have constraints, this paper designs a multiple kernel

method taking both strongly- and weakly-supervised models into consideration. We propose an approximate solution to dynamically find the important neural units from the multi-scale feature maps. The extensive experiments on benchmark datasets and the newly established Comics dataset show the superiority of our method.

6 ACKNOWLEDGEMENTS

This work was supported by the NSFC (NO.61876094), Natural Science Foundation of Tianjin, China (NO.18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Unaiza Ahsan, Munmun De Choudhury, and Irfan A. Essa. 2017. Towards using visual attributes to infer image sentiment of social events. In *IJCNN*.
[2] Soraia M. Alarcao and Manuel J. Fonseca. 2018. Identifying emotions in images from valence and arousal ratings. *Multimedia Tools Appl.* 77, 13 (2018), 17413–17435.
[3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM International Conference on Multimedia*.
[4] Victor Campos, Brendan Jou, and Xavier Giró i Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image Vision Comput.* 65 (2017), 15–22.
[5] Víctor Campos, Amaia Salvador, Xavier Giro-i Nieto, and Brendan Jou. 2015. Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In *International Workshop on Affect & Sentiment in Multimedia*.
[6] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intel. Syst. Tec* 2, 3 (2011), 1–27.
[7] Ming Chen, Lu Zhang, and Jan P. Allebach. 2015. Learning deep features for image emotion classification. In *IEEE International Conference on Image Processing*.
[8] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* (2014).
[9] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan Ying Chen, and Shih Fu Chang. 2014. Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia*.
[10] Yan-Ying Chen, Tao Chen, Taikun Liu, Hong-Yuan Mark Liao, and Shih-Fu Chang. 2015. Assistive Image Comment Robot – A Novel Mid-Level Concept-Based Representation. *IEEE Trans. Affect. Comput.* 6, 3 (2015), 298–311.
[11] Youngmin Cho and Lawrence K Saul. 2009. Kernel methods for deep learning. In *Annual Conference on Neural Information Processing Systems*.
[12] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
[13] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* 53, 4 (1987), 712.
[14] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. 2018. Emotional Attention: A Study of Image Sentiment and Visual Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[15] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2015. Compact bilinear pooling. *arXiv preprint arXiv:1511.06062* (2015).
[16] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision*.
[17] Alan Hanjalic. 2010. Extracting moods from pictures and sounds. *IEEE Sig. Proc. Mag.* 23, 2 (2010), 90–100.
[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*.
[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).
[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[21] Xuanyu He and Wei Zhang. 2018. Emotion recognition by assisted learning with Convolutional Neural Networks. *Neurocomputing* 291 (2018), 187–194.
[22] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. 2012. Can we understand van gogh's mood?: Learning to infer affects from images in social networks. In *ACM International Conference on Multimedia*.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*.

[24] Mingyuan Jiu and Hichem Sahbi. 2015. Semi supervised deep kernel design for image annotation. In *IEEE Int. Conf. Acou. Speech Signal Process.*

[25] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang Tuan Luong, James Z. Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and Emotions in Images. *IEEE Sig. Proc. Mag.* 28, 5 (2011), 94–115.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems*.

[27] Xin Lu, Poonam Suryanarayan, Reginald B. Adams, Jia Li, Michelle G. Newman, and James Z. Wang. 2012. On shape and the computability of emotions. In *ACM International Conference on Multimedia*.

[28] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. 2012. On shape and the computability of emotions. In *ACM International Conference on Multimedia*.

[29] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*.

[30] Julien Mairal, Piotr Koniusz, Zaïd Harchaoui, and Cordelia Schmid. 2014. Convolutional kernel networks. In *Annual Conference on Neural Information Processing Systems*.

[31] Scott Mccloud. 2007. Making comics: Storytelling secrets of comics, manga and graphic novels. *The Journal of Popular Culture* 40, 5 (2007), 890–892.

[32] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. 2005. Emotional category data on images from the international affective picture system. *Behavior Research Methods* 37, 4 (2005), 626–630.

[33] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[34] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury. 2018. Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In *European Conference on Computer Vision*.

[35] Kuan-Chuan Peng and Tsuhan Chen. 2015. Cross-layer features in convolutional neural networks for generic classification tasks. In *IEEE International Conference on Image Processing*.

[36] Kuan-Chuan Peng and Tsuhan Chen. 2015. A framework of extracting multi-scale features using multiple convolutional neural networks. In *IEEE International Conference on Multimedia and Expo*.

[37] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[38] Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. 2017. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* 261 (2017), 217–230.

[39] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. 2007. More efficiency in multiple kernel learning. In *International Conference on Machine Learning*.

[40] Tianrong Rao, Min Xu, and Dong Xu. 2016. Learning multi-level deep representations for image emotion classification. *arXiv preprint arXiv:1611.07145* (2016).

[41] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.

[42] Andreza Sartori, Dubravko Culibrk, Yan Yan, and Nicu Sebe. 2015. Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *ACM International Conference on Multimedia*.

[43] Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 640–651.

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

[45] Kaikai Song, Ting Yao, Qiang Ling, and Tao Mei. 2018. Boosting image sentiment analysis with visual attention. *Neurocomputing* 312 (2018), 218–228.

[46] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. 2006. Large scale multiple kernel learning. *J. Mach. Learn. Res.* 7, 7 (2006), 1531–1565.

[47] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. 2006. Large scale multiple kernel learning. *J. Mach. Learn. Res.* 7 (2006), 1531–1565.

[48] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen. 2016. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *IEEE International Conference on Multimedia and Expo*.

[49] Vladyslav Sydorov, Mayu Sakurada, and Christoph H. Lampert. 2014. Deep fisher kernels - End to end learning of the fisher kernel GMM parameters. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[50] Quoc-Tuan Truong and Hady W. Lauw. 2017. Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN. In *ACM International Conference on Multimedia*.
[51] Wang Wei-ning, Yu Ying-lin, and Jiang Sheng-ming. 2006. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *SMC*.
[52] Lifang Wu, Shuang Liu, Meng Jian, Jiebo Luo, Xiuzhen Zhang, and Mingchao Qi. 2017. Reducing noisy labels in weakly labeled data for visual sentiment analysis. In *IEEE International Conference on Image Processing*.
[53] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. 2018. Weakly supervised coupled networks for visual sentiment analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[54] Jufeng Yang, Dongyu She, Yu-Kun Lai, and Ming-Hsuan Yang. 2018. Retrieving and classifying affective Images via deep metric learning. In *AAAI Conference on Artificial Intelligence*.
[55] Jufeng Yang, Dongyu She, and Ming Sun. 2017. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In *International Joint Conference on Artificial Intelligence*.
[56] Jufeng Yang, Dongyu She, Ming Sun, Ming Ming Cheng, Paul Rosin, and Liang Wang. 2018. Visual Sentiment Prediction based on Automatic Discovery of Affective Regions. *IEEE Trans. Multimedia* (2018).
[57] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. 2017. Learning visual sentiment distribution via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence*.
[58] Quanzeng You, Hailin Jin, and Jiebo Luo. 2017. Visual sentiment analysis by attending on local image regions. In *AAAI Conference on Artificial Intelligence*.
[59] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI Conference on Artificial Intelligence*.
[60] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI Conference on Artificial Intelligence*.
[61] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. 2013. Sentribute: Image sentiment analysis from a mid-level perspective. In *International Workshop on Issues of Sentiment Discovery and Opinion Mining*.
[62] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*.
[63] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8, 4 (2018).
[64] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. 2016. Picking deep filter responses for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[65] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Approximating Discrete Probability Distribution of Image Emotions by Multi-Modal Features Fusion. In *International Joint Conference on Artificial Intelligence*.
[66] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Learning Visual Emotion Distributions via Multi-Modal Features Fusion. In *ACM International Conference on Multimedia*.
[67] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. 2018. Affective Image Content Analysis: A Comprehensive Survey. In *International Joint Conference on Artificial Intelligence*.
[68] Sicheng Zhao, Yue Gao, Guiguang Ding, and Tat Seng Chua. 2017. Real-Time Multimedia Social Event Detection in Microblog. *IEEE Trans. Cyber.* (2017).
[69] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia*.
[70] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. 2017. Continuous Probability Distribution Prediction of Image Emotions via Multi-Task Shared Sparse Regression. *IEEE Trans. Multimedia* 19, 3 (2017), 632–645.
[71] Sicheng Zhao, Hongxun Yao, You Yang, and Yanhao Zhang. 2014. Affective image retrieval via multi-graph learning. In *ACM International Conference on Multimedia*.
[72] Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer. 2018. EmotionGAN: Unsupervised Domain Adaptation for Learning Discrete Probability Distributions of Image Emotions. In *ACM International Conference on Multimedia*. 1319–1327.
[73] Honglin Zheng, Tianlang Chen, Quanzeng You, and Jiebo Luo. 2017. When saliency meets sentiment: Understanding how image content invokes emotion and sentiment. In *IEEE International Conference on Image Processing*.
[74] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. 2017. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. In *International Joint Conference on Artificial Intelligence*.
[75] Jinfeng Zhuang, Ivor W. Tsang, and Steven C. H. Hoi. 2011. Two-layer multiple kernel learning. *Journal of Machine Learning Research* 15 (2011), 909–917.