Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval

Xingxu Yao¹, Dongyu She¹, Sicheng Zhao², Jie Liang¹, Yu-Kun Lai³, Jufeng Yang¹
¹Nankai University ²University of California, Berkeley ³Cardiff University
{yxxhb, sherry656}@163.com, schzhao@gmail.com, liang27jie@163.com
LaiY4@cardiff.ac.uk, yangjufeng@nankai.edu.cn

Abstract

Images play a crucial role for people to express their opinions online due to the increasing popularity of social networks. While an affective image retrieval system is useful for obtaining visual contents with desired emotions from a massive repository, the abstract and subjective characteristics make the task challenging. To address the problem, this paper introduces an Attention-aware Polarity Sensitive Embedding (APSE) network to learn affective representations in an end-to-end manner. First, to automatically discover and model the informative regions of interest, we develop a hierarchical attention mechanism, in which both polarity- and emotion-specific attended representations are aggregated for discriminative feature embedding. Second, we present a weighted emotion-pair loss to take the inter- and intra-polarity relationships of the emotional labels into consideration. Guided by attention module, we weight the sample pairs adaptively which further improves the performance of feature embedding. Extensive experiments on four popular benchmark datasets show that the proposed method performs favorably against the state-of-the-art approaches.

1. Introduction

With the increasing popularity of online social networks, people are more likely to express their opinions through posting images on social platforms such as Flickr and Instagram. Recently, affective image analysis that studies the emotional response of humans on visual stimuli has drawn attention from both psychologists [38, 49, 32] and computer vision researchers [30, 63] due to its wide applicability, e.g. opinion mining [36, 39], image captioning [8, 31], etc.

How to search affective images based on human perception is a meaningful yet challenging task. Various emotion-based image retrieval (EBIR) systems have been proposed [54, 24, 34, 65]. Compared to content-based image retrieval (CBIR), EBIR involves high-level abstract semantics and human perception subjectivity. To bridge the “affective gap” between low-level features and high-level affective semantics, some hand-crafted features are proposed according to psychology and art theory [30, 68]. To capture the semantic similarity among affective images, Zhao et al. [72] employ multi-graph learning for affective image retrieval based on features of different levels including low-level color, texture, and other high-level features that contribute to expressing image emotions. More recently, deep learning has been harnessed to predict emotions evoked by images via embedding images into a feature space [58, 41, 52, 47], which results in breakthrough performance. Pang et al. [37] develop a unit density model over the multi-modal space using a deep Boltzmann machine, which enables emotion-oriented cross-modal retrieval. Yang et al. [57] propose a multi-task framework to simultaneously optimize the classification and retrieval losses, in which the performances of both tasks are boosted.

Figure 1. Illustration of retrieving affective images in the embedding space. The two regions in the space represent binary sentiment polarities, i.e. positive and negative. For the given query image, the retrieved image from exactly the same emotion category is shown in a green box, while the images from the same polarity but different category and the opposite polarity are in blue and red boxes, respectively.
However, there are two important characteristics in visual emotion (shown in Fig. 1), which are neglected in existing methods for affective image retrieval. On the one hand, informative regions of interest are crucial to image emotion (see the heat map of each sample image) [12, 3, 50], which can evoke emotional stimuli to people; on the other hand, there exist sentiment polarities in emotional label space other than concrete categories. Note that polarity indicates the coarse-level classes {positive, negative}, and the concrete-level emotions are defined as {amusement, contentment, awe, excitement, fear, anger, disgust, sadness} as per [32, 63]. In this paper, the term ‘class’ is utilized to mean both sentiment polarity and emotion category. Given a query image, our goal is to rank the retrieved images according to the relationship with the given image in the following order: the same emotion category, the same polarity but different emotion categories, different polarity.

In the paper, we propose an attention-aware polarity sensitive embedding (APSE) network for affective image retrieval according to aforementioned characteristics of visual emotion. In detail, there exists a correlation between sentiment polarity and low-level features [42, 29, 68], while specific emotion categories are mainly determined by semantic content. Therefore, in the attention module, we utilize the polarity-specific attention in lower layers of the network, and exploit emotion-specific attention in higher layers. In the embedding process, we introduce a polarity sensitive feature embedding strategy based on the proposed weighted emotion-pair (WEP) loss. We separate binary sentiment polarities in the embedding space, while also effectively distinguishing different emotions in the same polarity. Guided by the attention module, hard negative examples are imposed stronger penalty so as to improve the learning performance. The unified architecture is optimized by the total loss consisting of WEP and attention losses to learn discriminative feature embedding.

Our contributions are twofold. 1) We propose to take multi-level attended local features into account for affective image retrieval, based on the observation that low-level and high-level image features concern different levels of the emotion hierarchy. 2) We introduce an attention-aware polarity sensitive embedding (APSE) network, which takes the inter- and intra-polarity relationships of the emotional labels into consideration. Our proposed WEP loss effectively connects the attention module and embedding process for more effective learning. Extensive experiments demonstrate the effectiveness of the proposed method.

2. Related Work

2.1. Visual Emotion Analysis

In the field of visual emotion analysis, most existing methods focus on emotion prediction [73, 35, 70, 57, 62, 69, 40, 23]. Early work uses a variety of hand-crafted features [30, 60] including shape features [29] and principles-of-art features [68] to represent the emotions evoked by images. In addition, Borth et al. [2] propose adjective noun pairs (ANP) to bridge the affective gap between low-level features and high-level emotion semantics. With extensive applications of deep learning models, numerous methods [52, 41, 74] exploit convolutional neutral networks (CNNs) to extract deep features for emotion representations, which perform well on image emotion classification [6, 56, 58], emotion label distribution prediction [71, 67], and affective image retrieval [72].

While many methods have been devoted to image emotion prediction, far less attention is paid to affective image retrieval. Wang et al. [54] propose an EBIR system that allows users to perform retrieval using sentiment semantic words, and the system is further improved for different tasks [24, 34]. Zhao et al. [72] utilize multi-graph learning to retrieve affective images that are similar to the query image in emotion. A deep framework which simultaneously optimizes the classification and retrieval tasks is proposed in [57]. Different from the existing methods, we develop a polarity sensitive embedding method based on multi-level attended features for affective image retrieval.

2.2. Visual Attention Mechanism

Attention mechanism is widely used in various visual tasks [44, 55, 1, 66, 5, 4, 11], since it can find image regions that play a decisive role in networks. Wang et al. [51] train deep residual networks for image classification by introducing an attention based learning method. SCA-CNN network integrating spatial and channel-wise attention is proposed in [4] for image captioning. According to psychological theory [50, 12], affective content is easier to hold human attention than non-affective content. Unlike specific salient objects which have well-defined boundaries, the region arousing emotion may be ambiguous and abstract [56].

For affective images, prior methods [59, 61] detect emotional attention regions from numerous candidate bounding boxes, increasing the computational burden. Our method generates soft attention maps with the single shot based on the feature activations in an end-to-end manner. Moreover, we integrate features from multiple layers and build a hierarchical attention mechanism for learning robust representations in the embedding space. That is, both polarity-specific features from lower layers and emotion-specific features from higher layers are combined together in our framework.

2.3. Feature Embedding Learning

Recently, numerous methods have utilized embedding learning to measure image similarity for various tasks [28, 9, 17, 64, 53, 20]. Based on the popular pairwise loss [10],
Song et al. [33] utilize a matrix consisting of pairwise distances of the mini-batch to create a loss function which incorporates all samples to form a lifted embedding structure. In order to produce effective training samples, Harwood et al. [18] conduct a smart mining procedure to train the model effectively. In addition, Duan et al. [15] employ deep adversarial learning to generate hard negatives from easy negatives for building more robust models. Motivated by the fact that emotional classes have a hierarchical relationship, i.e., from coarse polarity to concrete emotions, we develop polarity-sensitive WEP loss to measure the similarity of the query and the retrieved images.

3. Methodology

We propose APSE network which can be trained in an end-to-end manner. It contains two main closely related components, as shown in Fig. 2. First, the proposed method integrates polarity- and emotion-specific attended features extracted by hierarchical attention mechanism (Sec. 3.1). Second, we learn polarity-sensitive and discriminative feature embedding by optimizing WEP loss guided by the attention module (Sec. 3.2).

3.1. Hierarchical Attention Mechanism

In addition to the regions for specific emotions obtained from higher layers in the deep network, we also learn the attended regions for specific polarities from lower layers. We propose a simple yet effective attention mechanism (Fig. 3), whose module consists of attention head and output head, which is applied to both attention levels.

The attention head receives the $l^{th}$ level feature activations $F^l \in \mathbb{R}^{c \times h \times w}$ as input, and outputs $K^l$ attention maps, where $c$, $h$, and $w$ are the number of channels, and the height and width of the feature activations, and $K^l$ represents the number of corresponding labels for layers at the $l^{th}$ level. First, we sum up the received feature activation tensor through the channel direction. Thus, an $h \times w$ 2-D aggregation map $A^l$ is derived from 3-D feature activations $F^l$, i.e., $A^l = \sum_{n=1}^{c} F^l_{n}$. Then a spatial attention mask $Z^l$ is obtained by spatial-wise softmax operation on $A^l$. Based on $Z^l$, we implement spatial-wise attention on the feature activations $F^l$ resulting in spatially-attended feature maps, i.e., $F^l_\text{att} = F^l \odot Z^l$, where $\odot$ denotes Hadamard Product by broadcasting, i.e. repeating $Z^l$ for each channel of $F^l$. Then a $1 \times 1$ conv layer is applied to reduce the dimension of $F^l_\text{att}$ to $K^l \times h \times w$, denoted as $S^l \in \mathbb{R}^{K^l \times h \times w}$, with each 2-D feature activation corresponding to a sentiment polarity or specific emotion category depending on the level. $S^l_\text{norm}$ is put through a global average pooling layer and a softmax layer successively, resulting in confidence score vector $C^l$ whose elements lie in the range of $[0, 1]$ and sum to 1.

The output head at the $l^{th}$ level receives 2-D feature activations $S^l$ and corresponding confidence scores. Each confidence score $c$ can be regarded as the degree of tendency towards the corresponding class. Therefore, final attention map $U$ is obtained by adding up all 2-D feature activations $S_j$ weighted by confidence scores:

$$U = \text{norm} \left( \sum_{j=1}^{K} c_j S_j \right),$$  \hspace{1cm} (1)$$

where $\text{norm}$ denotes the normalization operation. Note that
3.2. Polarity Sensitive Embedding Learning

In this section, considering the polarity characteristic of sentiment, we propose the polarity sensitive emotion-pair (EP) loss inspired by N-pair loss. In the embedding process, sample pairs are further adaptively weighted based on confidence scores from the attention module, generating WEP loss. Specifically, the harder anchor-negative pairs are to separate, the higher the weight of them should be, so as to augment their proportion when training the network.

**Review on N-pair loss.** Given \( N \) categories, the N-pair loss function proposed in [46] optimizes to identify a positive example from \( N - 1 \) negative examples. Define \( \{(f_1, f_i^+), \cdots , (f_N, f_i^-)\} \) as \( N \) pairs of convolution features from \( N \) different categories, where \( f_i \) denotes the \( i^{th} \) category anchor point, and \( f_i^\pm \) represents a positive example of the \( i^{th} \) category. Meanwhile, \( f_i^- \) can also be regarded as a negative example of the \( j^{th} \) category (\( \forall i \neq j \)). The value of \( f_i^+ f_i^- \) has positive correlation with the similarity between \( f \) and \( f^+ \). Therefore, the N-pair loss function can be formulated as

\[
\mathcal{L}_{np} = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \sum_{j \neq i} \exp(f_i^T f_j^- - f_i^T f_j^+)). \tag{3}
\]

**EP loss.** In general, N-pair loss can embed features effectively and efficiently. However, for affective image retrieval, the polarity characteristic cannot be considered by the approach directly. Therefore, it is essential to differentiate different negative examples based on their polarity when learning feature embedding. More specifically, image features from the same polarity should be more similar than those from opposite polarity. Therefore, our proposed inter-polarity loss is formulated as

\[
\mathcal{L}_{inter} = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(f_i^T f_j^+ - f_i^T f_j^-)) - \frac{1}{N_{P_i} N_{Q_i}} \sum_{j \in P_i, i \neq j} \sum_{j \in Q_i} (f_i^T f_j^+ - f_i^T f_j^-)), \tag{4}
\]

where \( P_i \) and \( Q_i \) denote the sets of emotion categories with the same and opposite polarities to the anchor from the \( i^{th} \) category, respectively. \( N_{P_i} \) and \( N_{Q_i} \) are the numbers of the corresponding categories.

The inter-polarity loss is very important for affective image retrieval, because it can largely avoid dramatic failure that the retrieved result has many images with opposite sentiment polarity, which may cause unpleasant user experience. That is, the inter-polarity loss ensures the returned images are consistent with query images in sentiment polarity. Further, the more challenging task is to learn discriminative feature embedding within the same polarity. To achieve this, we introduce a new intra-polarity loss to differentiate similar categories in the same polarity as follows:

\[
\mathcal{L}_{intra} = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \sum_{j \in P_i, i \neq j} \exp(f_i^T f_j^- - f_i^T f_j^+)). \tag{5}
\]

Therefore, the EP loss is obtained by combining inter-polarity loss and intra-polarity loss as:

\[
\mathcal{L}_{ep} = \mathcal{L}_{inter} + \mathcal{L}_{intra}. \tag{6}
\]
Weighting sample pairs. Given an affective image $I$, we can obtain its confidence scores regarding both polarity and emotion as demonstrated in Sec. 3.1. For an anchor $I_a^i$ from the $i$th category and one of its negative samples $I_a^j$ from the $j$th category, a higher confidence of $I_a^i$ w.r.t. the $j$th category or $I_a^j$ w.r.t. the $i$th category denotes that the pair is harder to separate. Consequently, we assign a stronger penalty term on this pair in the training process.

Specifically, $c_{ij}$ and $c_{ji}^p$ represent the confidences of $I_a^i$ w.r.t. the $j$th category in polarity- and emotion-level, while $c_{ij}^p$ and $c_{ij}^e$ represent the confidences of $I_a^j$ w.r.t. the $i$th category in polarity- and emotion-level. The weights are formed as

$$v_{ij} = \exp(c_{ij}^p) \cdot \exp(c_{ij}^e), \quad (7)$$

$$v_{ji} = \exp(c_{ji}^p) \cdot \exp(c_{ji}^e), \quad (8)$$

where $v_{ij}$ denotes the polarity-level weight of the pair constructed by $I_a^i$ and $I_a^j$, and $v_{ij}$ denotes the emotion-level weight of the pair constructed by $I_a^i$ and $I_a^j$. Note that $v_{ij}$ will be set to 1 if the $i$th and $j$th categories belong to the same polarity. Then $v_{ij}$ and $v_{ji}$ form the weight matrix $V$ and $V'$ respectively as shown in Fig. 2, whose diagonal elements are set to 1 (i.e. $v_{ii} = 1$, $c_{ii}^e = 1$). The final weight $\tilde{v}_{ij} = v_{ij} \cdot v_{ji}$. The value of $\tilde{v}_{ij}$ ($i \neq j$) determines the importance during learning. We set the weight of any anchor-positive pair to be 1, i.e., $\tilde{v}_{ii} = 1$. Therefore, we introduce WEP (weighted EP) loss:

$$L_{wep} = \frac{1}{N} \sum_{i=1}^{N} \log[(1 + \exp(\frac{1}{N_{qi}} \sum_{j \in P_i} \tilde{v}_{ij} f_i^T f_j^p)) (1 + \sum_{j \in P_i, i \neq j} \exp(\tilde{v}_{ij} f_i^T f_j^p)) - f_i^T f_i^p)]. \quad (9)$$

We define the total loss consisting of attention and WEP losses to optimize the unified embedding network simultaneously:

$$L_{total} = \lambda L_{wep} + (1 - \lambda) L_{att}, \quad (10)$$

where $\lambda$ is the weight to control the trade-off between two types of losses.

4. Experiments

In this section, we conduct extensive experiments on the most commonly used affective datasets to evaluate the proposed algorithm against the state-of-the-art methods.

4.1. Datasets

We perform our experiments on four popular datasets, including Flickr and Instagram (FI) [63], Subset A of IAPS (IAPSa) [32], Artistic dataset (ArtPhoto) [30], and Abstract paintings (Abstract) [30]. FI is collected from social websites by querying Mikels’ eight emotions as keywords, resulting in 23,308 labeled images. IAPSa consists of 395 images collected from International Affective Picture System (IAPS) [32], while ArtPhoto contains 806 artistic photographs searched by emotion categories. The Abstract is composed of 228 peer rated abstract paintings which contain abundant color and texture.

4.2. Evaluation Metrics

Following previous work [72, 57], we adopt the following metrics as our evaluation criteria. The mean precision of the retrieval results are represented by mean Average Precision (mAP). We concern both mAP of eight emotion-specific categories (mAP$e$) and mAP of the two polarities (mAP$\pm$). Nearest neighbor rate (NN) denotes the proportion of the rank-1 sample belonging to the same category with the query. First tier (FT) and second tier (ST) both represent the recall of the retrieval results. FT denotes the top-$n$ recall, while ST is defined as the top-$2n$ recall. Here, $n$ is the number of all the positive examples. Discounted cumulative gain (DCG) [21] measures the importance of different positions of relevant samples in the ranking sequence of returned results. $F_1$ score is a measure combining Precision and Recall, as their harmonious mean. Average normalized modified retrieval rank (ANMRR) [16] considers the ranking sequence of relevant images within the retrieved results. Smaller values of ANMRR represent better retrieval results, while for other evaluation metrics the larger the better.

4.3. Baselines

We compare the proposed method with different baselines. First, we extract low-level local descriptors (i.e. SIFT and HOG), whose dimensions are fixed to 1000. We also extract mid-level features, including 1200-dimensional ANP detectors of SentiBank [2], 2089-dimensional features of DeepSentiBank [7], and 4342-dimensional features of MVSO (English) [22]. For CNN methods, we fine-tune deep models with softmax loss based on different architectures, including AlexNet, VggNet, GoogleNet, and ResNet-50, and extract the features from the last FC layer for retrieval. Also, we train different embedding learning methods based on ResNet-50, including contrastive loss [10], triplet loss [43], N-pair loss [46], and retrieve images using 2048-dimensional features. We also compare with the state-of-the-art methods for affective image retrieval, including Yang et al. [57] and Multi-Graph [72].

4.4. Implementation Details

Following [57], each image in the test set of FI dataset is treated as a query image to retrieve relevant images in the training set. For small-scale datasets, we use each image to retrieve the rest of images. We rank the retrieved images based on the similarity between them and the query image.

The proposed framework is based on ResNet-50 [19] pre-trained on the ImageNet [14]. The original images are
Table 1. Retrieval performance on the FI dataset. We evaluate the proposed method against different algorithms, including traditional methods (TRA), existing CNN models (CNN), and embedding learning methods (EMB). Note that ‘S’ denotes using softmax loss for training, and ‘Dim.’ represents the dimension of features.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dim.</th>
<th>mAPs ↑</th>
<th>mAP↑</th>
<th>FT↑</th>
<th>ST↑</th>
<th>NN↑</th>
<th>DCG↑</th>
<th>ANMRR↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT [27]</td>
<td>1000</td>
<td>0.1705</td>
<td>0.5913</td>
<td>0.1830</td>
<td>0.3513</td>
<td>0.2462</td>
<td>0.4507</td>
<td>0.6553</td>
</tr>
<tr>
<td>HOG [13]</td>
<td>1000</td>
<td>0.2115</td>
<td>0.6002</td>
<td>0.1926</td>
<td>0.3620</td>
<td>0.3225</td>
<td>0.4639</td>
<td>0.6424</td>
</tr>
<tr>
<td>Sentibank [2]</td>
<td>1200</td>
<td>0.2337</td>
<td>0.6168</td>
<td>0.2422</td>
<td>0.4232</td>
<td>0.3900</td>
<td>0.5223</td>
<td>0.5934</td>
</tr>
<tr>
<td>DeepSentiBank [7]</td>
<td>2089</td>
<td>0.2559</td>
<td>0.6247</td>
<td>0.2658</td>
<td>0.4468</td>
<td>0.4583</td>
<td>0.5509</td>
<td>0.5655</td>
</tr>
<tr>
<td>AlexNet (S) [25]</td>
<td>4096</td>
<td>0.2709</td>
<td>0.6328</td>
<td>0.2795</td>
<td>0.4693</td>
<td>0.5038</td>
<td>0.5633</td>
<td>0.5636</td>
</tr>
<tr>
<td>VggNet (S) [45]</td>
<td>4096</td>
<td>0.3013</td>
<td>0.6552</td>
<td>0.3007</td>
<td>0.4887</td>
<td>0.5511</td>
<td>0.5860</td>
<td>0.5161</td>
</tr>
<tr>
<td>GoogleNet (S) [48]</td>
<td>2048</td>
<td>0.3583</td>
<td>0.6773</td>
<td>0.3571</td>
<td>0.5619</td>
<td>0.5816</td>
<td>0.6403</td>
<td>0.4517</td>
</tr>
<tr>
<td>ResNet (S) [19]</td>
<td>2048</td>
<td>0.3830</td>
<td>0.7068</td>
<td>0.4286</td>
<td>0.6079</td>
<td>0.6084</td>
<td>0.6816</td>
<td>0.3998</td>
</tr>
<tr>
<td>WSCNet [56]</td>
<td>4096</td>
<td>0.5060</td>
<td>0.7381</td>
<td>0.4653</td>
<td>0.6223</td>
<td>0.6558</td>
<td>0.6910</td>
<td>0.3872</td>
</tr>
<tr>
<td>Contrastive loss (ResNet) [10]</td>
<td>2048</td>
<td>0.3842</td>
<td>0.6972</td>
<td>0.3768</td>
<td>0.5702</td>
<td>0.5711</td>
<td>0.6508</td>
<td>0.4396</td>
</tr>
<tr>
<td>Triplet loss (ResNet) [43]</td>
<td>2048</td>
<td>0.5130</td>
<td>0.7120</td>
<td>0.4864</td>
<td>0.6216</td>
<td>0.5710</td>
<td>0.6843</td>
<td>0.3860</td>
</tr>
<tr>
<td>N-pair loss (ResNet) [46]</td>
<td>2048</td>
<td>0.5217</td>
<td>0.8062</td>
<td>0.4785</td>
<td>0.7075</td>
<td>0.5341</td>
<td>0.7310</td>
<td>0.3089</td>
</tr>
<tr>
<td>Yang et al. (GoogleNet) [57]</td>
<td>640</td>
<td>0.4885</td>
<td>0.8098</td>
<td>0.4834</td>
<td>0.6978</td>
<td>0.6023</td>
<td>0.7802</td>
<td>0.3135</td>
</tr>
<tr>
<td>Yang et al. (ResNet) [57]</td>
<td>544</td>
<td>0.6395</td>
<td>0.8081</td>
<td>0.5995</td>
<td>0.7354</td>
<td>0.6164</td>
<td>0.7866</td>
<td>0.2518</td>
</tr>
<tr>
<td>APSE (ours)</td>
<td>512</td>
<td>0.7344</td>
<td>0.9079</td>
<td>0.6985</td>
<td>0.7817</td>
<td>0.6613</td>
<td>0.8114</td>
<td>0.2201</td>
</tr>
</tbody>
</table>

Figure 4. Retrieval performance on the three small datasets (Artphoto, Abstract, and IAPSa).

Figure 5. Effect of $\lambda$ for total loss on mAPs and mAP↑ testing on FI dataset. Note that $\lambda$ is the weight of $\mathcal{L}_{wep}$, and $1 - \lambda$ is the weight of $\mathcal{L}_{att}$.

We evaluate the retrieval performance with different methods on four affective datasets. As shown in Tab. 1, resized to 256 $\times$ 256 followed by a center 224 $\times$ 224 cropping. We initialize the learning rate as 0.001 and drop down one-tenth every 40 epochs. The gross number of epochs is 100 for fine tuning all layers by stochastic gradient descent (SGD) with a batch size of 32 ensuring images from each emotion. We optimize the parameters of the framework by SGD with the weight decay of 0.0005 and a momentum of 0.9. Considering both effectiveness and consumption of parameters, we choose the features from last layer of conv3 and conv5 to represent the low-level and high-level features, respectively. For contrastive and triplet losses, we set the margin $\gamma$ to 0.4 and 0.2 respectively. We adopt the semi-hard triplet sampling method in triplet loss. In our architecture, the dimension of the output embedding feature after being compacted is 512 according to the experience from [26]. The FI dataset is split randomly into 80% training, 5% validation, and 15% testing sets. For small-scale datasets, we transfer the parameters of the network fine-tuned on FI to them. 5-fold validation is performed and the average performance is reported.

4.5. Retrieval Performance

We evaluate the retrieval performance with different methods on four affective datasets. As shown in Tab. 1,
Table 2. Ablation experiments on the FI dataset. The fundamental framework is ResNet-50 pre-trained on ImageNet. Here, AT represents the attention loss consisting of two softmax losses. HA denotes hierarchical attention, and SA denotes the emotion-specific attention on the last convolutional layer. CLB represents cross-level bilinear operation. SO means using the feature from the last layer, and MO means using the feature from the last layer of both conv3 and conv5, respectively. When CLB is not selected, the features from different layers are concatenated directly. The weights of all parts in the combined loss are the same.

<table>
<thead>
<tr>
<th>AT</th>
<th>N-pair</th>
<th>EP</th>
<th>WEP</th>
<th>SA</th>
<th>HA</th>
<th>CLB</th>
<th>SO</th>
<th>MO</th>
<th>mAP$_5$↑</th>
<th>mAP$_2$↑</th>
<th>FT↑</th>
<th>ST↑</th>
<th>NN↑</th>
<th>DCG↑</th>
<th>ANMRR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.4380</td>
<td>0.7068</td>
<td>0.4286</td>
<td>0.6079</td>
<td>0.6084</td>
<td>0.6816</td>
<td>0.3998</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.5217</td>
<td>0.8062</td>
<td>0.4785</td>
<td>0.7075</td>
<td>0.5341</td>
<td>0.7310</td>
<td>0.3089</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.5680</td>
<td>0.8558</td>
<td>0.5247</td>
<td>0.7187</td>
<td>0.5623</td>
<td>0.7602</td>
<td>0.2789</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.6225</td>
<td>0.7816</td>
<td>0.5779</td>
<td>0.7255</td>
<td>0.5975</td>
<td>0.7451</td>
<td>0.2623</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.6630</td>
<td>0.8241</td>
<td>0.6036</td>
<td>0.7485</td>
<td>0.6110</td>
<td>0.7863</td>
<td>0.2551</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.6680</td>
<td>0.8325</td>
<td>0.6365</td>
<td>0.7504</td>
<td>0.6278</td>
<td>0.7885</td>
<td>0.2421</td>
</tr>
</tbody>
</table>

Table 2. Ablation experiments on the FI dataset. The fundamental framework is ResNet-50 pre-trained on ImageNet. Here, AT represents the attention loss consisting of two softmax losses. HA denotes hierarchical attention, and SA denotes the emotion-specific attention on the last convolutional layer. CLB represents cross-level bilinear operation. SO means using the feature from the last layer, and MO means using the feature from the last layer of both conv3 and conv5, respectively. When CLB is not selected, the features from different layers are concatenated directly. The weights of all parts in the combined loss are the same.

we compare our proposed method with traditional methods, CNN-based methods and other embedding learning methods on FI. We can see that current popular deep representations outperform the hand-crafted features. In general, embedding learning methods get remarkable improvements in all evaluation metrics other than NN as presented in Tab. 1, compared with the CNN architectures trained by softmax loss. This is because softmax loss only concerns the location of single data rather than a holistic distribution in metric space. In addition, we compare our method with other competitive and influential embedding learning approaches as well as the state-of-the-art algorithms. For fair comparison, we also implement the state-of-the-art [57] using ResNet-50 architecture as this work. Our framework improves about 10% on mAP$_5$ and mAP$_2$ respectively as compared to state of the art. The other evaluation metrics are also improved obviously.

For other three small-scale datasets, we transfer the model trained on the FI dataset for fine-tuning on the target datasets. As reported in Fig. 4, we draw similar conclusions on the small-scale datasets as FI, where the proposed method still obtains the best retrieval results. This illustrates that our framework has robust generalization ability.

4.6. Influence of Parameter $\lambda$

In Eqn. (10), the value of $\lambda$ controls the relative importance between the WEP loss and attention loss. The bigger the value of $\lambda$ is, the more important the WEP loss is. We use the two essential metrics, which are mAP$_5$ and mAP$_2$, on FI dataset to demonstrate how $\lambda$ influences the performance of total loss on FI. Note that the two losses are not isolated absolutely, so we only concern the results with $\lambda$ ranging from 0.1 to 0.9. As shown in Fig. 5, we can find through the curves that: (1) mAP$_5$ is more sensitive than mAP$_2$ for the variation of $\lambda$; (2) When $\lambda = 0.5$, mAP$_5$ and mAP$_2$ both achieve the best performance. On the whole, the values of the two metrics are stable, which demonstrates that our method is robust for affective image retrieval.

4.7. Ablation Study

In order to demonstrate the contribution of different components in the proposed method, we further examine the advantage of each component through ablation experiments on FI dataset. First, AT is the attention loss consisting of two softmax losses on conv3 and conv5, respectively. As shown in the first part of Tab. 2, our EP loss has obvious superiority compared with the softmax and N-pair losses in all criteria. The results on mAP$_5$ and mAP$_2$ illustrate the architecture optimized by the EP loss improves the precision of retrieved images considering sentiment polarities other than specific emotions. As can be seen, integrating the AT and EP losses can enhance the performance on all the evaluation criteria other than mAP$_2$, because they benefit each other in the process of training. On the one hand, the AT provides category-specific cues for EP loss; on the other hand, the AT in the last convolution layer neglects the distinction between polarity, resulting in a weak decline on mAP$_2$, which can be recovered in our attention mechanism and multi-level output.

In addition, experiments are also performed to verify the effect of attention mechanism as shown in the second part of Tab. 2. The result of only using SA exceeds about 3% on both mAP$_5$ and mAP$_2$ compared with the performance of framework without any attention. Furthermore, hierarchical attention mechanism also has obvious benefits compared with SA, when both of them utilize features from both conv3 and conv5. It demonstrates that the attended features from different levels are complementary, resulting in improvement on overall retrieval performance.

In order to make the features from different levels interact effectively, the cross-level bilinear (CLB) is exploited to integrate multi-level information, leading to further per-
Figure 6. Top 5 results of sample query images from the FI dataset. (a) are sample query images from FI. (b-c) are the retrieval results of networks trained by the N-pair loss and our method, respectively. Image frames with different colors represent different emotions.

Figure 7. Visualization of attention maps from different levels. The images from the FI dataset are presented in column (a), and the visualizations of polarity- and emotion-specific attention results are presented in column (b) and column (c), respectively. The classes of the two sample images are disgust and sadness, respectively.

We present some attention visualization results of samples in Fig. 7. The polarity-specific attention considers the distinct color or texture details which can represent certain emotional tendency. Although these regions scatter in the image, they carry significant information which contributes to the specific emotion involved in the image. In the first image, the polarity-specific attention regions cover a great mass of blood. It guides to disgust emotion as the cue and enhance the high-level attention features in some ways. The ragged and shabby wall in the second image is attended by polarity-specific attention, while the region containing the person is drawn more attention in the emotion-specific attention map. Therefore, the polarity-specific attention can supplement this deficiency of emotion-specific attention.

5. Conclusion

In this paper, we propose an attention-aware polarity sensitive embedding network for affective image retrieval. The polarity- and emotion-specific attended features are integrated effectively. We present a weighted emotion-pair (WEP) loss, which constrains features from inter- and intra-polarity respectively. Then the sample pairs are weighted based on confidence scores derived from attention module adaptively. Finally, the total loss consisting of WEP and attention losses is exploited to optimize the architecture. Extensive experiments on four datasets indicate that our method outperforms the state-of-the-art approaches.

Acknowledgment

This work was supported by the NSFC (No.61876094, 61701273, U1933114), Natural Science Foundation of Tianjin, China (No.18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), the Fundamental Research Funds for the Central Universities, and Berkeley Deep-Drive.
References


[35] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Con-


