The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# Retrieving and Classifying Affective Images via Deep Metric Learning

**Jufeng Yang,[1] Dongyu She,[1] Yu-Kun Lai,[2] Ming-Hsuan Yang[3]**
[1]College of Computer and Control Engineering, Nankai University, Tianjin, China
[2]School of Computer Science and Informatics,
Cardiff University, Cardiff, United Kingdom
[3]School of Engineering, University of California, Merced, USA

## Abstract

Affective image understanding has been extensively studied in the last decade since more and more users express emotion via visual contents. While current algorithms based on convolutional neural networks aim to distinguish emotional categories in a discrete label space, the task is inherently ambiguous. This is mainly because emotional labels with the same polarity (*i.e.*, positive or negative) are highly related, which is different from concrete object concepts such as cat, dog and bird. To the best of our knowledge, few methods focus on leveraging such characteristic of emotions for affective image understanding. In this work, we address the problem of understanding affective images via deep metric learning and propose a multi-task deep framework to optimize both retrieval and classification goals. We propose the sentiment constraints adapted from the triplet constraints, which are able to explore the hierarchical relation of emotion labels. We further exploit the sentiment vector as an effective representation to distinguish affective images utilizing the texture representation derived from convolutional layers. Extensive evaluations on four widely-used affective datasets, *i.e.*, Flickr and Instagram, IAPSa, Art Photo, and Abstract Paintings, demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods on both affective image retrieval and classification tasks.

## Introduction

Psychological studies have demonstrated that visual contents (*e.g.*, images and videos) can evoke a variety of emotional responses for human observers (Detenber, Simons, and Bennett Jr. 1998). As such, it is intriguing and important to understand the emotion of a given image due to its broad potential applications including emotion semantic image retrieval (ESIR) (Wang and He 2008), aesthetic quality categorization (Lu et al. 2014), and opinion mining (Qian, Zhang, and Xu 2016; Zhao et al. 2016), to name a few.

Numerous methods have been developed for classifying (Machajdik and Hanbury 2010; Zhao et al. 2014a) and retrieving (Zhang et al. 2013; Zhao et al. 2014b) affective images, where the main focus is to extract low-level features (*e.g.*, texture, color and composition) that can well represent the evoked emotion from visual contents. Instead of designing visual features manually, convolutional neural networks (CNNs) provide end-to-end feature learning frameworks. Several CNN-based methods for affective image classification demonstrate the effectiveness of deep representations over hand-crafted features (You et al. 2016; Yang, She, and Sun 2017).


Figure 1: Examples from the Flickr and Instagram dataset where different colors indicate different sentiments. The Mikels' emotion wheel suggests that pairwise emotion correlation can be defined as the reciprocal of the corresponding distance. Clearly the emotion labels with the same polarity (*i.e.*, positive emotions in the top row, or negative emotions in the bottom row) are highly related.

However, compared to conventional vision tasks (*e.g.*, object recognition), affective image understanding is inherently ambiguous due to the following two challenges, namely subjectivity and complexity of emotions. First, the emotions are not semantically independent of each other, which is drastically different from concrete object concepts (*e.g.*, cat, dog and bird). Figure 1 shows that there exists a clear hierarchical relation that the emotions with the same polarity (*e.g.*, both positive or negative) are highly related. Moreover, the correlation between emotions with the same polarity can be defined by the reciprocal of pairwise distance in Mikels' wheel (Mikels et al. 2005), where the neighboring emotions are more related to each other. However, existing studies have rarely focused on leveraging the relation of emotions for affective image understanding since they are trained in the discrete label space (You et al. 2015). Second, most CNN-based methods rely on the feature representation ability of CNNs, especially fully-connected lay-

ers (Xu et al. 2014). There exists an underlying assumption that the CNN's features, which are good at distinguishing high-level semantic contents (*e.g*., objects), are also good at distinguishing the image emotion. However, this is not necessarily true for emotions and it is possible that these features are insufficient to characterize them. In fact, some studies reveal that texture information is one of the most important elements related to visual emotion (Machajdik and Hanbury 2010; Rao, Xu, and Xu 2016), which is not emphasized in recent deep models for affective image understanding.

In this work, we address emotion relations via deep metric learning (Hoffer and Ailon 2015), which allows incorporating similarity constraints (*e.g*., triplets) to learn the feature embedding. However, compared with classification models which emphasize the softmax loss, the learned embedding provides sub-par results when used as features for classification. To address this, this paper proposes a multi-task framework to simultaneously optimize the classification and retrieval tasks by combining the softmax loss and sentiment loss in the CNN training. Specifically, we present novel sentiment constraints by considering the relations among emotional categories demonstrated in the Mikels' wheel, which extend triplet constraints to a hierarchical structure. Moreover, inspired by the Gram matrix (Gatys, Ecker, and Bethge 2015), we propose a sentiment vector based on the texture information from the convolutional layer, which calculates the correlations between feature responses. Our framework uses the sentiment vector rather than the fully-connected feature vector to measure the difference between affective images, which is more effective for characterizing emotions.

The contributions of this work are summarized as follows: First, we address the challenges of affective image understanding via deep metric learning, and propose a unified framework to simultaneously optimize the retrieval and classification goals. We propose the sentiment constraint generalized from the triplet constraint to incorporate the relation of emotional categories to the CNN learning process. Second, we propose the sentiment vector to measure the difference between affective images, which captures the texture information from multiple convolutional layers. The experimental results on four popular affective datasets (*i.e*., Flickr and Instagram, IAPSa, Art Photos and Abstract Paintings) demonstrate that our proposed framework can effectively retrieve similar images based on emotions and also outperform the state-of-the-art methods for emotion classification.

## Related Work

In this section, we review the affective image understanding methods and metric learning algorithms which are most related to this work.

**Affective Image Understanding.** Most affective image understanding methods focus on the classification problem using hand-crafted features or discriminative representations from deep learning. Numerous methods based on low-level features (Yanulevskaya et al. 2008) or mid-level representations (Chen et al. 2014b; Zhao et al. 2017a; 2017b) have been developed. Recently, the relationship between CNN features and human emotions has also been demonstrated on photographs. The DeepSentiBank (Chen et al. 2014a) method constructs a detector for visual sentiment concept based on the classification on adjective-noun pairs. This representation encodes statistical cues for detecting emotions depicted in images effectively. Some methods aim to incorporate the model weights learned from a large-scale general dataset (Deng et al. 2009) and fine-tune the state-of-the-art CNNs for the task of visual emotion prediction. You *et al*. (You et al. 2015) propose a novel progressive CNN architecture to utilize large-scale web data, and further perform benchmarking analysis on the Flickr and Instagram dataset (You et al. 2016).

Most existing CNN-based methods for affective image classification employ the softmax loss to maximize the probability of the correct class, which fails to consider the relations between different emotional categories since these models are trained in a discrete label space. Moreover, in (Zhang et al. 2013), Zhang *et al*. apply a multiple kernel learning framework to retrieve and classify affective images, whereas Zhao *et al*. (Zhao et al. 2014b) evaluate the performance of different features on affective image retrieval in a multi-graph learning framework. Both models lead to sub-optimal performance when locating images at the affective level, since off-the-shelf features are used as input without end-to-end feature learning. In this work, we propose to incorporate the sentiment constraints to the CNN by employing deep metric learning for both affective image classification and retrieval tasks.

**Deep Metric Learning.** Metric learning has been widely studied in pattern recognition and image analysis within the past decades (Bellet, Habrard, and Sebban 2013). Recent methods employ CNNs with either pairwise (contrastive) (Chopra, Hadsell, and LeCun 2005) or triplet constraints (Chechik et al. 2010) to learn feature embeddings capturing the semantic similarity among images. Deep learning methods are verified to perform favorably against conventional methods based on hand-crafted features. As such, deep metric learning methods have been successfully applied to a variety of domains, *e.g*., face verification (Taigman et al. 2014), image retrieval (Wang et al. 2014), and geo-localization (Lin et al. 2015). Other than using classification constraints alone, a few schemes incorporate similarity constraints to generate discriminative features (Wang et al. 2014; Zhang et al. 2016). The most related work is (Zhang et al. 2016) which jointly optimizes the softmax and triplet losses for the fine-grained task. Our framework differs from that paper in two aspects: (1) The relation between emotional categories is different from that of fine-grained categories. The fine-grained labels are explicitly distinguished, while the emotion relations with the same / different polarity are also taken into consideration. (2) Existing schemes use features from the last few fully connected layers of CNNs as feature embeddings (You et al. 2016), while we adopt the sentiment vector utilizing the texture information from the convolutional layers for better affective image understanding. Extensive experiments show the effectiveness of our framework for retrieving and classifying affective images.

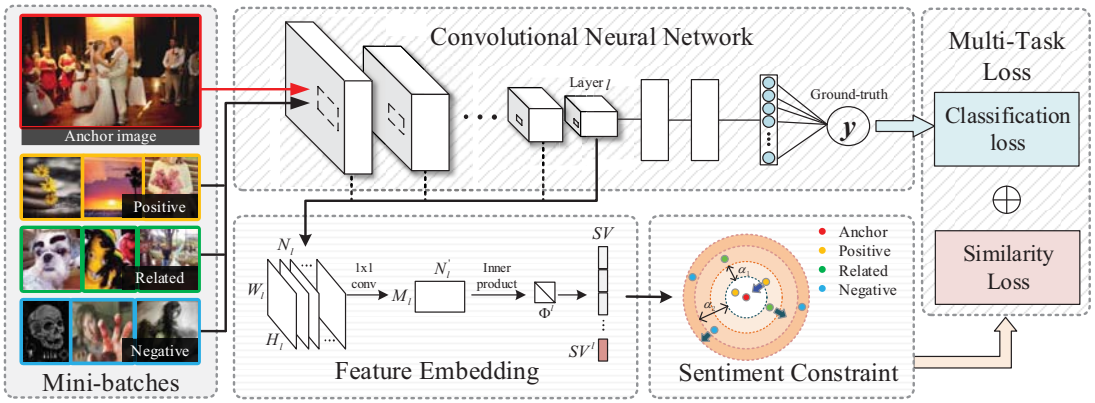



Figure 2: Illustration of the proposed algorithm. Given the mini-batches containing images with different emotions, we first generate the sentiment vector to measure the emotional difference with the $1 \times 1$ convolutional layer followed by the inner product operation. Our framework simultaneously optimizes the classification loss (*i.e.*, softmax) and similarity loss (*i.e.*, sentiment).

## Proposed Algorithm

Figure 2 shows the main steps of the proposed algorithm. Given the mini-batches of affective images, sentiment vectors (SVs) can be generated based on the Gram matrix from multiple convolutional layers. The sentiment constraint is then embedded in the training as the sentiment loss to uncover the relation of emotions. Then the unified framework is simultaneously optimized with the softmax loss as well as the sentiment loss for affective image understanding.

### Sentiment Metric Learning

We propose to learn a sentiment metric by comparing image pairs according to the Euclidean distance $D$ of their texture representation with unit norm:

$$D(x_i, x_j) \mapsto \|SV_i - SV_j\|_2^2, \tag{1}$$

where $x_i$ and $x_j$ denote the different anchor images from the training set $\Gamma$, and $SV_i$ and $SV_j$ refer to the sentiment vectors computed from the convolutional layers. The distance between different emotional images is arguably subjective, but the general relationship is clear and should be well satisfied: images with the same polarity are close to each other while those of the opposite polarity should be further apart. Thus, we generalize the triplet constraint to the sentiment constraint taking emotion relationships into consideration.

**Triplet Constraints.** Existing algorithms (Schroff, Kalenichenko, and Philbin 2015) usually generate mini-batches of triplets, *i.e.*, an anchor $a_i$, a positive instance $p_i$ of the same class, and a negative instance $n_i$ of a different class. The goal is to learn an embedding function that assigns a smaller distance to more similar image pairs, which can be expressed as:

$$D(a_i, p_i) + \alpha < D(a_i, n_i), \forall (a_i, p_i, n_i) \subset \Gamma, \tag{2}$$

where $\alpha > 0$ is a margin that is enforced between positive and negative pairs.

**Sentiment Constraints.** There exists a hierarchical relation between sentiment labels. For example, Mikels' eight emotions have four positive and four negative emotions (Mikels et al. 2005). We ensure that an image $a_i$ (anchor) of a specific emotion is closer to all images $p_i$ (positive) of exactly the same emotion, which is again closer than it is to any related images $r_i$ with emotion of the same polarity, while images with the opposite polarity $n_i$ remain the furthest distance away. The difference between these two constraints is illustrated in Figure 3. Formally, a sentiment constraint can be denoted as

$$\begin{cases} D^*(a_i, p_i) + \alpha_1 < D^*(a_i, r_i) \\ D^*(a_i, r_i) + \alpha_2 < D^*(a_i, n_i) \end{cases}, \forall (a_i, p_i, r_i, n_i) \subset \Gamma, \tag{3}$$

where $\alpha_1, \alpha_2 > 0$ control the margins between different sentiment labels. For images $x_i$ and $x_j$ with emotional labels $y_i$ and $y_j$, we define $D^*(x_i, x_j) = \theta D(x_i, x_j), \theta \propto \frac{1}{dis(y_i, y_j)}$, where $dis(y_i, y_j)$ denotes 1 + "the number of steps required to reach one emotion from another by the Mikels' wheel". With a weak form of prior knowledge, the sentiment constraint is able to utilize the natural emotion relation that not only the emotions can be divided into two polarities, but also the correlation of the same polar emotion is different. Therefore, the sentiment metric is learned by minimizing the sentiment loss function:

$$\begin{aligned} E_{sml} = &\sum_{i=1}^{N} [D^*(a_i, p_i) - D^*(a_i, r_i) + \alpha_1]_+ \\ &+ \sum_{i=1}^{N} [D^*(a_i, r_i) - D^*(a_i, n_i) + \alpha_2]_+ , \end{aligned} \tag{4}$$

where $N$ is the number of training images. $[\cdot]_+ = \max(0, \cdot)$ since we only need to optimize the situation when the sentiment constraint is violated.

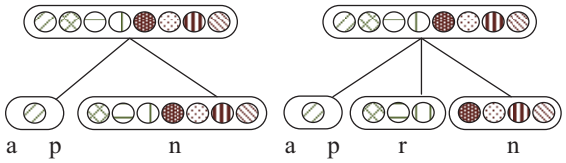


Figure 3: Different from the existing triplet constraint (a) involving anchor, positive and negative samples, our sentiment constraint (b) consists of the anchor, positive, related and negative samples considering the natural polarities of sentiment labels. Moreover, the correlation between the labels with the same polarity is also considered in (3).

## Sentiment Vector for Feature Embedding

To effectively represent image texture, which has been proven as one of the important low-level visual features related to image emotion categorization (Machajdik and Hanbury 2010), we utilize deep representations and propose the sentiment vector consisting of multiple Gram layers, each of which computes inner products of filter responses in a convolutional layer (Gatys, Ecker, and Bethge 2015).

Specifically, we first feed the image into a CNN and compute the response of each intermediate convolutional layer $l \in \{1, 2, \cdots, L\}$. Assume that layer $l$ contains $N_l$ filters and therefore $N_l$ feature maps, with each map of the size of $W_l \times H_l$. Since a CNN usually has tens of filters or even more in a layer, there would be thousands of elements in the Gram matrix which causes heavy computation burden. To improve the generalization ability of the proposed framework, We employ the $1 \times 1$ convolutional layer adding the non-linear activation while shrinking the size of the Gram matrix. Thus the response in layer $l$ can then be stored in a matrix $F^l \in \mathbb{R}^{M_l \times N'_l}$, where $F^l_{ij}$ is the activation of the $j^{\text{th}}$ filter at position $i$ in the layer $l$, $N'_l$ is the number of filters and $M_l = W_l \times H_l$. These feature maps can be represented as a two-dimensional matrix $\Phi^l \in \mathbb{R}^{N'_l \times N'_l}$. Each element in the Gram matrix denotes the correlation between each pair of feature maps, where $\Phi^l_{ij} = \sum_{k=1}^{M_l} F^l_{ki} F^l_{kj}$ is the inner product between the $i^{\text{th}}$ and $j^{\text{th}}$ vectorized feature maps in the layer $l$.

Since the matrix is a symmetrical matrix, the number of independent elements is $N'_l(N'_l + 1)/2$. We define the sentiment vector $SV$ from layer $l$ as

$$SV^l = [\Phi^l_{1,1}, \Phi^l_{2,1}, \Phi^l_{2,2}, \cdots, \Phi^l_{N'_l,1}, \cdots, \Phi^l_{N'_l,N'_l}]. \quad (5)$$

The sentiment vectors from multiple convolutional layers are then concatenated as $SV = [SV^1, SV^2, \cdots, SV^L]$, which is then normalized to unit $l_2$ norm to form the sentiment vector used in sentiment metric learning.

## Multi-Task Framework

In the standard training process, traditional classification constraints such as the softmax loss are optimized to maximize the probability of the correct class. Given a training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$, here, $x^{(i)}$ is the $i^{\text{th}}$ affective image and $y^{(i)} \in \{1, 2, \cdots, C\}$ is the corresponding sentiment label. Let $\{h^{(i)}_j | j = 1, 2, \cdots, C\}$ be the activation values of unit $j$ in the last fully connected layer for $x^{(i)}$, then the fine-tuning of the last layer is done by minimizing the softmax loss:

$$E_{cls} = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{C} \mathbf{1}(y^{(i)} = j) \ln p^{(i)}_j\right], \quad (6)$$

where the indicator function $\mathbf{1}(\delta) = 1$ if $\delta$ is true, otherwise 0. In addition, $p^{(i)}_j$ indicates the probability that the label of $x^{(i)}$ is $j$, which is given by $p^{(i)}_j = \frac{\exp(h^{(i)}_j)}{\sum_{k=1}^{C} \exp(h^{(i)}_k)}$.

The loss of softmax can be seen as the sum of the negative log-likelihood over all training images $\{x_i\}_{i=1}^N$, which penalizes the classification error for each class equally and thus ignore the intra-class variance.

Thus, given the sentiment triplets and the labels of images as input, we explicitly train the deep model to optimize the classification and similarity constraints. Our loss function is integrated with two losses via a weighted combination:

$$E = (1 - \omega)E_{cls} + \omega E_{sml}, \quad (7)$$

where $E_{cls}$ and $E_{sml}$ denote the classification loss and sentiment loss, respectively. $\omega$ is the weight to control the trade-off between the two losses.

# Experimental Results

To evaluate the effectiveness of our proposed method for affective image understanding, we conduct thorough experiments on the affective datasets. In particular, we demonstrate that our learned feature embeddings can be used for affective level image retrieval, with significantly better performance than the state of the art. Meanwhile, our framework also achieves promising classification accuracy compared with several baseline methods.

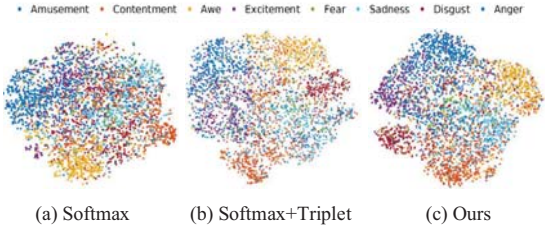
## Datasets

We perform our experiments on four datasets, including Flickr and Instagram (FI) (You et al. 2016), IAPSa, ArtPhoto and Abstract Paintings (Machajdik and Hanbury 2010). FI is collected from social websites by querying with Mikels' eight emotions as keywords. 225 Amazon Mechanical Turk workers were then hired to label the images, which ended up with 23,308 images receiving at least three agrees.[1] The International Affective Picture System (IAPS) (Lang, Bradley, and Cuthbert 2008) is a common stimulus dataset which is widely used in visual emotion understanding research, from which IAPSa selects 395 pictures annotated with the same eight emotion categories. ArtPhoto includes 806 artistic photographs from a photo sharing site. Abstract Paintings contains 228 peer rated abstract paintings consisting of color and texture.

[1]We have 22,713 manually labeled images as some images no longer exist on the Internet.

Table 1: Performance of classification and retrieval tasks on the FI dataset. Here, '$*$', '$\bullet$' and '$\diamond$' denote using different sampling methods (*i.e.*, random sampling, hard sampling, semi-hard sampling). Different embeddings are employed to represent sentiments, *i.e.*, fully-connected layers (FC), sentiment vectors ($SV$).

| Constraints | | | | Feature | Acc.(%) | $\text{mAP}_8$ | $\text{mAP}_2$ |
|---|---|---|---|---|---|---|---|
| Softmax | Center | Triplet | Senti | | | | |
| √ | | | | FC | 65.18 | 0.3583 | 0.6773 |
| | √ | | | FC | 63.14 | 0.3695 | 0.7254 |
| | | √* | | FC | 63.46 | 0.3712 | 0.7278 |
| √ | | √* | | FC | 64.08 | 0.4016 | 0.7456 |
| √ | | √• | | FC | 64.58 | 0.4228 | 0.7532 |
| √ | | √◇ | | FC | 65.35 | 0.4426 | 0.7592 |
| √ | | | √◇ | FC | 65.84 | 0.4575 | 0.7913 |
| √ | | | √◇ | SV | **67.64** | **0.4885** | **0.8098** |




(a) Softmax (b) Softmax+Triplet (c) Ours

Figure 4: Visualization of feature embeddings using t-SNE on the testing set of the FI. Different colors represent different sentiment labels. (a) and (b) show the feature space using the FC feature, while (c) denotes the feature space of using the sentiment vector. As can be seen, our framework can separate emotional categories more effectively.

## Implementation Details

We build our framework based on the GoogleNet-Inception (Szegedy et al. 2015), which achieves state-of-the-art performance in large-scale image classification on ImageNet. First, the network is initialized with the weights trained for the large-scale dataset, and then fine-tuned on the FI dataset with the FC8 layer changed to 8, which is split randomly into 80% training, 5% validation and 15% testing set. The learning rates of the convolutional layers and the last fully-connected layer are initialized as $10^{-4}$ and $10^{-3}$, respectively. We fine-tune all layers by stochastic gradient descent (SGD) through the whole net using batches of 128, which ensures that at least 8 images are contained for each emotion. A total of 100 epochs are run to update the parameters, which is enough for our framework to converge. We set the margin $\alpha$ in the triplet loss to 0.2, while $\alpha_1$ and $\alpha_2$ in the sentiment loss are set to 0.2, 0.1, respectively. We set the weight $\omega$ as 0.7; discussions regarding its sensitivity is given in the following subsection. The feature dimension for the triplet loss is 512. The sentiment vector from each layer is 136 using 16 filters with kernel size of $1 \times 1$, resulting in a total of 680-dimensional features as embeddings. All our experiments are carried out on two NVIDIA GTX 1080 GPUs with 32 GB CPU memory on-board. With the help of transfer learning, we also employ our framework on three datasets with limited training examples. In details, we transfer the parameters of the network fine-tuned on the FI as well as the hyper-parameters to the small-scale datasets, which are split into 80% training and 20% testing set randomly. We conduct 5-fold validation and report the average performance.

## Baseline

We focus on the comparison of different methods, including methods using low-level and mid-level features as well as deep methods. We extract three low-level features including local descriptors like SIFT, HOG and Gabor. The 1,200-dimensional ANP detectors of SentiBank as well as the 2,089-dimensional features from the DeepSentiBank are exploited as mid-level features. LIBSVM (Chang and Lin 2011) is employed for classification. For the CNN-based methods, we focus on the comparison of models trained with different constraints, and various architectures are also evaluated in our experiments, *i.e.*, AlexNet, VGGNet, and GoogleNet. We compare the performance of models pre-trained on the ImageNet as well as models fine-tuned on the affective datasets, where softmax loss is employed for optimization. We show the results of using LIBSVM trained on features extracted from the last FC layer with dimensions reduced by employing PCA. In practice, we find that different cost values (parameter $C$ in LIBSVM) produce similar accuracies, so we just use the default value. Moreover, we compare the CNN directly employing the similarity loss (*e.g.*, triplet loss, center loss) for learning the representation, and also the models jointly optimizing the softmax loss and triplet loss (Zhang et al. 2016).

**Evaluation** To search for an image which has a similar emotion as a given query image, we determine the nearest neighbors in terms of the feature representation as (Wang et al. 2014). For the FI dataset, we use each image in the test set as input to retrieve all the relevant images from the training set. For the three small sets, each image is used as input to retrieve relevant images from the remaining ones following (Zhao et al. 2014b). We evaluate the performance of retrieval using commonly used measurements. Nearest neighbor rate (NN) evaluates the retrieval precision of the first returned result. First tier (FT) and Second tier (ST) denote the recall of the top $m$ and $2m$ retrieval results, where $m$ is the number of the relevant images in the whole dataset. Mean average precision (mAP) denotes the mean precision of the retrieval results. We focus on the mAP of eight emotions (*i.e.*, $\text{mAP}_8$) as well as the mAP of the binary polarities (*i.e.*, $\text{mAP}_2$). Discounted cumulative gain (DCG) measures the importance of different positions of relevant results, assuming that users are more likely to consider the frontal results. Average normalized modified retrieval rank (ANMRR) takes the ranking sequence of relevant images within the retrieved images. All these retrieval measurements range

Table 2: Classification and retrieval performance on the FI dataset. We compare different baselines for learning the sentiment representation, including the traditional methods and CNN based methods. Here, 'S + T' denotes using softmax and triplet loss to jointly train the model. Note that ImageNet is the only model without fine-tuning, while the others have been fine-tuned.

| Algorithm | | Acc.(%) | Retrieval Performance | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $\text{mAP}_\mathbf{8}$ ↑ | $\text{mAP}_\mathbf{2}$ ↑ | FT ↑ | ST↑ | NN ↑ | DCG↑ | ANMRR↓ |
| Baseline | SIFT | 37.56 | 0.1705 | 0.5913 | 0.1830 | 0.3513 | 0.2462 | 0.4507 | 0.6553 |
| | HOG | 44.67 | 0.2115 | 0.6002 | 0.1926 | 0.3620 | 0.3225 | 0.4639 | 0.6424 |
| | Gabor | 36.33 | 0.1724 | 0.5942 | 0.1768 | 0.3395 | 0.2641 | 0.4434 | 0.6770 |
| | SentiBank | 49.09 | 0.2337 | 0.6168 | 0.2422 | 0.4232 | 0.3990 | 0.5223 | 0.5934 |
| | DeepSentiBank | 56.15 | 0.2559 | 0.6247 | 0.2658 | 0.4468 | 0.4583 | 0.5509 | 0.5655 |
| FC-CNN | ImageNet (Softmax) | 47.15 | 0.2376 | 0.6240 | 0.2480 | 0.4309 | 0.4695 | 0.5284 | 0.5863 |
| | AlexNet (Softmax) | 58.13 | 0.2709 | 0.6328 | 0.2795 | 0.4693 | 0.5038 | 0.5633 | 0.5463 |
| | VggNet (Softmax) | 64.55 | 0.3013 | 0.6552 | 0.3007 | 0.4887 | 0.5511 | 0.5860 | 0.5161 |
| | GoogleNet (Softmax) | 65.18 | 0.3583 | 0.6773 | 0.3571 | 0.5619 | 0.5816 | 0.6403 | 0.4517 |
| | GoogleNet (Triplet) | 63.46 | 0.3951 | 0.6981 | 0.3932 | 0.6081 | 0.5578 | 0.6762 | 0.4082 |
| | GoogleNet (S + T) | 65.35 | 0.4426 | 0.7592 | 0.4435 | 0.6513 | 0.5866 | 0.7119 | 0.3603 |
| Ours | | **67.64** | **0.4885** | **0.8098** | **0.4834** | **0.6978** | **0.6023** | **0.7802** | **0.3135** |

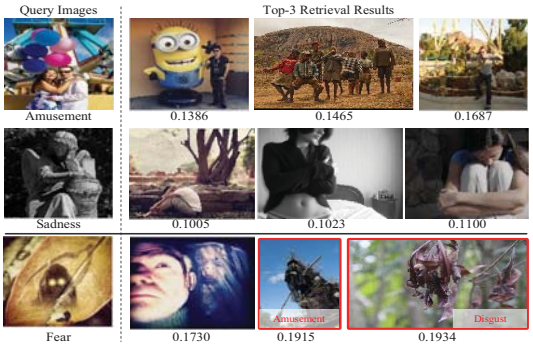


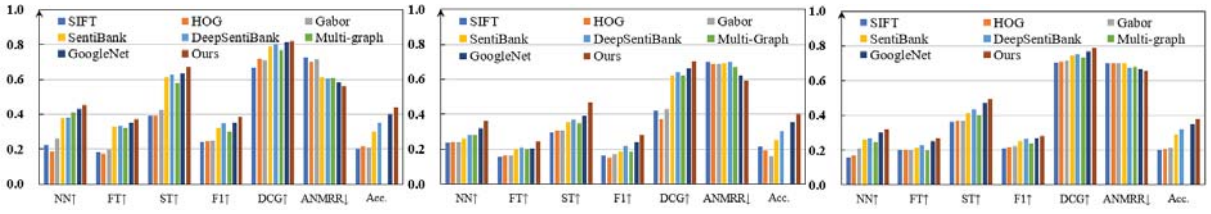

Figure 5: Sample retrieval results of the proposed method from FI. Given the query images (first column), the top-3 retrieval results and the corresponding Euclidean distances are shown. Moreover, a failure retrieval case is shown in the last row with red boxes.

from 0 to 1. A higher value represents better performance for the first five measurements and a lower value indicates better performance for ANMRR.

## Performance Evaluation on the FI dataset

We first evaluate the proposed algorithm against different methods on the current largest FI dataset (You et al. 2016).

**Constraints and Sampling Methods** Table 1 shows the classification accuracy and the retrieval mAP using the feature representations extracted by models trained with different loss functions. Compared to the softmax loss, results of using the metric learning constraints (*e.g.*, triplet and center losses) are more effective for retrieval. However, their classification performance is inferior to the model trained with the softmax loss. For the triplet-based CNN, the triplet sampling method is also crucial since there are $\mathcal{O}(N^3)$ possible triplets on a dataset with $N$ training data. A good triplet sampling method can ensure stable convergence. We observe that during the training process, given the anchor image, randomly selecting the violated triplets leads to slow convergence. When employing the hard sampling that only selects the hardest negatives may unstably lead to bad local minimal, since the hard cases may contain noise and cause overfitting. In this paper, we employ the semi-hard sampling, where all the positive images are considered and the semi-hard negative images are randomly selected in a mini-batch inspired by (Schroff, Kalenichenko, and Philbin 2015) leading to the $\mathcal{O}(N^2)$ sentiment triplets. This sampling method converges more quickly while being less aggressive than the hard sampling, and so we employ the semi-hard sampling in the remaining experiments. Table 1 shows that with the semi-hard sampling, the scheme based on jointly optimizing the softmax and triplet losses achieves better results on both tasks. The model trained with softmax loss and our sentiment constraint further improves $\text{mAP}_\mathbf{8}$ and $\text{mAP}_\mathbf{2}$, which illustrates that our sentiment constraint is able to capture the relation of emotions. Moreover, the performance of our proposed framework using sentiment vectors achieves 67.64% and outperforms the fine-tuned CNN models by 2.5%, which illustrates that utilizing the texture information from the CNN can be more discriminative than the fully-connected layer. The retrieval performance also demonstrates that using the proposed sentiment vectors the embedding is more effective for distinguishing the affective images than the FC feature, by capturing the texture information from the convolutional layers. To provide insights of our promising results for affective image understanding, we use the sentiment vectors from our proposed framework to visualize the feature space after dimensionality reduction (Maaten and Hinton 2008). Figure 4 shows that the features from the sentiment constraint are consistently much better separated than




(a) IAPSa (b) ArtPhoto (c) AbstractPaintings

Figure 6: Classification and retrieval performance on IAPSa, Artphoto and Abstract Paintings. We compare different baselines for learning the sentiment representation, including traditional methods as well as CNN-based methods. Note that the Multi-graph (Zhao et al. 2014b) method is only proposed for affective image retrieval, thus the classification result is not applicable.

ones from the conventional softmax loss, while our method further enlarges the inter-class variance benefited from incorporating emotion relations into feature learning process.

**Affective Image Retrieval** We compare the retrieval performance with different methods on the FI dataset. Table 2 shows that the low-level generic features are not effective for affective image retrieval. The mid-level representations achieve similar performance as the features from ImageNet trained for object classification. The fine-tuned classification models improve the retrieval performance since the models can learn discriminative representations to distinguish emotions, while the CNN trained with the triplet constraint is more effective to locate the images at the affective level. Compared with jointly optimizing the softmax loss and triplet loss, our proposed method improves $\text{mAP}_\mathbf{8}$ and $\text{mAP}_\mathbf{2}$ by a large margin. We also show our top-3 retrieval results from the FI dataset in Figure 5, which illustrate the efficacy of our proposed framework for retrieving affective images. The distances between the retrieved images and the query images are also given reflecting the emotional differences. In the first example, the top-1 retrieved image with smallest distance obviously belongs to amusement with the Minion figure whereas the remaining two with larger distances evoke the amusement emotion in a subtle way. For the second example, all three retrieved images have consistently small distances, and all of them arouse similar emotion as the queried image. The last example shows that since emotion evoked by images may involve higher abstraction, there are also cases that the retrieval results fail to capture the affective-level semantic parts of images.

**Affective Image Classification** We report the classification accuracy of different methods in Table 2. The deep representations outperform the hand-crafted features designed based on several small-scale datasets for specific domains. The fine-tuned CNNs show the discriminative ability to recognize emotions, and the models with deeper architecture can achieve better performance. Training with the triplet loss achieves worse performance than the fine-tuned GoogleNet, as it is more suitable for retrieval rather than classification. Our algorithm optimizing both losses performs favorably against the fine-tuned CNN models by about 2% improvement, which is higher than these compared methods.

**Hyper-parameters** The margins $\alpha$ and $\alpha_1$ are set to 0.2 as a trade-off between the performance and stable training (Schroff, Kalenichenko, and Philbin 2015). Experiments were also conducted to study the influence of changing $\alpha_2$, where stable performance was achieved for $\alpha_2$ in the range of [0.1,0.2]. The effect of parameter $\omega$ in (7) is analyzed, where $\omega$ indicates the weight of similarity constraint term in the optimization objective function. Since the softmax loss may contain more information than a triplet in each iteration, it is sensible to assign a higher weight. Our experiments show that the performance is not sensitive to small variations of $\omega$, *i.e.*, within 0.5% difference in a range of $[0.6, 0.8]$.

### Results on Small-Scale Datasets

We report the results on three widely-used datasets with comparisons to several other state-of-the-art methods. Figure 6 shows the performance of classification and retrieval. Since emotion anger only contains 8 and 3 images in the IAPSa and Abstract Paintings datasets, they are not enough to perform the 5-fold cross validation. Overall, the deep visual features contribute significantly to achieving better results over the manually crafted visual features in both tasks, while our framework achieves the best results illustrating the generalization ability to the small-scale datasets.

## Conclusions

In this work, we propose to incorporate the hierarchical relation of emotions via deep metric learning, and present a multi-task framework that jointly optimizes the classification loss and sentiment loss in an end-to-end manner. We exploit the sentiment constraint for utilizing the emotion relations, and further propose the sentiment vector based on the Gram matrix for the distance comparison between affective images. Extensive experiments show that our algorithm performs favorably against the state-of-the-art approaches on four popular affective datasets for both affective image classification and retrieval tasks.

## Acknowledgments

This research was sponsored by NSFC (61620106008, 61572264), CAST (YESS20150117), Huawei Innovation Research Program (HIRP), and IBM Global SUR award.

## References


Bellet, A.; Habrard, A.; and Sebban, M. 2013. A survey on metric learning for feature vectors and structured data. *arXiv:1306.6709*.

Chang, C., and Lin, C. 2011. LIBSVM: A library for support vector machines. *ACM TIST* 2(3):27:1–27:27.

Chechik, G.; Sharma, V.; Shalit, U.; and Bengio, S. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11:1109–1135.

Chen, T.; Borth, D.; Darrell, T.; and Chang, S. F. 2014a. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. In *arXiv:1410.8586*.

Chen, T.; Yu, F. X.; Chen, J.; Cui, Y.; Chen, Y.-Y.; and Chang, S.-F. 2014b. Object-based visual sentiment concept analysis and application. In *ACM MM*.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.

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

Detenber, B. H.; Simons, R. F.; and Bennett Jr., G. G. 1998. Roll 'em!: The effects of picture motion on emotional responses. *J. Broadcast. & Electr. Media* 42(1):113–127.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. Texture synthesis using convolutional neural networks. *Febs Letters* 70(1):51–55.

Hoffer, E., and Ailon, N. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*.

Lang, P. J.; Bradley, M. M.; and Cuthbert, B. N. 2008. International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. *Tech. Rep. A-8, U. Florida*.

Lin, T.; Cui, Y.; Belongie, S. J.; and Hays, J. 2015. Learning deep representations for ground-to-aerial geolocalization. In *CVPR*.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. RAPID: Rating pictorial aesthetics using deep learning. In *ACM MM*.

Maaten, L. V. D., and Hinton, G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Research* 9:2579–2605.

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

Mikels, J. A.; Fredrickson, B. L.; Larkin, G. R.; Lindberg, C. M.; Maglio, S. J.; and Reuter-Lorenz, P. A. 2005. Emotional category data on images from the International Affective Picture System. *Behavior Res. Methods* 37(4):626–630.

Qian, S.; Zhang, T.; and Xu, C. 2016. Multi-modal multi-view topic-opinion mining for social event analysis. In *ACM MM*.

Rao, T.; Xu, M.; and Xu, D. 2016. Learning multi-level deep representations for image emotion classification. *arXiv:1611.07145*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*.

Wang, W., and He, Q. 2008. A survey on emotional semantic image retrieval. In *ICIP*.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*.

Xu, C.; Cetintas, S.; Lee, K.-C.; and Li, L.-J. 2014. Visual sentiment prediction with deep convolutional neural networks. *arXiv:1411.5731*.

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

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

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

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

Zhang, H.; Yang, Z.; Gönen, M.; Koskela, M.; Laaksonen, J.; Honkela, T.; and Oja, E. 2013. Affective abstract image classification and retrieval using multiple kernel learning. In *ICONIP*.

Zhang, X.; Zhou, F.; Lin, Y.; and Zhang, S. 2016. Embedding label structures for fine-grained feature representation. In *CVPR*.

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

Zhao, S.; Yao, H.; Yang, Y.; and Zhang, Y. 2014b. Affective image retrieval via multi-graph learning. In *ACM MM*.

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

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

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; and Ding, G. 2017b. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia* 19(3):632–645.