Image sentiment analysis based on hierarchical semantic concept ontology and convolutional neural network

Zichen Chao
Department of Computer Science
Shanghai JiaoTong University
Shanghai, 200240 China
zichen.chao@gmail.com

Honglei Guo
IBM Research - China
19 Zhongguancun Software Park
Beijing, 100193 China
guohl@cn.ibm.com

Zhong Su
IBM Research - China
19 Zhongguancun Software Park
Beijing, 100193 China
suzhong@cn.ibm.com

Yong Yu
Department of Computer Science
Shanghai JiaoTong University
Shanghai, 200240 China
yyu@cs.sjtu.edu.cn

ABSTRACT

More and more people enjoy sharing their emotion and sentiment through posting various images in the social network. Since limited explicit text description are provided by authors, it is a very challenging task to detect people's emotion and sentiment orientation hidden in various complicated images. Some previous existing works try to leverage the low-level visual features to predict sentiment, but they may not work when semantic content in images plays a crucial role in arousing sentiment. In this work we propose a novel visual sentiment analysis approach based on high-level semantic ontology and convolutional neural network (CNN) model, which can detect sentiment orientation of the images through analysis and understanding of semantic concepts presented in images. We first construct hierarchical sentiment-related concept ontology (HCO) from over 500k+ images using a statistic-based semantic mining approach. Then we build CNN models to classify these images in 3 layers of the HCO. CNN models in different layers serve as the classifier of certain layer and the pre-trained model for the next layer. Our main contribution is: firstly, we are the first to leverage CNN and HCO to handle general visual sentiment prediction on all kinds of complicated images in various scene, CNN gives great performance and interpretable results when combined with HCO. Secondly, by providing multidimensional classification, our proposed HCO framework indicates implicit steps in which people perceive sentiment on complicated scenes, and figures out main semantic topics related to sentiment shared in web. By combining HCO and CNN, our image sentiment prediction model are proved by experiments to distribute promising performance over different domains.

WWW '2015 Florence, Italy

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.6 [**Artificial Intelligence**]: Learning—*Concept Learning*; I.2.7 [**Artificial Intelligence**]: Natural language processing—*Text Analysis*

General Terms

Algorithms, Experimentation

Keywords

Image sentiment prediction, Image classification, Semantic, Concept Detection, Ontology, Social Multimedia, Convolutional Neural Network

1. INTRODUCTION

Nowadays, an increasing number of people are willing to share their sentiment and emotions in social network such as Twitter and Facebook. Many works has been proposed to extract sentiment from the text over the web. However, according to a recent study, about 36 percent of all the shared links on Twitter are images, which indicates that images are playing a more and more vital role in sentiment sharing and information changing. For example, people may post photos about a birthday party to express positive sentiment such as happiness and excitement. This trend leads the analysis of visual sentiment to be an active and interesting research area.

Many exsiting works [15, 14, 23, 25, 8] leverage low-level visual features to directly predict sentiment. In [14] it is believed that human perception and understanding of image sentiment is so subjective that it is supposed to be handled in *affective level*. In *affective level*, the analysis and prediction of sentiment mimics human intuition about images, and takes psychological and aesthetic components into consideration. Many psychological and aesthetic theories [5, 21, 10]

suggests a potential bridge between low-level features and sentiment. For example, warm-toned color may arouse positive sentiment. Thus these work mainly focus on designing powerful low-level features such as color and texture to build this bridge.

On contrast, some other existing works [2, 1, 26] suggest a huge *affective gap* between low-level features and sentiment, similar to the well-known *semantic gap* in content-based image retrieval(CBIR) between low-level features and image semantics. Hence, mid-level representations, which has been proved to distribute great performance bridging the *semantic gap*, are designed to fill the *affective gap* and infer the sentiment, and more importantly, make the prediction more interpretable. Thus, these works mainly concentrate on designing suitable mid-level representations that tightly related to sentiment semantic concept.

In this paper we propose a novel approach to determine the sentiment orientation hidden in the various images with rich semantic scenes and objects. We primarily focus on images with strong sentiment and aim to figure out whether the image would arouse positive or negative sentiment. We crawled images from Flickr, a widely-used photo sharing platform. By applying a statistic-based semantic mining approach to the meta data of the images, we crawled over 500k+ images and simultaneously construct Hierarchical Concept Ontology(HCO), a tree-structure ontology, which implies different steps and degrees in which people perceive sentiment and reveals the most popular sentiment-related semantic concepts in web. Meanwhile, A convolutional nearual network(CNN)[11] with 4 convolutional layers is designed. CNN has been leveraged to tasks like object and pattern recognition[11, 12, 13, 16]. However, few works apply CNN to tasks to recoginize high-level concepts. In our work, the CNN models are designed to recognize concepts in different layers of HCO. The CNN models are trained based on the structure of HCO and serve as classifiers in different layers and the potential pre-trained models for the next layer. In summary, our contributions are- **first**, a hierarchical sentiment-related semantic concept ontology founded by statistic-based web semantic mining; **second**, a convolutional neural network that is applied to general sentiment analysis for the first time; **third**, the combination of HCO and CNN which distributes promising sentiment prediction accuracy and gives interpretable results.

In the rest of the paper, we first discuss related work(Sec.2). Then a overview of our image sentiment model with the HCO and CNN(Sec.3) is presented. Next, the detailed introduction about hierarchical concept ontology for image sentiment prediction is discussed(Sec.4). We then elaborately introduce the CNN-based image sentiment prediction framework with HCO(Sec.5). Finally, a series of application and experiments of our approach in sentiment prediction over different domains is dicussed(Sec.6).

2. RELATED WORK

Many previous works [15, 14, 23, 25, 8]predict the emotions aroused by images mainly through extracting appropriate low-level visual features. Founded by related aesthetic and psychology theories, their systems were proposed to construct bridge between low-level features and *affective level*. Various kinds of basic component of images are extracted and combined to train the final sentiment classifier. In [15] color, texture, composition of the images are leveraged, and

Figure 1: Sample of images with similar low-level features arousing opposite sentiment. Although similar in terms of color distribution, the left image shows a terrible fire disaster, while the right image shows awesome sunrise, they arouse different emotions

their approach was proved to work fine in classifying art and abstract pictures, while in [14], shape features of the images were studied. In this paper, this type of works are refered as *low-level based approaches*.

On the other hand, some works are aligned with the development of semantic concept detections. They take advantages of the progress made in areas like concept detection or pattern recognition[3, 19, 6, 17], and design mid-level features to train their classifiers[2, 26]. Mid-level representations reveals the semantic features of the images, which may lead to various sentiment orientations. [2] defines adiective and noun pairs(ANP) as the mid-level representations. Each ANP is a sentiment-related concept, the adjective implies sentiment orientation while the noun indicate specific concept. Low-level features are extracted to train the detector of mid-level features, which serve as the training input of the final emotion classification. In, [26]102 pre-defined mid-level attributes were selected. In this paper, this kind of works are refered as *mid-level based approaches*.

In spite of great numbers of existing works, research about image sentiment prediction is still faced with big chanllenges. *Low-level based approaches* are potentially fill the *affective gap* in some cases, but it is likely that images with similar low-level features are in fact presenting opposite sentiments(See Fig.1), espically when dealing with large scale images shared in social web. In constrast, based on concept learning and detection, *mid-level based approaches* could perform stably fine when appied to images from all kinds of domains. Nevertheless, *mid-level based approaches* still show drawbacks in interpreting the classification result. Although founded by semantic concept, previous mid-level representations are defined either arbitrarily or subjectively, which lead the representation to be less interpretable. Another problem is about the sentiment metrics, due to the variety of psychological sentiment theories, previous works define vairious sentiment categories for their research, which may cause confusion. In this work, to avoid confusion and inconvinience, we concentrate on images reflecting strong sentiment, and aim to classify the images into two basic sentiment orientation – positive and negative.

3. IMAGE SENTIMENT MODEL WITH HCO AND CNN

In this work, we are eager to analyze the sentiment aroused by the images posted in social web. We are not talking

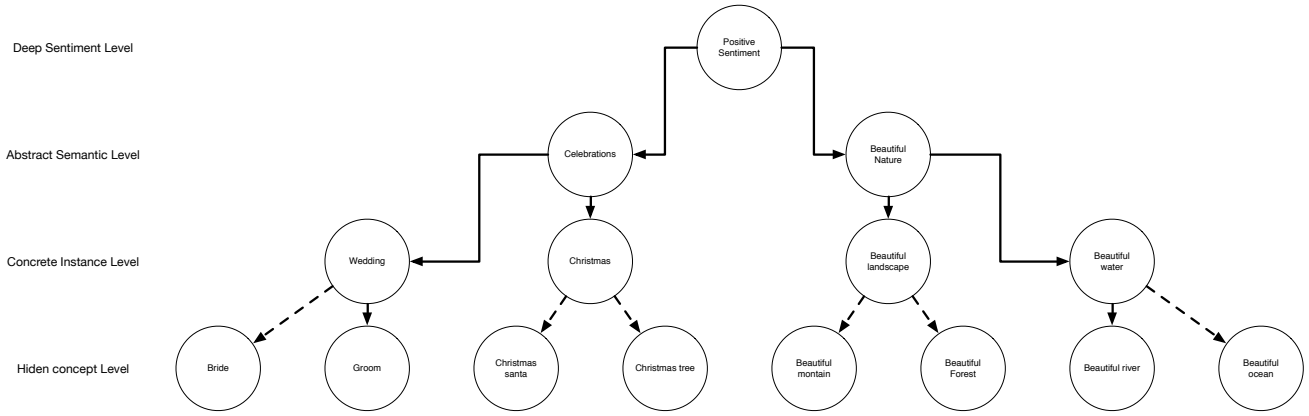

Figure 3: Illustration of the HCO structure, the concepts under the dashed lines are hiden to avoid redundacy and confusion. Only part of the HCO is presented here, the complete HCO consists of 2 concepts in deep sentiment level, 11 concepts in abstract semantic layer and 28 concepts in concrete instance level.

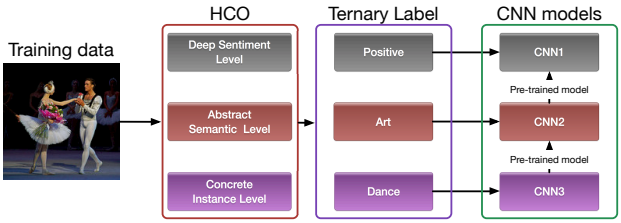

Figure 2: Overview of the our sentiment model intergrating HCO with CNN

about extract the face emotion image, instead, we consider all kinds of images in different scenes, and evaluate the sentiment presented by all the contents. However, the images in social web varies and are complicated. Moreover, since sentiment is subjective and complicated, merely analyzing the low level features of the image is not enough, which lead this task to be tough.

Despite the difficulties, we could still get some clues from some obeservations. Imaging you are browsing the pictures posted by your friends in Twitter. When you see a photo about a wedding, you may get happy, while if you see a picture about an earthquack, you get depressed. Let us consider how these emotions are aroused. In the first case, what we directly see might be that the bride is kissing the groom and surrounding people are smiling, then we realize this is a scene of a wedding, which could be categorize into *ceromony* or *celebration*, and celebrations always make people happy, that is the whole process. In social network, users post photos mainly in order to share their status and sentiment, which make the photos contain sentiment-related concepts. Therefore, for images posted in social web, we could constuct a common pattern for the sentiment arousing process like this. Thus, we introduce the HCO, a tree concept structure which helps us abstract the process of sentiment cognition.

To automatically extract the concepts in the HCO, we can leverage many approaches that work on object or pattern recognition. However, the problem is that the concepts in the HCO such as *wedding* might be far beyond a pattern or object. Thus, we introduce convolutional neural network(CNN), which has been proved to be effective in extracting hiden features. By leveraging CNN, we are likely to effectively extract certain concepts without explicitly designing features. Thus, CNN seems to be perfect for our task.

The essential part of our work integrates the HCO and CNN(See Fig.2). The HCO(hierarchical concept ontology) is a tree structure which abstracts the sentiment-related concepts in different cognitive level(See Fig.3). Large scale images crawled from the web are used to train our model. The images in the training set are assigned a ternary label based on the HCO, indicating the corresponding sentiment-related concept the images represent in different level. Next, 3 CNN models are trained sequently to serve as the classfiers in 3 levels. The model in higher level finetunes from the pretrained model in lower level. Our sentiment prediction system consists of all these 3 CNN classifiers. Given an image, the systems input it into the 3 classifiers, and obtains a ternary classification result. In summary, our system is trained in 3 levels, and is able to predict sentiment orientations, and the corresponding lower-level concept interpretation as well.

4. HIERARCHICAL CONCEPT ONTOLOGY FOR IMAGE SENTIMENT PREDICTION

Based on the senario described in Sec.3, our hierarchical concept ontology comes out(See Fig.3). In the 3-layers tree sturcture, each leaf node stands for a relativelt concrete concept that represents strong sentiment, like *wedding* in our previous example, this level are refered as **Concrete Instance Level**. The upper nodes give abstractions over all the kid nodes, like that *celebration* include *wedding* and other concepts, namely, *new year* and *birthday*, we refer this level as **Abstract Semantic Level**. Finally, the root of the tree, or the top level, is called **Deep Sentiment Level**, which abstracts all the concepts that arouse certain sentiment. In fact, we can still create nodes below the concrete instance level, for example, *bride* and *exchanging rings* under *wedding*. However, these nodes are omitted to keep the structure less redundant and highly abstracted.

4.1 Key Semantic Concept Extraction in the Image

Table 1: Example of top-ranked tags

Base keywords	Top-related tags
happy	birthday, flower, smile
sad	crying, alone, child
fear	dark, death, scared
disgust	blood, food, gross

In order to construct HCO which presents sentiment-related concepts in different levels, we firstly need to extract various related concepts from large scale web image sets to support the HCO. The steps for the concept extraction could be summarized as following:

Input: image set with metadata (tags, descriptions, title)

Step 1: Tag ranking: For each tag, we count its occurrence frequency over all the metadata of the images, and then rank all the tags based on the frequency

Step 2: Tag filtering: wordnet2 is leveraged to filter out meaningless tags and we pick up top-n ranked tags to be our candidate tags.

Step 3: Tag scoring and reranking: we randomly crawl images for each tag, and annotators label the images with these candidate tags based on 2 criterions: **first**, the image strongly arouse the corrsponding sentiment; **second**, the image is related to the *base keyword*. We then select the final top n tags according to their score. The score for a certain tag x is

$Score(x) = \sum_{i \in V} 1 * \{\sum_{j \in Img(x)} 1 * \{S_{i,j}(x) = (1,1)\} > 7\}$

(V:set of all the annotators, $Img(x)$:the image set of the tag x, $1 * \{P\}$ equals to 1 if P is satisfied, otherwise, 0). We put the result tags into **Step 1** and repeat all the steps.

Since our goal is to determine the sentiment orientation aroused by images, we select some keywords to serve as the **base keywords** in our mining process. Specifically, based on the 7 fundamental human emotions(*angry, disgust, netural, fear, happy, surprise, sad*) we select 4 from them which represent strong emotions – *happy* for positive and *angry, disgust, sad* for negative. We refer positive and negative sentiments as the **base concepts**, which are leveraged later in the construction of the HCO

Base keywords serve as the query to get metadata of the images, we extract metadata of 4000 images with Flickr API[1], including human-labeled tags and the title and descriptions. Then, a tag ranking mechanism is implemented. For each tag, we count its occurrence frequency over all the metadata of 4000 images, and rank all the tags over the frequency. Next, *wordnet*[2] is leveraged to filter the tags since there are some noise in the raw tag list, such as some meaningless tags. We only keep the tags that can be returned by *wordnet*. Thus, we get a list of tags which is potentially tightly related to our base concepts, we pick up the top-100 ranked tags to be our candidate tags.Then we randomly crawl 10 images for each candidate tag, and build an annotated dataset based on the label criterion. Each annotator score certain image as (s_1, s_2), where $s_1, s_2 \in \{0, 1\}$, indicating if the image meets certain criterion. From the annotated dataset, we calculate the score for each tag using forluma in **Step 3** and get the final top n tags

Through this process, we could filter out a great number of tags which may not perform well when search for sentiment-related images which present the base concepts. For example, *Canon* and *Nikon* are always top-ranked tags but they are obviously not what we expect. We finally get a highly-selected list of tags(See Table.1). These tags are combined with their original base keywords, and are leveraged to be our new base keywords. Then the previous steps recurs, which means all these tags are used to extract new tags. Finally, after 4-5 recursions, we obtain over 100 keywords.

In summary, during this period, based on a recursive mining approach, we extract a series of keywords from the web that arouse corresponding sentiment. Taking objective statistics and human evaluation into consideration, this approach provides a stable fundament for the our future training process.

4.2 HCO Construction for Image

The construction of the HCO is based on many observations. We get some inspirations from our previous recurisive mechanism. By recursively set the base keywords, we can easily get a tree-like concept structrue from the process. The root of the tree is the initial base concept, without loss of generality, we set the base concept to be *positive sentiment*, the corresponding base keyword is *happy*. In the first iteration, the top-3 ranked tags are *birthday, flower, smile*, then these concepts are the child node of the root, and they serve as base keywords for the new iteration. Thus, we could potentially build a tree structure based on our previous process.

However, one of the problem is that it is likely that a keyword occurs in various layers, which breaks the tree-structure. To solve this problem, annotators are asked to score the keywords from 1 to 3 based on how abstract the keyword is(1 for the highest abstraction), then the final score for the tag x defined as:

$Score(x) = argmax(sum(i))$

where $sum(i) = \sum_{j \in V}\{1 * \{S_j(x) = i\}\}$

Here, V stands for the set of annotator, $S_j(x)$ represents the score annotator j gives to x. Some of the label results are shown in table.

Thus, keywords are classified into 3 layers. Then we build our tree-structured concept ontology based on the following ideas: **Connect**, Build a direct edge from i to j if i is one layer upper than j and i severs as base keyword to obtain j(See Fig.4(a)); **Merge**, merge i and j if they are in the same layer and they have a common child(See Fig.4(b)(c)). To make this idea more clear, we provide the following pseduocode:

```
for (i, j) in keyword_set{
        If (layer(i)+1 == layer(j)
            && Base(i, j) == true){
                Connect(i, j)
        }
}
for layer in (3, 2, 1){
        for (i, j) in layer{
                if has_common_kid(i, j){
                        merge(i, j)
                }
        }
}
```

[1]https://www.flickr.com/services/api/

[2]http://wordnet.princeton.edu/wordnet/

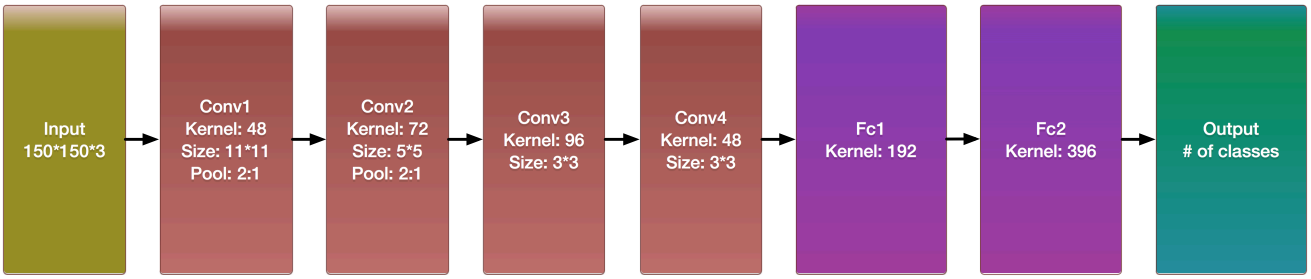

Figure 5: Our CNN model with 4 convolutional layers and 2 full-connected layers

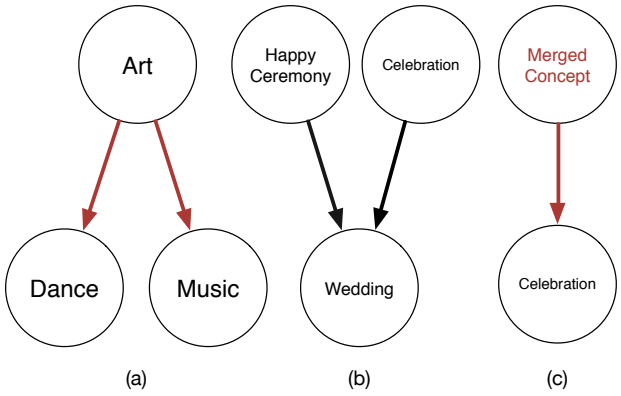

Figure 4: a. Example of Connect. *Dance* and *Music* are obtained when *Art* serves as the base keywords. Thus, edges are build between *Art* and others. b. Example of Merge. *Happy Ceremony* and *Celebration* have one common child *Wedding*, thus, these two concepts are merged(c)

Through the above process, we are able to get a forest structure, in which all the nodes(keywords) arouse same sentiment, and each 3-layers tree represents distinct sentiment-related concept, so do all the subtrees. The concept represented by each node abstract all its child, or subtrees. Finally, we connect all the trees to a super node, which stands for the concept of sentiment. We also hiden all the nodes in the bottom layer to make the structure simpler. Therefore, our final hierarchical concept ontology comes out. We refer the layers from top to bottom as **Deep Sentiment Layer**, **Abstract Semantic Layer**, and **Concrete Instance Layer**. Each image in our ontology is assigned a 3-dimensional label, which indicates its classification in each layer. The HCO not only provide an ontology for certain sentiment orientation, but also maps the images in various semantic concept level, leading the sentiment classification more interpretable.

5. CNN-BASED IMAGE SENTIMENT PREDICTION FRAMEWORK WITH HCO

Being one of the most popular deep learning framework, convolutional nerural network(CNN) have been proved to be highly effective and promising when applied to image classification and regression. Many works introduce the fundanmental structure and components of deep learning and CNN[11, 7], and the corresponding performance over various image classification problems as well. Nevertheless, previous works mainly leverage CNN to classify image in object level, such as pattern and object recognition. Few CNN-based approaches has been proposed to predict image sentiment. In this section, we innovatively suggest a 4-convolutional-layers CNN to classify the images into distinct sentiment orientations, and then introduce how the CNN models are well combined with the HCO to generate an interpretable result for sentiment prediction.

5.1 CNN classification model for complicate image sentiment

Our CNN has 4 convolutional layers, the structure of which is illustrated in Fig.5. Some approaches get great performance over 3-layers CNN, but our task is to predict sentiment in complicated images, which is more tougher, thus we increase the number of the convolutional layers. Our CNN model is trained based on the HCO so that the classification over different concept layers could be realized. The size of input color image is 150× 150, while the output size varies among the layers in the HCO. Different configuration of the network is tried, including the number of the layers, the number of the filters in each layer, and the kernel size for the filters. Eventually, we obtained the most promising structure with 4 convolutional layers, and the number of filter for each layer are 48, 72, 96, 192, respectively. We also introduce 2 full-connected layer to better train the model. Moreover, 2 max-pooling layers is set after the first 2 convolutional layers to make the data less redundant.

Two basic issues related to convolutional neural network are *forward pass estimation* and *back-propagation* The training process of the CNN model can be illustrated as follows: **first**, the input images are passing forwardly through the whole network and obtain a classification prediction. The results is compared to the ground truth label, and a loss value of the estimation is computed based on a predefined loss function. In our case, the softmax loss function is used, which is defined as:

$$l(x^n, y^n) = -log(\frac{e^{Wx^n_{y^n}+b}}{\sum_j e^{Wx^n_j+b}})$$

second, the loss value is used to update the the parameters backwardly. After several iterations, the loss would converge to a minimal value, which means that the classification concept has been learned.

In forward pass estimation of our model, the images are processed and computed sequently in 4 convolutional layers, which is the most essential part of a CNN. A group of filters are set in each layer to process the input images and generate output images. The output images of layer i serve as the

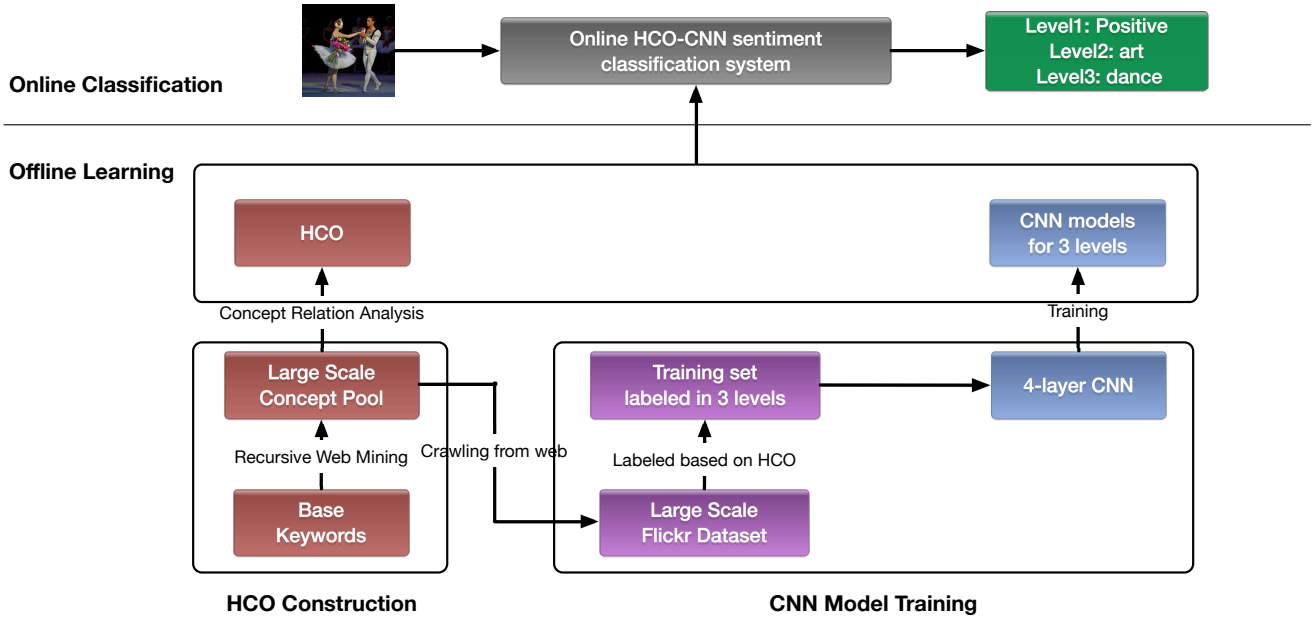

Figure 6: Overview of the proposed main framework for the construction of the HCO and the training of the CNN models, and the relevant application as well

input images of layer $i+1$, so on and so forth.

Apart from the convolutional layers, some other components are also crucial. Rectified Linear Units(ReLU) replaces the traditional tanh function to serve as a non-linear activation function, which reduces the operation time. Pooling layers are introduced to down-sample the output images. For example, by generate a single pixel out of 4 adjacent pixels, the size of the data is reduced. Moreover, normalization layers redistributes the local areas of the image to balance the image values and the layer weights.

In backward propagation, the parameters are refined to minimize the loss value. Specifically, the conlutional filters are adjusted, which are initially set as standard guassian filters. Stochastic gradient decent(SGD) is inplemented to realize the refinement for the weights of the filters. Generally, for a loss function f, the update rule based on SGD is defined as:

$$x^{(t)} = x^{(t-1)} - \eta \bigtriangledown f(x^{(t-1)})$$

where $x^{(t)}$ indicates the parameter at time t, η is the step size. Another important factor is the learning rate. By adjusting the learning rate, we can control the step size in SGD, so that the pace of the convergence could be managed. Generally, large learning rate are set firstly to quickly learn the basic concept, then the learning rate is slowed down to guarantee the accuracy. We present detailed analysis over learning rate in experiments in Sec.6.

5.2 CNN and HCO combination framework

Based on HCO-CNN model, the framework for image sentiment prediction is designed as Fig.6. Our main framework could be divided into 2 parts – the offline learning, and the online sentiment classification. In the offline learning period, the HCO is constructed and the CNN models are trained, which serve together as our HCO-CNN sentiment classification system. In the period of online sentiment classification, test images from all domains are input to our HCO-CNN system, and the HCO-CNN system returns a ternary classification for certain image, indicating the concept it presents in different levels.

The first part of the offline learning period is HCO construction. We start with some base keywords, crawling the corresponding keywords from the web. By using a recursive tag mining approches, we obtain a large scale sentiment-concept pool. Next, the concepts are analyzed in terms of their correlation and the degree of abstraction. Finally, the HCO is constructed, which is a 3-layer tree structure, revealing different levels in which people percive sentiment concepts.

The next part is training the CNN models. Images are crawled from the web by using the large scale sentiment concept pool. These images are then assigned a ternary label based on the HCO. We use these images to train 3 CNN models for each concept level in the HCO, and meanwhile, the CNN models in lower level serves as a pre-trained model for the model in higher level. Finally, the HCO as well as the 3 CNN models constitute our HCO-CNN sentiment classification system.

In the online sentiment classification period, we introduce images from different domains to our HCO-CNN system. For certain image, each of the 3 models classifys the image into a class in certain level, thus the image obtain a ternary classification, which suggests the sentiment interpretation for the image in different levels.

6. EXPERIMENTS

In this section, we firstly leverage the Flickr testing set to realize a validation within the image domain, the total classification accuracy over 3 HCO layers are evaluated, and we particularly analyze the performance including precision, recall and F-scores of the model in deep sentiment level. Then we introduce some evaluation cross the domain to predict binary sentiment orientation on a Twitter dataset and an

Table 2: Data settings in the experiments

	Positive	Negative	total
Training(Flickr)	278349	154836	433185
Testing(Flickr)	69623	38731	108354
Testing(Twitter)	463	133	599
Testing(Artistic)	378	428	806

Table 3: Total accuracy of Base-CNN and HCO-CNN over 3 levels

	B-CNN	HCO-CNN
Deep Sentiment Level	0.79	0.81
Abstract Semantic Layer	0.66	0.68
Concrete Instance Level	0.60	-

artistic photograph dataset, the accuracy of the binary sentiment classification and the comparision with the state of the art is discussed.

6.1 Experiment settings

We crawled 500k+ images from Flickr based on the construction of the HCO. According to the HCO, the images are not just labeled in terms of binary sentiment orientation. Each image is assigned a 3-dimension label in terms of the concept level. Thus, we trained 3 CNN models to serve as 3 classifiers in different levels. The Flickr dataset is splited into 2 part, serving as the training and testing set. The ratio of training size to testing size is 4:1. The ratio of positive size to negative size is about 1.79. All the data settings is shown in Table.2. For the implementation of the CNN, we use the CAFFE lib[9], and a Geforce GTX 460 gpu helps accelerate the training.

6.2 Experiment results

6.2.1 performance of HCO-CNN model

Based on the HCO, the number of the conpepts in each layer are 2, 11 and 28. Thus, the output number for the CNN in each concept layer equals to these number respectively. We set the basic learning rate to be 0.001, and slow down it to 0.0001 after 100000 iterations. We firstly train the CNN models in different HCO layers sperately, which are refered as B-CNN(Base-CNN) models, and then use the B-CNN models in lower levels to serve as pre-trained models to training another set of models in higher levels, which are refered as HCO-CNN models. The classfication accuracy in different concept layers are showed in Table.3. The accuracy in *abstract semantic level* for HCO-CNN is about 68% and the accuracy for the HCO-CNN in *deep sentiment level* reaches to 81%, indicating that our HCO-CNN model could give high accuracy over all 3 levels in the HCO, and HCO-CNN also outperforms B-CNN in different levels. It may cause confusion since that HCO-CNN distribute even higher accuracy in *deep sentiment level*. But since the number of class in *concrete instance level* is much more than in *deep sentiment level*, this result is reasonable. More particularly, we further investigated the detailed performance including precision, recall and F-scores in deep sentiment level obtained by B-CNN and HCO-CNN(See Table.4). In Table.4, the precison, recall and F-scores for HCO-CNN models are better than the B-CNN models, in average, an enhancement of 2.5% from 0.79 to 0.81 in F-score is obtained. Moreover,

Table 4: Detailed performance of Base-CNN and HCO-CNN over deep sentiment level

B-CNN	Precision	Recall	F-score
Positive	0.81	0.89	0.84
Negative	0.76	0.63	0.69
Total	0.79	0.79	0.79
HCO-CNN	Precision	Recall	F-score
Positive	0.85	0.84	0.85
Negative	0.72	0.73	0.73
Total	0.81	0.81	0.81

HCO-CNN gives more balanced precision and recall. Particularly, it is true that people tend to post positive images to social web rather than negative ones, thus the prediction for negative sentiment is tougher since the images are more complicated and diverse. But with HCO-CNN, the test over negative set gain much improvement, the recall increases 16% from 0.63 to 0.73 and the F-score increases 6% from 0.69 to 0.73(The enhancement of HCO-CNN over B-CNN is defined as score(HCO-CNN)-score(B-CNN)/score(B-CNN)). Therefore, we can conclude that HCO-CNN gain a better overall performance than B-CNN in determining the sentiment orientation of certain image.

Additionally, we evaluate the loss change during the training periods. The loss value is used to update the parameters backwardly. After several iterations, the loss would converge to a minimal value, which means that the classification concept has been learned. Since B-CNN supports HCO-CNN by providing pre-trained models, we firstly illustrate the loss change of training process of B-CNN in different levels in Fig.7-Fig.9, which present the loss value change within 200000 iterations in concrete instance level, abstract semantic level, deep sentiment level, resepectively. We can see from the figures that the loss value fluctuates during the process, but with a tendency to drop gradually. Moreover, we can figure out a relatively drastic drop around the 100000th iteration, which corresponds to our reducing the learning rate to 0.0001. This indicates our correct settings for the learning rate, which enables the training process to run fluently and quickly, without being stuck at some point far away from convergency. In fact, we try different settings for the CNN, and finally find the best learning rate for sentiment classification is between 0.001 and 0.00001. As for the loss curve for deep sentiment level(Fig.9), we can still see a tendency of dropping, but the process is much more slow, which indicates that learning the concept of certain sentiment orientation is tougher than recognizing the concepts in low levels.

On the other hand, The loss curves of the HCO-CNN from lower levels are shown in Fig.10-Fig.11, in which the the curve of the corresponding CNN models are presented to compare. We only record the first 100000 iterations. It turns out that, with the pre-trained B-CNN model from the lower level, the HCO-CNN models in the high level learn much more faster, the loss value could quickly converge to a small number. For deep sentiment level, the loss value even drops to less than 0.1, which could not be accessed without the pre-trained B-CNN model from abstract semantic level. We can conclude from the above observation that the idea of the HCO enables the HCO-CNN to quickly learn the sentiment-related concept in different levels, and benefits the

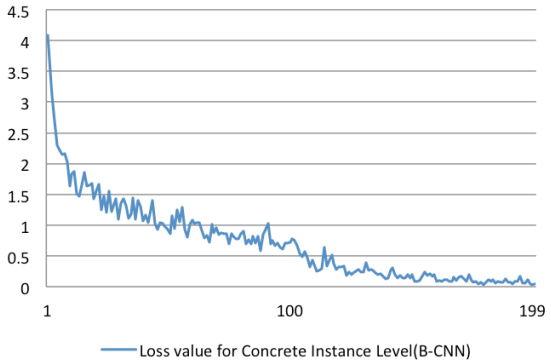

Figure 7: Loss changing of B-CNN in Concrete Instance Level(We record the loss every 1000 iterations, the loss values before 200000 iters are shown)

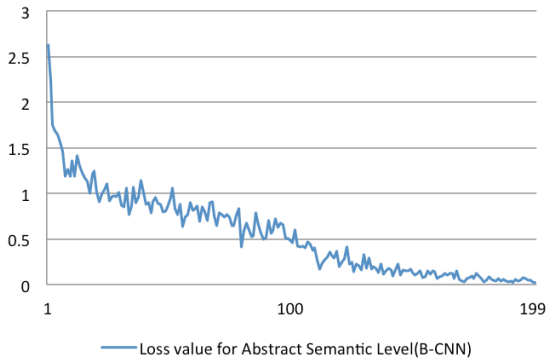

Figure 8: Loss changing of B-CNN in Abstract Semantic Level(We record the loss every 1000 iterations, the loss values before 200000 iters are shown)

performance as well.

In summary, supported by pre-trained B-CNN models, our HCO-CNN learn quickly in 3 concept levels of HCO, and give promising accuracy for sentiment prediction(81%) and corresponding interpretations(68%).

6.2.2 *Effectiveness of HCO-CNN model cross domains*

In this part, we evaluate the performance and rubostness of HCO-CNN in a new domain which is different from the training domain. We use the HCO-CNN model in *sentiment* level to implement sentiment prediction for images across domains. [2] published a group of data from Twitter, in which the images are labeled as positive and negative sentiment. The dataset consists of the images and their corresponding text metadata, and the ground truth is obtained by introducing Amazon Mechanical Turk(AMT) evaluation, each Turker is asked to label the image in terms of the image itself, the text, and the combination of image and text, and the final labels integrate all the 3 parts. [2] predict the sentiment by training both text and images, but in our work we only consider image sentiment prediction.

We do experiments on both the low-level based approach stated in [15] which leverage color, texture, composition as the features, and the mid-level based approach using

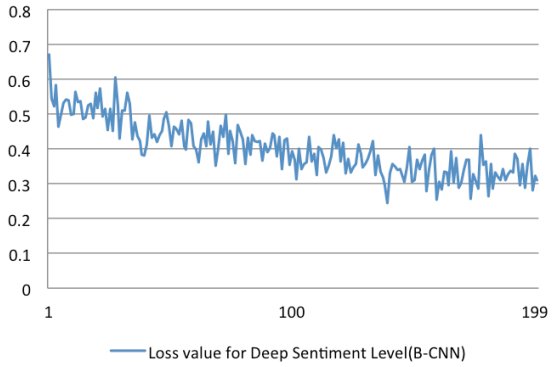

Figure 9: Loss changing of B-CNN in Deep Sentiment Level (We record the loss every 1000 iterations, the loss values before 200000 iters are shown)

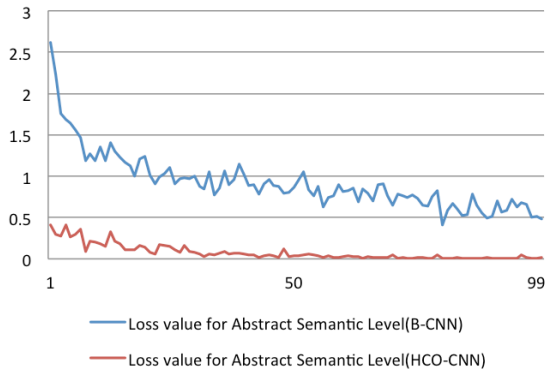

Figure 10: Loss changing of B-CNN and HCO-CNN in Abstract Semantic Level(We record the loss every 1000 iterations, the loss values before 100000 iters are shown)

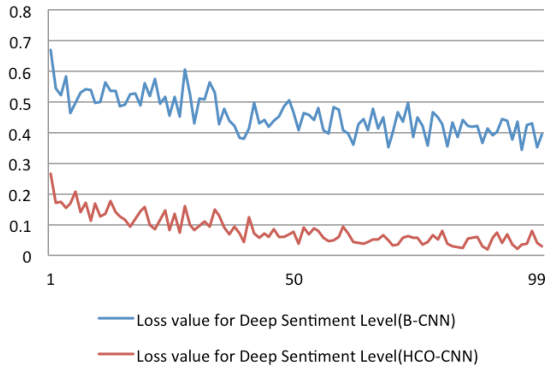

Figure 11: Loss changing of B-CNN and HCO-CNN in Deep Sentiment Level(We record the loss every 1000 iterations, the loss values before 100000 iters are shown)

Table 5: Total accuracy over the Twitter set using low-level based approach, mid-level based approach and our HCO-CNN approach, the enhancement of HCO-CNN over method i is defined as (accuracy(HCO-CNN)-accuracy(i))/accuracy(i)

Methods	Accuracy(%)	Enhancement(%)
Low-level(SVM)	55	33
Low-level(LR)	67	9
Mid-level(SVM)	57	28
Mid-level(LR)	70	4
HCO-CNN	**73**	-

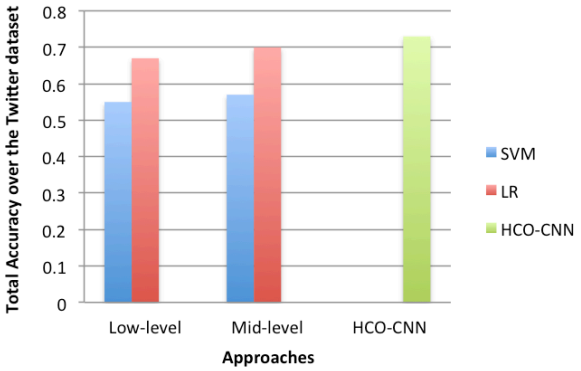

Figure 12: Total accuracy over the Twitter dataset obtained by low-level, mid-level and HCO-CNN approaches

Adjective-Noun as mid-level features proposed by [2]. Both these two works are representative methods to predict sentiment. For each of the approach, either linear SVM and logistic regression(LR) are leveraged for training to find the best classifier for each approach, the experiment data of these works are refered from [2]. We finally use our HCO-CNN to predict the sentiment. The perfomances are presented in Table.5 and Fig.12. It turns out the low-level based approach shows its weakness in analyzing complicated images in socail web, since it just involves the image-level features without considering relative semantic concepts. As for the mid-level based approach, even if it includes some semantic related features to enhance the performs, it faces difficulties in finding the hiden sentiment behind the complicated images, because no systematic semantic levels like HCO are extracted. In general, training over in-domain data are more likely to get better performance since the system could figure out the data distribution better. However, according to the results, even if our model is trained on Flickrs, which differs from the domain of the test images – Twitter, we still give better performance than another two approaches which actually implement a in-domain training, which suggests the robust performance of our HCO-CNN system over various domains. Specifically, compared to the highest accuracy other approches give, the HCO-CNN system give an enhancement of 4% on mid-level approach, while an enhancement of 9% over low-level approach.

6.2.3 *Effectiveness of HCO-CNN model on complicate artistic photographs*

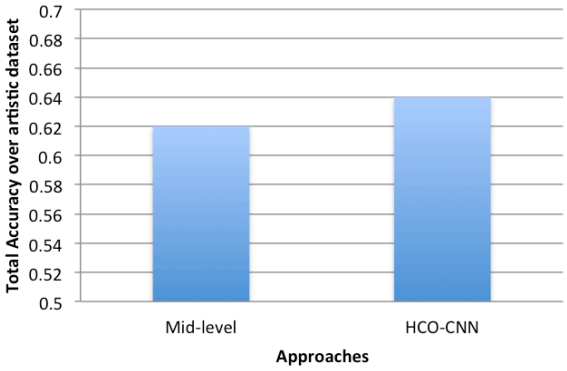

Figure 13: Total accuracy over the artistic dataset obtained by mid-level, HCO-CNN approaches

In Sec6.2.2, our models is proved to work fine on dataset from different social web. We are now introducing a tougher task.

In this experiment, we evaluate the capability of HCO-CNN on more complicate artistic images. 806 art photos crawled from the art share site[3] published by [15] are employed as test dataset. These photos are obtained by using the emotion categories as the search terms and are categorized into 8 emotion classes(*amazement, anger, awe, contentment, disgust, excitement, fear, sad*). To map these emotions to the binary sentiment orientations, we leverage *Sentiwordnet*[4], which gives a ternary score for each query in terms of positivity, negativity and objectivity of the word. The 8 emotion word is mapped to positive if the positivity score is larger than negativity, vice versa. Thus the art photo dataset is splited into positive and negative set.

We still leverage the mid-level approach proposed in [2] to compare with our work. The mid-level approach uses SVM as the classifier. The total accuracy is reported in Fig.13. Generally, the art photos are relatively more abstract and complicated than the social web, and the content of the images varies. Thus, the training data of our HCO-CNN model differs a lot from the artistic photos. Despite of these differences, it turns out that our HCO-CNN models still outperforms the in-domain trained mid-level approach, specifically, an enhancement of 2 points from 0.62 to 0.64 is gained. This again indicates the rubostness and the powerfulness of our HCO-CNN framework to extract sentiment-related concepts from complicated images across different image domain.

7. CONCLUSIONS

In this work we propose a novel framework to predict the sentiment orientation of the images shared in social web. A statistic based web mining approach is leveraged to extract various sentiment-related concepts. A hierarchical concept ontology(HCO) is then constructed. By applying convolutional neural network to the HCO dataset, we obtain a HCO-CNN model to predict sentiment orientations of images from various domain, which outperforms most state of the art. Meanwhile, the HCO reveals the most popular sentiment

[3]www.deviantart.com
[4]http://sentiwordnet.isti.cnr.it

related concept, and interprete the sentiment prediction in different concept levels.

Furthermore, some exciting directions are open for investigation. On the one hand, the HCO could still be expanded whether in width or height, so that a wider range of concepts and a more accurate concept relation structure could be obtained. On the other hand, the trained CNN model has the potential to be the pre-trained model serving different sentiment classification, so that more detailed sentiment categories could be analyzed.

8. REFERENCES

[1] D. Borth, T. Chen, R. Ji, and S.-F. Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 459–460. ACM, 2013.

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

[3] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, et al. Ibm research and columbia university trecvid-2011 multimedia event detection (med) system. In *NIST TRECVID Workshop*, 2011.

[4] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. Predicting viewer affective comments based on image content in social media. In *Proceedings of International Conference on Multimedia Retrieval*, page 233. ACM, 2014.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[7] G. E. Hinton. Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10):428–434, 2007.

[8] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can we understand van gogh's mood?: learning to infer affects from images in social networks. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 857–860. ACM, 2012.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

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

[12] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *Neural Networks, IEEE Transactions on*, 8(1):98–113, 1997.

[13] S.-C. B. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun. Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7):1201–1214, 1995.

[14] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang. On shape and the computability of emotions. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 229–238. ACM, 2012.

[15] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010.

[16] C. Nebauer. Evaluation of convolutional neural networks for visual recognition. *Neural Networks, IEEE Transactions on*, 9(4):685–696, 1998.

[17] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

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

[19] C. G. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2008.

[20] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1891–1898. IEEE, 2014.

[21] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.

[22] W. Wang and Q. He. A survey on emotional semantic image retrieval. In *2008 15th IEEE International Conference on Image Processing*, pages 117–120, 2008.

[23] W.-N. Wang and Y.-L. Yu. Image emotional semantic query based on color semantic description. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 7, pages 4571–4576. IEEE, 2005.

[24] X. Wang, J. Jia, P. Hu, S. Wu, J. Tang, and L. Cai. Understanding the emotional impact of images. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1369–1370. ACM, 2012.

[25] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 4, pages 3534–3539. IEEE, 2006.

[26] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM, 2013.