Learning Socially Embedded Visual Representation from Scratch

Shaowei Liu, Peng Cui, Wenwu Zhu and Shiqiang Yang
Computer Science Department, Tsinghua University, China
Tsinghua National Laboratory for Information Science and Technology
liu-sw11@mails.tsinghua.edu.cn, cuip/wwzhu/yangshq@tsinghua.edu.cn

ABSTRACT

Learning image representation by deep model has recently made remarkable achievements for semantic-oriented applications, such as image classification. However, for user-centric tasks, such as image search and recommendation, simply employing the representation learnt from semantic-oriented tasks may fail to capture user intentions. In this paper, we propose a novel Socially Embedded VIsual Representation Learning (*SEVIR*) approach, where an Asymmetric Multi-task CNN (*amtCNN*) model is proposed to embed user intention learning task into semantic learning task. Specifically, to address the sparsity and unreliability problems in social behavioral data, we propose to use user clustering, reliability evaluation, random dropout in output layer in our *amtCNN*. With its the partially shared network architecture, the learnt representation can capture both semantics and user intentions. Comprehensive experiments are conducted to investigate the effectiveness of our approach in applications of user favoring prediction, personalized image recommendation, and image reranking. Compared to the state-of-the-art image representation techniques, our approach achieves significant improvement in performance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation, Performance

Keywords

user intention, user behavior, multi-task CNN, image search and recommendation

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806247.

1. INTRODUCTION

Image representation learning is a fundamental task in various image-related applications. In past years, a tons of research works have been conducted to learn image representations with the goal of bridging the semantic gap between low-level features and high-level semantics. The learnt image representations have gained great success in content-centric applications, such as image classification, object detection, motion tracking, etc. In recent years, user-centric image platforms (such as Flickr and Pinterest) are more and more prevalent, where images are generated, disseminated and consumed by users. How to proactively provide images according to users' intentions (*e.g.*, interests) has become a critical problem, which subsequently posits a significant challenge to image representations: besides semantics, can image representations also reflect user intentions?

Recently, a paucity of works attempt to embed social signals into visual representations to capture user intention, where social signals such as user behaviors and social relations are commonly used to capture user intentions [15, 28]. They usually learn a linear or kernel variation of traditional hand-crafted descriptors, such as SIFT, HOG, bag-of-features representations, *etc*, to make the learnt representations consistent to social signals. However, these methods can hardly bridge the gap between low-level features and high-level user intentions because the hand-crafted features pose serious limit on how much intention information can be embedded. These hand-crafted features are designed for identifying image semantics, and cannot work well in reflecting user intentions, because these two factors cover different aspects of image contents.

Then can we learn image representations incorporating both semantics and user intentions from scratch rather than hand-crafted features? The recent progress in deep learning models, such as Convolutional Neural Network, offers an optimistic answer. These methods have demonstrated their superiority in learning image representations from scratch and attained significant improvement than hand-crafted features. However, existing deep learning based methods are usually designed for semantic-oriented tasks, *e.g.* image classification. How to embed both semantic-related information and intention-related information into deep models is still an unexplored problem.

Learning image representations to capture both semantics and user intentions still entail following challenges:

- **Multiple and asymmetric learning tasks.** In user-centric image applications, such as image search and recommendation, semantics is an important factor to determine whether a user will like an image or not. Other factors in user intention, such as visual style and emotion will also play important roles. These two factors are complementary, but they have different characteristics and thus need different learning paths from raw data to supervised information. How to design an asymmetric architecture to jointly fuse semantics learning and intention learning tasks is challenging.

- **The sparsity and unreliability of social behavioral data.** In social media platform, the number of images and users is huge. But a user can only see a tiny proportion of the whole images. Thus, the real user-image behaviors are very sparse. Furthermore, the social behavioral information is sometimes unreliable due to some "fake" users and noisy information. How to make the right use of social behavioral data is critical and challenging for image representation learning.

To address the above challenges, we propose a novel *Socially Embedded VIsual Representation learning (SEVIR)* method. It aims at learning mid-level image representation that can capture both semantics and user intention based on Convolutional Neural Network. The framework is illustrated in Figure 1. In our approach, user intention is evaluated based on social behavioral data. An Asymmetric Multi-task CNN (*amtCNN*) model is proposed to embed user intention learning task into semantic learning task. In *amtCNN*, semantic meaning is learnt from labeled image classification datasets, such as ImageNet. And user intention is learnt from the social images with quantified intention labels. In particular, we consider data reliability issue and conduct dropout on the output layer to address the sparsity and unreliability problems in social behavioral data, and thus avoid overfitting problem. As the output, the activations in mid-layers of *amtCNN* can be utilized as image representations, where both semantics and user intention are incorporated. We perform extensive experiments to demonstrate the effective of our approach. Figure 2 is a showcase, in which the curves in the left part denote the performance in *Precision@k* on our image recommendation dataset. And the representations learned for a image by different methods are visualized in the right part. Intuitively, we can observe from Figure 2 that the representations learned by *SEVIR* and semantic-oriented CNN are apparently different, and *SEVIR* performs significantly better than traditional representations, which is attributed to its capability of fusing semantics and intentions into image representations.

The contributions of the proposed approach can be summarized as follows:

1) In contrast with traditional semantic-oriented representation learning methods, we investigated an unexplored problem of learning socially embedded visual representations from scratch, where the learned representation can well capture both the semantics of images and users' intentions over the images.

2) We propose a novel Asymmetric Multi-task CNN (*amtCNN*) model to address the challenges of multiple and asymmetric learning tasks, as well as the sparsity and unreliability of social behavioral data, where two different pathways are designed to learn semantics and user intentions respectively with partially shared network architecture and data reliability in user intention learning pathway is considered.

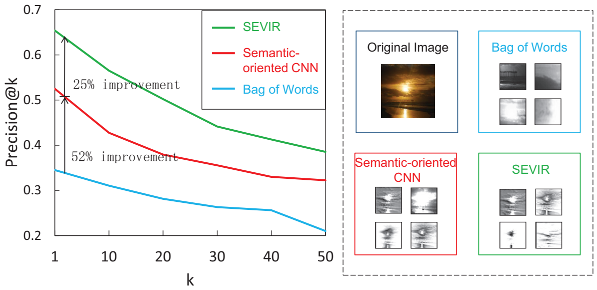


Figure 2: A showcase of the performance in *Precision@k* for image recommendation tasks (left) and the visualized features (right, best viewed in color).

3) We conducted comprehensive experiments on real application scenarios. Compared to the state-of-the-art baseline methods, our approach can achieve significant improvement, i.e., it performs at least 15% better in image recommendation and reranking tasks, which demonstrates its superiority in user-centric image applications.

The rest of the paper is organized as follows: Section 2 gives a brief introduction and comparison of related works. Section 3 introduces the framework of the proposed approach. In Section 4, we present the architecture of the proposed asymmetric multi-task CNN model and the strategies for training. Then, we report the experimental results to show the effectiveness of our approach in Section 5. Finally, Section 6 summarizes the paper.

2. RELATED WORK

2.1 CNN based Image Representation Learning

To bridge the gap between low-level features and high-level semantics, a series of works focus on learning image representation to capture semantic meaning of images. In image classification task, Convolutional Neural Network (CNN) [12, 25] has shown is superiority to traditional methods based on hand-crafted features such as Bag-of-Visual-Word features (BoW) [33] since 2012. Furthermore, in other computer vision problems, such as tracking and object detection, CNN also showed its strength [4, 6]. Not only it performs well in the above areas, but also the layer before the output layer can naturally be used as image representation. However, to date, there is few deep learning based approaches designed for image retrieval or recommendation tasks. Although some works [23, 1, 27] explore to use the codes learnt from image classification datasets for image retrieval, they can only retrieve the images in object level, where user intention, an important factor for human, is ignored.

2.2 Learning User Intention for Image Representaion

Aiming at capturing user intention to improve the performance of image retrieval and recommendation, query log analysis [8, 7, 21] and relevance feedback [24, 32] are proposed in the past years. However, users' query log in image search engines can be hardly accessed for common users and researchers. Besides, the frequent operation in relevance

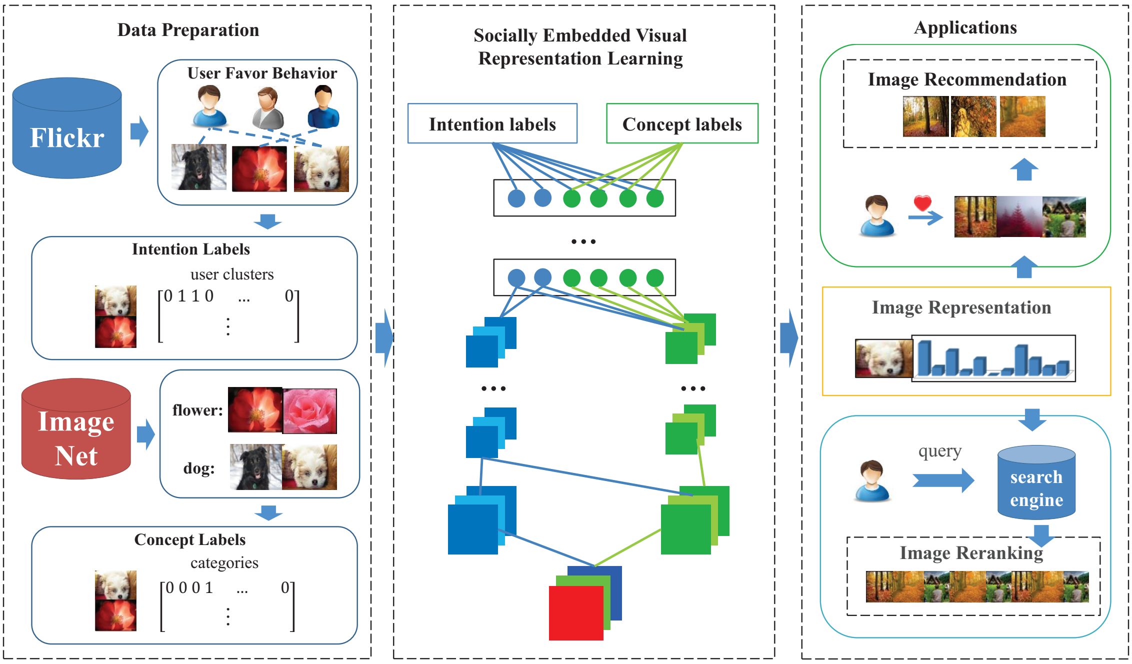


Figure 1: The conceptual framework of the proposed Socially Embedded VIsual Representation Learning approach.

feedback methods may sometimes reduce user satisfaction. With the development of social media, the information in social platforms, such as image tags [13], user behaviors [30] and user relationships, are utilized to analyze user interests. Cui *et al.* [3] proposed a social-sensed image search framework, which first summarizes user interests based on his favorite images in Flickr, and then reranks the search results based on his interests to realize personalized search. Liu *et al.* [15] learns an image distance metric based on social behavioral information to evaluate image similarity of user intention. However, traditional works based on social information usually use "shallow models". Thus, their ability in bridging semantic gap and intention gap can still be strengthened by deep models. Although in [30], Yuan *et al.* explore to learn latent features of social entities (*e.g.* users, images, tags) by deep model, this work only focuses on learning the relationship between pair-wised social entities (*e.g.* user-image and image-tag). For an image that has multi-modal information, such as faves, tags, there is no common representation for this image. In another word, when user intention is learnt based on user-image relationship, the semantic meaning in image-tag relationship will be ignored. Furthermore, as other social-sensed works show [14], the pair-wised relationships in social platforms are usually unreliable because social information is very sparse and noisy. To make image representation robust, it must completely capture both semantics and user intention.

2.3 Multi-task Deep Learning

Multi-task learning is typically applied when there are multiple related learning tasks on multiple datasets. Deep model based multi-task learning has been proven effective in many computer vision problems, such as face recognition [34], human tracking [26, 29], ontology concept learning [5], and video semantic embedding [16]. The main idea of multi-task is to share part of network structures or add some common constraints to the parameters in each task. Most of previous works aim at learning multiple subproblems in a given problem. However, in our problem, the goal is to learn two different aspects in image representation, *i.e.*, semantic aspect and intention aspect. Therefore, we cannot simply apply traditional multi-task models but have to consider the different characteristics in different task.

3. THE FRAMEWORK OF SEVIR

3.1 Overview

In this paper, we propose a Socially Embedded VIsual Representation Learning *SEVIR* approach to capture both semantic meaning and user intention in image representation. In our approach, we first prepare and pre-process two training datasets: an image classification dataset for semantic learning and a user favor behavior[1] dataset in social platforms for user intention learning. Then, a Symmetric Multitask CNN (*amtCNN*) is designed to embed user intention learning task into traditional semantic learning CNN model. Therefore, the image representation learnt in mid-layers in *amtCNN* can be applied to user-centric applications, such as image search and recommendation.

In image classification task, we usually use category label vector to represent the semantic meaning of images. To make user behavior learning task aligned to classification task, we need to restructure the favor behavior into the form of label vector which is called intention label in data pre-

[1] In Flickr, user can click "favor" button for an image, which denotes this is his/her favorite image.

processing stage. For an image, its intention label is defined as the distribution of user interests. In our approach, we divide user into different clusters with respect to the images they favor. Thus, each user cluster can represent a user interest. We can evaluate whether the current image will be liked for each user cluster based on our user favor behavior data. Furthermore, we evaluate the confidence score of the intention label to solve the sparsity and unreliability problems in social information.

After the data are pre-processed, we have two datasets: a classification dataset $\mathcal{D}_1 = \{I_1, \cdots, I_m\}$ with category label z_i for each image I_i; a social dataset $\mathcal{D}_2 = \{I'_1, \cdots, I'_n\}$ with intention y_i and confidence score t_i for each image I_i. Then, we use the proposed *amtCNN* model to learn an image representation $f(\cdot)$, so that: 1) for any $I_i \in \mathcal{D}_1$, its features $f(I_i)$ includes the knowledge in z_i, i.e., there exists $g(\cdot)$, so that $g(f(I_i)) \approx z_i$; 2) for any $I'_i \in \mathcal{D}_2$, its features $f(I'_i)$ includes the knowledge in y_i, i.e., there exists $h(\cdot)$, so that $h(f(I_i)) \approx y_i$.

In this section, we introduce how to conduct user clustering, as well as generate intention labels y_i and confidence score t_i for social images in detail. The proposed amtCNN model will be introduced in Section 4.

3.2 Intention Label for Social Images

In most of social multimedia platforms, such as Flickr, a user can "favor" some images, which indicates that the user likes this image. By collecting all images that a user favors, we can estimate his/her interests. By dividing n users into k clusters, we can regard each cluster as a interest. Thus, a social image can be represented as the distribution of the clusters whose users like it, which is called intention labels in our work. Intuitively, the images that are liked by similar user clusters should be similar in user intention aspect. Therefore, the image representations that are learnt from intention labels can capture user intention aspect. Here, we generate intent labels based on user clusters rather than independent users due to two main reason. First, social behavior information is very sparse, using independent users may make the label vector very sparse and high-dimensional. Second, when new users come, we hope the intent label to be stable in dimensionality. We first introduce how to divide users into k clusters in this section.

Inspired by [15], given n users and their favorite images, we conduct user clustering to partition the users into k clusters. First, each user u_i is represented by the set of his favorite images $\mathcal{I}m(u_i)$. Then, the pair-wised user similarity of u_i and u_j can be evaluated by the *Jaccard* similarity of $\mathcal{I}m(u_i)$ and $\mathcal{I}m(u_j)$:

$$sim(u_i, u_j) = \frac{|\mathcal{I}m(u_i) \cap \mathcal{I}m(u_j)|}{|\mathcal{I}m(u_i) \cup \mathcal{I}m(u_j)|}. \tag{1}$$

Then, we can conduct Spectral Clustering [19] on the similarity graph. Finally, we can obtain the users in each cluster, which is denoted as C_i.

However, we cannot judge that users in the same cluster must have the same interests, and vise versa. It is mainly because the data in social platforms are usually sparse and unreliable. For some active users who favor a lot of images that he/she really likes, we can understand his/her interests well. Oppositely, for the ones who favor few images or randomly favor some images, the clustering results may have some error. Therefore, for each cluster, we give a reliability score to each user in it to denote the confidence that it belongs to the cluster. In our approach, we use the centrality score calculated by PageRank [20] model as the reliability score. r_i denotes the vector of reliability scores of these users. Then, r_i can be iteratively updated by:

$$r_i(t) = d \cdot P_i \cdot r_i(t-1) + (1-d)e, \tag{2}$$

where $r_i(t)$ is the pagerank score in the t^{th} iteration. P_i is the transaction matrix, which is normalized from pair-wised user similarity of the users in $\mathcal{C}_i$ to make the sum of each column to be 1, d is the damping factor to guarantee the connectivity of the similarity graph, and e is a normalized n-dimension vector whose elements are all $1/n$. In traditional PageRank model [20], the empirical value of d is about 0.8. Thus, we adopt this value in our approach. After $r_i(t)$ converged to $\tilde{r}_i$, the final reliability score r_i is normalized to make the maximum value to be 1:

$$r_i = \frac{\tilde{r}_i}{max(\tilde{r}_i)}. \tag{3}$$

Thus, for user u_i, we have known that he/she should belong to cluster c_i with reliability r_i.

For an image I_i, we use $\mathcal{U}_i$ to denote the set of users who favor it. Thus, we can map $\mathcal{U}_i$ to a k-dimensional vector y_i to represent the intention label for I_i:

$$y_{ij} = \begin{cases} 1, & \mathcal{U}_i \cap \mathcal{C}_j \neq \phi \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In this equation, if there is at least one user in cluster j that favors image I_i, the j^{th} element in y_i will be 1. However, for different 1 in y_i, we have different confidence because the users have different reliability. We use t_{ij} to denote the confidence score that image I_i is favored by cluster j. If $y_{ij} = 1$, we can know that users $\mathcal{U}_i \cap \mathcal{C}_j$ favored I_i in $\mathcal{C}_j$. Thus, the confidence score t_{ij} is defined as the average of these users' reliability score. Thus, the confidence score t_{ij} is defined as follows,

$$t_{ij} = \begin{cases} \frac{\sum_{u \in \mathcal{U}_i \cap \mathcal{C}_j} r_j(u)}{|\mathcal{U}_i \cap \mathcal{C}_j|}, & y_{ij} = 1 \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Note that, although we define $t_{ij} = 1$ when $y_{ij} = 0$, we just know that there is no user in cluster j that favors image i. However, we do not know that the users in cluster j dislike image i. This problem will be discussed in detail in the next section.

Based on the above formulation process, for an image I_i, we can obtain two k-dimensional vectors y_i and t_i. y_i denotes which clusters of users favor the image, and t_i represents the corresponding confidence.

4. ASYMMETRIC MULTI-TASK CNN

In the last section, we introduced how to pre-process the social behavioral datasets to construct intention labels. Meanwhile, for traditional image classification task, we usually use a 1-of-k vector to denote the category that an image belongs to. Therefore, we have two datasets after data preprocessing stage: classification dataset D_1 with category labels z for semantic learning and user favor behavior dataset D_2 with intention labels y and confidence scores t for user intention learning. We now use the proposed Asymmetric Multi-task CNN (*amtCNN*) model to learn an effective image representation, in which semantic meaning and user intention are captured.

4.1 Network Architecture

We design the multi-task convolutional neural network based on the previous problem definition, which is illustrated in Figure 3. In Figure 3, there are two pathways for two tasks. The pathway in the top dash box is designed for image classification task, and the other is designed for user intention learning task. "conv", "pool" and "fc" denote convolutional layer, pooling layer, and fully connected layer in CNN correspondingly. The layer name that ends with "i_j" denotes it is the i^{th} layer in the j^{th} pathway. For example, "conv3_1" indicates that it is the third layer in the first pathway (for image classification task), and it is a convolutional layer.

In the first pathway, there are five convolutional layers and three fully connected layers, where the first, second and fifth convolutional layers are followed by pooling layers. This setting refers to AlexNet [12], which is one of typical CNN architectures in image classification area. The size of convolution kernels in the convolution layers and the number of neurons in the fully connected layers are illustrated over each layer. The setting of stride and padding is also similar to AlexNet. In this pathway, we do not adjust much from AlexNet because it is for typical image classification task. Therefore, in the forward propagation stage, the transaction function from the $(l-1)^{th}$ layer to the l^{th} layer can be formulated as follows,

$$x_l^1 = \sigma(W_{l-1}^1 x_{l-1}^1 + b_{l-1}^1), \qquad 1 < l \leq 8, \tag{6}$$

where $\sigma(\cdot)$ is the activation function, where we use RELU [12] in our network, where $\sigma(x) = x$ when $x > 0$ and 0 otherwise. Here superscript "1" refers to the first pathway; x_l^1 denotes the output of the l^{th} layer; W_{l-1}^1 denotes the weights from the $(l-1)^{th}$ layer to the i^{th} layer; b_{l-1}^1 is the bias. Equation is suitable for both convolutional layer and fully connected layer. For convolutional layers, x and b are $q_l \times 1$ vectors and W is a $q_{l-1} \times p_{l-1} \times p_{l-1} \times q_l$ tensor, where q_l is the number of feature maps and p_l is the size of convolutional kernel in the l^{th} layer. For fully connection layers, we can regard each neuron as a 1×1 convolution. Thus W is a $q_{l-1} \times q_l$ matrix. Different from the previous layers, the final output layer $\tilde{z}$ is defined as:

$$\tilde{z} = softmax(\sigma(W_8^1 x_8^1 + b_8^1)), \tag{7}$$

where $softmax(\cdot)$ is a function which can generate a distribution from a given vector.

The second pathway is designed to embed social favor behavior. In this pathway, the types of layers are similar to those in the first pathway, but the connections are quite different. A convolutional or fully connected layer in the second pathway is connected to the previous layers in both of the first and the second pathway, *i.e.*,

$$x_l^2 = \sigma(W_{l-1}^{2,1} x_{l-1}^1 + b_{l-1}^{2,1} + W_{l-1}^{2,2} x_{l-1}^2 + b_{l-1}^{2,2}), 1 < l \leq 8, \tag{8}$$

where $W_{l-1}^{2,i}$ $(i = 1, 2)$ refers to the weights from the i^{th} pathway's $l-1$ layer to the second pathway's l layer, and $b_{l-1}^{2,i}$ is the corresponding bias. In the second pathway, the output layer $\tilde{y}$ is defined as:

$$\tilde{y} = \sigma'(W_8^{2,1} x_8^1 + b_8^{2,1} + W_8^{2,2} x_8^2 + b_8^{2,2}), \tag{9}$$

where $\sigma'(\cdot)$ is sigmoid function, which is defined as $\sigma'(x) = \frac{1}{1+e^{-x}}$. Compared to Equation 7, we do not use softmax+RELU but sigmoid function. It is mainly because in the first pathway, the groudtruth label is z_i, which is a 1-of-k indicator. However, in the second pathway, the groundtruth is y_i, which have multiple 1s. Therefore, the softmax (a probability distribution) is not suitable.

In this work, we call the layers in the first pathway "semantic layers" and the second pathway "social layers". Therefore, we can observe that a semantic layer is only determined by the last semantic layer, while a social layer is determined by both of the last semantic layer and the last social layer. This is because user intention, *i.e.*, which kind of users will like this image, is first determined by the semantic meaning of the image, and also determined by other emotional or social aspects. Based on this setting, we can embed the user intention learning task into traditional image classification task. From Figure 3, we can observe that the final size of each social layer is half of the size of corresponding semantic layer to achieve the best performance. There are two reasons for this phenomenon. First, semantic layer has included a lot of useful information which will effect user intention. Thus, we do not need so much social nodes. Second, for image classification task, ImageNet is a very big dataset to pre-train our network. On the other hand, the Flickr dataset is crawled by ourselves and thus does not have a very huge amount. If the size of social layers is too big, the network may face under-fitting problem due to the lack of training samples.

4.2 Training

Based on the architecture in Figure 3, our target is to derive the optimal weights in the network. For each image, the input data x_1 is the pixel data of RGB channels, *i.e.*, $x_1 =< I_i^R, I_i^G, I_i^B >$. In the first pathway, *i.e.*, image classification task, we expect the output layer $\tilde{z}_i$ to be close to the classification label vector z_i. In ImageNet ILSVRC-2012 dataset, there are 1000 categories in total. Thus, z_i is a 1000-D vector where only one element is 1 and the others are 0. We define the loss function for this task as L_1, which can be computed according to the loss function of regression:

$$L_1 = \frac{1}{2}\sum_{i=1}^{N}(z_i - \tilde{z}_i)^2 + \frac{\lambda}{2}\sum_{l=1}^{8}\|W_l^1\|_F^2, \tag{10}$$

where $\tilde{z}_i$ is the output $\tilde{z}$ in Equation 7 for image I_i; λ is the factor to balance the loss and regularization to previous overfitting.

For the second user intention learning task, the supervised label is y_i. Here, the loss function should be re-defined because we have different confidence for the elements in y_i. In Section 3, we compute the confidence score t_i for y_i. As mentioned, when $y_{ij} = 1$, t_{ij} denotes the confidence that users in cluster j like image i. However, $y_{ij} = 0$ only denotes "unknown" but not "dislike". If we use all 0s in y_i for training, it may bring a lot of noise. Here we refer to the idea of dropout to solve this problem. For all the js that make $y_{ij} = 0$, we only select part of them for back propagation. Therefore, we first modify the confidence score in Equation 5 as follows,

$$t'_{ij} = \begin{cases} t_{ij}, & y_{ij} = 1 \\ 1, & y_{ij} = 0 \text{ and } j \text{ is selected} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

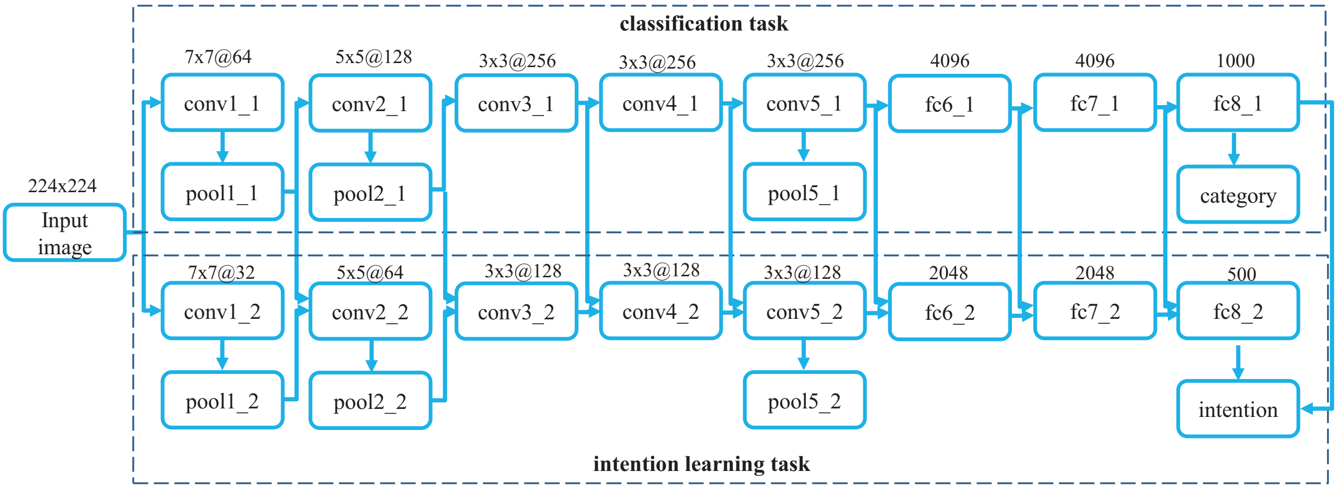


Figure 3: The architecture of the proposed network. The top dash line box is designed for image classification task and the bottom one is for favor behavior learning task. The arrow line between two layers denotes that these two layers are fully connected.

In this work, the dropout rate is 0.2, *i.e.*, we only select 20% 0s in y_i for back propagation. Then, the loss function for this task is defined as,

$$L_2 = \frac{1}{2}\sum_{i=1}^{N} t'_i(y_i - \tilde{y}_i)^2 + \frac{\lambda}{2}\sum_{l=1}^{8}(\|W_l^{2,1}\|_F^2 + \|W_l^{2,2}\|_F^2). \quad (12)$$

Using Equation 12, only the weights connected to the selected neuron in the output layer will be updated in the back propagation stage.

We use batch mode in the training stage. We first select a batch of images in classification dataset D_1, and use L_1 to update the weights W^1. Second, we select a batch of images in social image dataset D_2, and user L_2 to update the weights $W^{2,1}$ and $W^{2,2}$. We iteratively repeat the above two operations until the errors converge on both of the validation sets for two tasks.

When the training processed is finished, we can use the combination of mid-level fully connect layers, *i.e.*, the $fc6$ to $fc8$, as the learnt representation. In our approach, we find $fc8$ performs the best on the validation set. Therefore, for image I_i, its representation in our network $f(I_i)$ is $< x_{i,8}^1, x_{i,8}^2 >$.

4.3 Algorithm

We summarize the algorithm of the whole Socially Embedded Visual Representation Learning approach as described in Algorithm 1.

Our approach has three main steps. First, we divide n users into k clusters. Then, we compute the k-dimensional vector y and t for each social image. Finally, we train our *amtCNN* model based on the pre-processed data.

5. EXPERIMENTS

In this section, we first introduce the experimental settings in this work. Then we evaluate the proposed approach in three application scenarios: user favor behavior prediction, personalized image recommendation, as well as image reranking. Finally, we give some observations of the learnt image representation.

5.1 Experimental Setup

5.1.1 Datasets

In our experiments, we have prepared four datasets, including:

Training dataset. The training dataset consists of two sub sets: D_1 for classification and D_2 for user intention learning. In this work, D_1 is a subset of ILSVRC-2012 dataset in ImageNet and D_2 is crawled from Flickr. We do not use the whole ILSVRC-2012 dataset as D_1 bacause its amount is much more than D_2. We first use the whole ILSVRC-2012 to pre-train the weights in classification pathway in our network. Then, we use D_1 and D_2 for fine-tuning. We randomly select 128,000 images from ILSVRC-2012 dataset as D_1, where there are 1000 categories in total and each category has 128 images. To construct D_2, we select $128,000$ images from *Yahoo Flickr Creative Commons WebScope* dataset [2] with popular tags [3]. Then, we crawl the users who favor them through Flickr API. Finally, there are 97,513 users in total and each image is favored by 4 users in average. For each of D_1 and D_2, we use 90% for training and remained 10% for test.

Favor behavior prediction dataset. We crawl other 10,000 images that are favored by the users in training dataset in Flickr. The user intention labels y_i computed by Equation 4 is used as groundtruth to predict which clusters of users will favor them. This dataset can be regarded as the test dataset after training.

Recommendation dataset. We also prepare a recommendation dataset for image recommendation. We crawled $1,000$ users' favorite images in Flickr. The number of the images in total is 17,935. All of the selected users have at

[2] http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67
[3] https://www.flickr.com/photos/tags/

```
Algorithm 1: SEVIR
  Input: Image Classification dataset $\mathcal{D}_1 = \{I_1, \cdots, I_m\}$
         with label $z_i$ for each $I_i$;
  User favorite images $\mathcal{D}_2 = \{I'_1, \cdots, I'_n\}$ with favored
  users $\mathcal{U}_i$ for each $I'_i$;
  Output: The trained CNN model, which can obtain
          the representation $f(I_i)$ for any image $I_i$.
  for pair-wised users $(u_i, u_j)$ do
      Compute user similarity $sim(u_i, u_j)$ by Eq. 1;
  end
  Conduct spectral clustering on the similarity graph;
  for each cluster $C_i$ do
      for user $u \in C_i$ do
          Compute the reliability score $r(u)$ by Eq. 3;
      end
  end
  for each image $I_i \in D_2$ do
      Compute intention labels $y_i$ by Eq. 4;
      Compute confidence score vector $t_i$ by Eq. 5;
  end
  repeat
      Select batch of images $B_1 \subseteq D_1$;
      Compute loss $L_1$ along the first pathway by Eq. 10;
      Conduct back propagation for the first pathway;
      Select batch of images $B_2 \subseteq D_2$;
      Compute loss $L_2$ along the second pathway by Eq.
      12 with dropout;
      Conduct back propagation for the second pathway;
  until errors for both two pathways converged;
```

least 80 favorite images. For each user, we randomly select 40 of his favorite images as ground truth for test and the remained favorite image for training. Then we randomly sample other 160 images from the whole image dataset as candidates for testing.

MSR dataset. We utilize Bing Image Retrieval Grand Challenge (MSR) dataset [7] to prove that our distance learning method can improve the performance of image reranking. In this dataset, the images under a given query are labeled as “Excellent”, “Good” or “Bad” with respect to the click count. Different from traditional image retrieval datesets, MSR dataset is based on user click data, which can capture not only semantic relevance but also human cognition.

To the best of our knowledge, there is no public benchmark including user behavioral information for image recommendation. Thus, we crawled the data from Flickr for the first three datasets. In our image reranking method, a public dataset is utilized to make our approach comparable to others.

5.1.2 Model Implementation

We train the network introduced in Section 4 on our training dataset. Our training strategy follows the practice of the previous works on CNN [12, 31]. In our training set, each image is resized to 256×256. Five 224×224 crops are cropped from the center and the four corners of the resized image. We also conduct horizonal flipping and vertical flipping for each cropped sample. Following AlexNet [12], dropout with probability 0.5 is used in the first two fully connection layers in each task in learning process. The learning rate starts from 0.02 for all layers. It is divided by 10 when the error rate stops reducing. Our model is modified based on the public code of Caffe [9]. It is trained on a single GeForce Tesla K40 GPU with 12GB memory. The training process will cost about 3 days.

5.2 Favor behavior Prediction

In our CNN model, for any image I, we can compute the output $\tilde{y}$ in the second pathway using Equation 9. For an image in our favor behavior prediction dataset, y means which clusters of users will like it. Therefore, we can use $\tilde{y}$ to predict users' favor behavior. In this experiment, for each image I in favor behavior prediction dataset, we evaluate the similarity between the prediction result $\tilde{y}$ and groundtruth y in metrics of $RMSE$ (Root Mean Square Error), $Precision@k$, and $Kendall - \tau$ [11]. Here $RMSE$ reflects the error in value; $Precision@k$ reflects the accuracy when we recommend the image to k clusters of users; and $Kendall - \tau$ evaluates the performance in ranking point of view. For $RMSE$, lower value is better. For the other metrics, the higher the better.

To compare with our approach, we use the following methods as baselines:

- **Logistic Regression with Bag-of-Words feature (BOW) [33].** In this method, we first train a logistic regression model on the training dataset D_2. Bag-of-Words features are extracted based on SIFT descriptors [17]. Then we use the favor prediction dataset to evaluate the performance using the trained logistic regression model.
- **Logistic Regression with SIDL (SIDL).** SIDL [15] is a distance learning method which incorporates social behavioral information. The learned image distance metric can map the original image feature to a new space. Like the previous baseline, we train a logistic regression model based on the mapped features.
- **AlexNet.** AlexNet [12] is a CNN model proposed for image classification on ImageNet which captures semantic information. We use the last layer before the output layer as image representation.
- **Social task in SEVIR (SEVIR_soc).** Here we only use the second pathway in our proposed SEVIR Network. The network is trained on Flickr dataset D_2 and the last layer before the output layer is used as image representation.

In our experiments, we mark our proposed method as $SEVIR\_soc + sem$ because it captures social information and semantic information at the same time. Table 1 shows the results.

Table 1: The performance on favor behavior prediction dataset in $RMSE$, $Precision@k$, and $Kendall - \tau$ for different image representation methods.

	$RMSE$	P@3	P@10	$Kendall - \tau$
BoW [33]	0.4368	0.1201	0.2855	0.057
SIDL [15]	0.3345	0.2347	0.3451	0.3939
AlexNet [12]	0.347	0.2099	0.3362	0.3467
SEVIR_soc	0.3199	0.2289	0.353	0.3751
SEVIR_soc+sem	**0.2857**	**0.2556**	**0.4061**	**0.4277**

From Table 1, it can be observed that the proposed method $SEVIR\_soc + sem$ achieves the best performance in all of the metrics. BoW performs the worst because it is extracted only based on visual contents. Thus, it does not include adequate information for semantic and intention. Although SIDL also embeds social behavioral information, it performs worse than $SEVIR\_soc$ because the linear model (logistic regression) cannot bridge low-level feature and high-level favor behavior well. Intuitively, favor behavior is more related to user intention than semantic meaning of images. However, we can observe that AlexNet performs better than $SEVIR\_soc$. To our understanding, it is mainly because the social data are very sparse and unreliable. Thus the model that fits the training social data very much may reduce the performance on test data on the contrary due to overfitting.

5.3 Image Recommendation

In the image recommendation dataset, we have 1,000 users' favorite images in Flickr. We divide part of them for training and rest for test. In test stage, our task is to recommend 40 images from 200 candidate images for each user. In this experiment, we use the following recommendation methods to show the effectiveness of our approach:

- **Content-based Filtering using Bag-of-Words Features (BoW).** Content-based Filtering [22] is one of the most popular content based recommendation methods. The idea is to rank the candidate images according to their similarity to the training images of a given user. Here we use Bag-of-Words features to evaluate image similarity.
- **Content-based Filtering based on SIDL, AlexNet, SEVIR_soc, SEVIR_soc+sem.** Respectively, we use SIDL, AlexNet, SEVIR_soc, SEVIR_soc+sem to evaluate image similarity for Content-based Filtering.
- **Content-boosted Collaborative Filtering based on SIDL, AlexNet, SEVIR_soc, SEVIR_soc+sem.** Collaborative Filtering [2] is a typical recommendation method which uses user behavioral information. To demonstrate that our content-based method and CF are competitive. We adopt the approach in [18] to combine Content-based Filtering with Collaborative Filtering. Here we use the proposed SEVIR to evaluate content similarity of images.

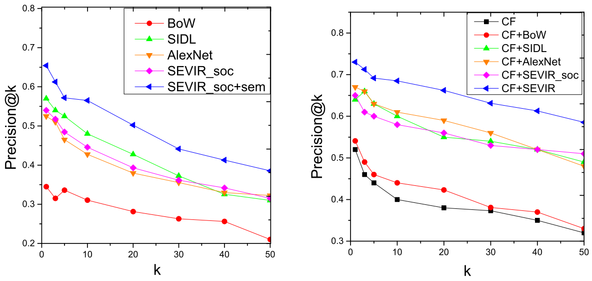
We use $Precision@k$ to evaluate the performance of the above recommendation methods. Figure 4 shows the results for image recommendation. In Figure 4, (a) illustrates the performance of content-based methods of our proposed SEVIR and baseline methods; (b) illustrates the performance of combining content-based methods with collaborative filtering. From Figure 4 (a), we can see that the proposed SIVIR_soc+sem performs the best, which is consistent to the previous experiments. When combined with Collaborative Filtering (CF), we can observe from Figure 4 (b) that the hybrid method CF+SEVIR (here SEVIR means SIVIR_soc+sem) still performs the best. It indicates that although the pure CF produces relatively good performance, our proposed SEVIR is still an effective image representation which is competitive to CF.

5.4 Image Reranking

In this experiment, we follow the method in [10] to rerank the images for a given query in MSR dataset. In this method, pair-wised image similarity is computed to build the similarity graph. Then, PageRank [20] is used to compute the centrality scores for each image. Therefore, the image with high PageRank score is similar to most of other images. Following the measurements in ACM Multimedia Grand Challenge 2013, we use Discounted Cumulated Gain of the top 25 images ($DCG@25$) to evaluate the performance. When the rank order is given, the $DCG@25$ for each query is calculated as:

$$DCG@25 = 0.01757 \sum_{i=1}^{25} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \quad (13)$$

where rel_i is the relevance score of the i^{th} ranked image (Excellent=3, Good=1, Bad=0), 0.01757 is to normalize the value of $DCG@25$ up to 1. Note that the $DCG@25$ of a perfect ranking may also be less than 1 when there are some non-excellent images in top 25. For the queries with fewer than 25 images, we simply supply some "Bad" images after the original ranking list. We use the baseline methods similar to favor behavior prediction task to evaluate image similarity. Figure 5 shows the results of our experiments. From Figure 5, we can observe that our approach has remarkable superiority to the other baseline methods. SIDL, AlexNet, and SEVIR_soc produce similar performance because they include different aspects of information: SIDL uses both users' favor behavior and image tags, but it is a shallow learning model; AlexNet considers semantic categories; SEVIR_soc only uses favor behavior but it is a deep learning model. Therefore,



(a) (b)

Figure 4: The performance on recommendation dataset in $Precision@k$ with different number of top k images using (a) content-based methods (b) hybrid methods of content-based filtering and collaborative filtering.

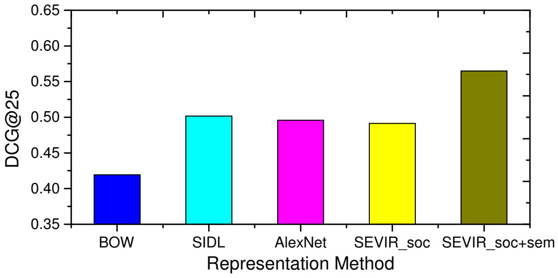


Figure 5: The performance on the MSR dataset in terms of $DCG@25$ with using different image representations in the PageRank model.

it is reasonable that our approach, which uses deep learning model and captures both user intention and semantic information achieves 0.5649 in *DCG*@25. Note that, here we just use the development dataset in MSR. Thus, the reranking method is unsupervised. However, the performance is comparable to the state-of-the-art methods that are supervised by training dataset. It indicates that our approach trained by Flickr images can indeed capture user intention information, which is independent to training data.

5.5 Discussions

In our approach, we are very curious about what we have learned from the classification task and the behavior learning task. Therefore, we visualize some representative activations in our CNN model. We first select 3 images that have the similar semantic meaning of "flower". Then, we compute the activations of bottom convolutional layers (conv1) and top convolutional layers (conv5) for social task and semantic task. Figure 6 illustrates representative activations that we obtain.

Although the selected 3 images are all about flowers, they have quite different favored users in Flickr. The first two images are from a flower lover's album, so that the favored users are similar. The third image is a photo in a party. Thus, it has no overlap with the previous two images in user dimension. From Figure 6, we have some interesting observations. For conv1, the activations in semantic task are very close to the original image, but the ones in social task have been abstract. It indicates that the features extracted in behavior learning task are relatively latent but not intuitive. For conv5, the activations in semantic task describe the outline of the flowers, while the ones in social task seem to be some very detailed aspects, such as texture of pistil or petal. For the activations of conv5 in semantic task, three images are very similar, which leads to the same classification result "flower". While when we observe the activations of conv5 in social task, the third image is obviously different with the first two images. Intuitively, for the third image, the activations of conv5 in social task deliver the information about "many" because there are many black "dots" in them. However, for the first and the second images, the activations are relatively pure. Therefore, we can find that we really learned some extra information in social task that differs from classification task.

6. CONCLUSION

In this paper, we explore learning image representation to capture both semantics and user intention for user-centric applications, such as image search and recommendation. A Socially Embedded VIsual Representation Learning (*SEVIR*) approach is proposed, in which an Asymmetric Multitask Convolutional Neural Network (*amtCNN*) model is designed to embed user intention learning task into semantic learning task. In user intention learning task, we first compute intention labels with confidence scores based on favor behavior in social platforms. Then, the loss function is specifically designed to tackle the sparsity and unreliability challenges in social behavioral information. The experimental results in user behavior prediction and image reranking applications indicate that our representation learning approach includes more intention level information than baseline methods. The experiment for image recommendation demonstrate that the learnt representation can make the performance of content-based filtering method comparable to collaborative filtering methods. In all of the experiments, our approach performs at least 15% better than baseline methods, which is remarkably better than baseline methods. From the visualization of the learning results, we observed that learnt features in social task are quite different with the ones in classification task.

This work is an effort to embed social information into deep CNN model. We have a long way to go because the quality of social data is much worse than well-labeled classification data. To our understanding, this is the main reason that we can just observe the difference of the features in visualization part, but we cannot give a good intuitive explanation. In the future, we will explore to better organize social information and design a more reasonable deep model for embedding to make image representation learning more intelligent and more personalized. Furthermore, we can incorporate multi-modal social information to better understand user intention.

7. ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program of China, No. 2015CB352300; National Natural Science Foundation of China, No. 61370022 and No. 61210008; International Science and Technology Cooperation Program of China, No. 2013DFG12870. Thanks for the support of NExT Research Center funded by MDA, Singapore, under the research grant, WBS:R-252-300-001- 490 and the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

8. REFERENCES

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599. Springer, 2014.

[2] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[3] P. Cui, S.-W. Liu, W.-W. Zhu, H.-B. Luan, T.-S. Chua, and S.-Q. Yang. Social-sensed image search. *ACM TOIS*, 32(2):8, 2014.

[4] J. Fan, W. Xu, Y. Wu, and Y. Gong. Human tracking using convolutional neural networks. *TNN*, 21(10):1610–1623, 2010.

[5] X. Geng, H. Zhang, Z. Song, Y. Yang, H. Luan, and T.-S. Chua. One of a kind: User profiling by social curation. In *ACM Multimedia*, pages 567–576. ACM, 2014.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE, 2014.

[7] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In *ACM Multimedia*, pages 243–252, 2013.

[8] B. J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 28(3):407–432, 2006.

[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] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *PAMI*, 30(11):1877–1890, 2008.

[11] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 1938.

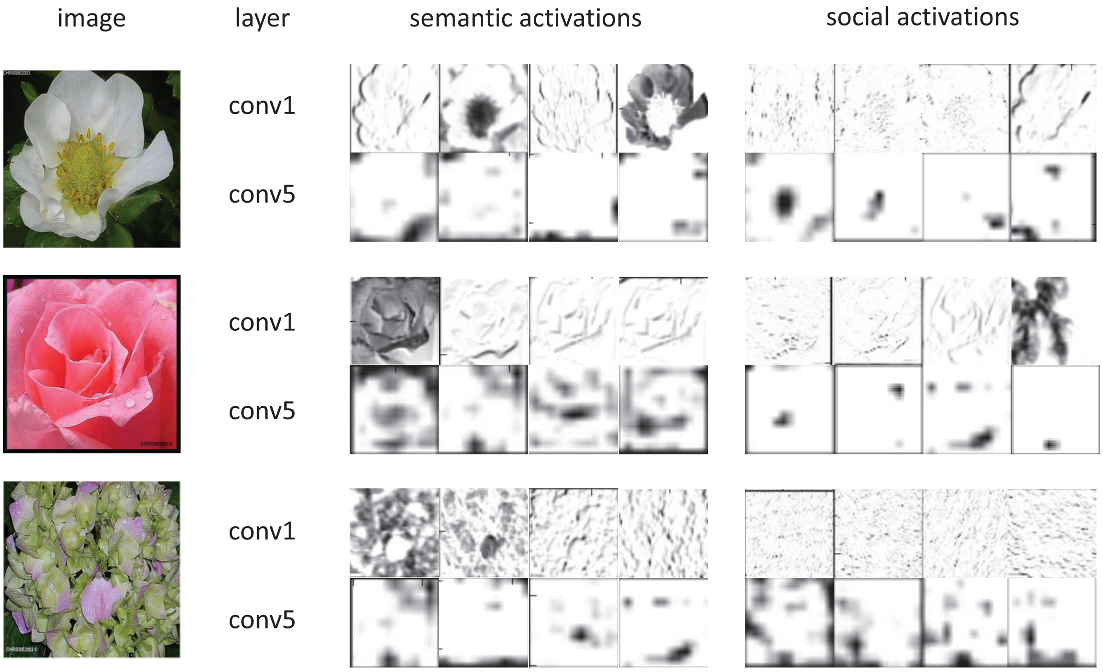


Figure 6: The illustration of representative activations of bottom convolutional layers (conv1) and top convolutional layers (conv5) for social task and semantic task.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[13] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *TMM*, 11(7):1310–1322, 2009.

[14] S. Liu, P. Cui, H. Luan, W. Zhu, S. Yang, and Q. Tian. Social visual image ranking for web image search. In *MMM*, pages 239–249. Springer, 2013.

[15] S. Liu, P. Cui, W. Zhu, S. Yang, and Q. Tian. Social embedding image distance learning. In *ACM Multimedia*, pages 617–626. ACM, 2014.

[16] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, pages 3707–3715, 2015.

[17] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157. Ieee, 1999.

[18] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *AAAI*, pages 187–192, 2002.

[19] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *NIPS*, 2:849–856, 2002.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[21] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui. Click-through-based cross-view learning for image search. In *SIGIR*, pages 717–726. ACM, 2014.

[22] M. J. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5-6):393–408, 1999.

[23] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshop*, pages 512–519. IEEE, 2014.

[24] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *CSVT*, 8(5):644–655, 1998.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014.

[26] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014.

[27] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *ACM Multimedia*, pages 157–166. ACM, 2014.

[28] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using fast kernel machines. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2177–2188, 2012.

[29] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In *NIPS*, pages 809–817, 2013.

[30] Z. Yuan, J. Sang, Y. Liu, and C. Xu. Latent feature learning in social media network. In *ACM Multimedia*, pages 253–262. ACM, 2013.

[31] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[32] H. Zhang, Z.-J. Zha, S. Yan, J. Bian, and T.-S. Chua. Attribute feedback. In *ACM Multimedia*, pages 79–88. ACM, 2012.

[33] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multimedia*, pages 75–84. ACM, 2009.

[34] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108. Springer, 2014.