Deep Cascaded Cross-modal Correlation Learning for Fine-grained Sketch-based Image Retrieval

Yanfei Wang , Fei Huang , Yuejie Zhang , Rui Feng , Tao Zhang , Weiguo Fan

PII: S0031-3203(19)30449-2
DOI: https://doi.org/10.1016/j.patcog.2019.107148
Reference: PR 107148

To appear in: *Pattern Recognition*

Received date: 24 April 2019
Revised date: 4 November 2019
Accepted date: 4 December 2019

Please cite this article as: Yanfei Wang , Fei Huang , Yuejie Zhang , Rui Feng , Tao Zhang , Weiguo Fan , Deep Cascaded Cross-modal Correlation Learning for Fine-grained Sketch-based Image Retrieval, *Pattern Recognition* (2019), doi: https://doi.org/10.1016/j.patcog.2019.107148

Highlights

- A simple yet effective pipeline for FG-SBIR is created through combining all the beneficial multimodal cues involved in sketches and annotated images.
- A deep cascaded neural network architecture with deep representation, embedding, and ranking is established for revealing multimodal relationships.
- Two extended image datasets are collected to validate the generalization ability of our scheme, which demonstrates its effectiveness for both SBIR and FG-SBIR.

Deep Cascaded Cross-modal Correlation Learning for Fine-grained Sketch-based Image Retrieval

Yanfei Wang[1], Fei Huang[1], Yuejie Zhang[1, *], Rui Feng[1], Tao Zhang[2, *], and Weiguo Fan[3]

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2]School of Information Management and Engineering, Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai, China
[3]Department of Business Analytics, Tippie College of Business, University of Iowa, USA

Yuejie Zhang and Tao Zhang are the corresponding authors, yjzhang@fudan.edu.cn, taozhang@mail.shufe.edu.cn

Abstract

Fine-grained Sketch-based Image Retrieval (FG-SBIR), which utilizes hand-drawn sketches to search the target object images, has recently drawn much attention. It is a challenging task because sketches and images belong to different modalities and sketches are highly abstract and ambiguous. Existing solutions to this problem either focus on visual comparisons between sketches and images and ignore the multimodal characteristics of annotated images, or treat the retrieval as a one-time process. In this paper, we formulate FG-SBIR as a coarse-to-fine process, and propose a Deep Cascaded Cross-modal Ranking Model (DCCRM) that can exploit all the beneficial multimodal information in sketches and annotated images and improve both the retrieval efficiency and the top-*K* ranked effectiveness. Our goal concentrates on constructing deep representations for sketches, images, and descriptions, and learning the optimized deep correlations across such different domains. Thus for a given query sketch, its relevant images with fine-grained instance-level similarities in a specific category can be returned, and the strict requirement of the instance-level retrieval for FG-SBIR is satisfied. Very positive results have been obtained in our experiments by using a large quantity of public data.

Keywords: Fine-grained Sketch-based Image Retrieval (FG-SBIR), Deep Cascaded Cross-modal Correlation Learning, Deep Multimodal Representation, Deep Multimodal Embedding, Deep Triplet Ranking.

1. Introduction

With the development of touch-screen technology, drawing sketches has become a simple and efficient way for people to express their visual perceptions and query intentions [1][2]. Thus sketch, as a new modality, has attracted wide interests in computer vision research, notably in Sketch-based Image Retrieval (SBIR) which utilizes query sketches to retrieve relevant color images in large-scale image collections [3][4]. Most existing SBIR approaches focus on the category-level matching between sketches and images, which can be viewed as a classification problem, i.e., given a query sketch to retrieve the images with the same category label [5]. However, such methods decrease the descriptive power of sketches and cannot

capture the important visual properties of intra-category variation at a fine-grained level, such as pose, viewpoint, texture, shape, etc. [6]. Thus, similar to Fine-grained Image Classification [7], it is imperative to pay close attention to mine the category-specific correspondences between human-drawn sketches and natural color images at the fine-grained instance level, that is, Fine-grained SBIR (FG-SBIR). An illustration of FG-SBIR vs. SBIR is shown in Fig. 1.

Fig. 1. An illustration of FG-SBIR vs. SBIR. Given a query sketch, SBIR aims to retrieve the relevant images with the same category label, e.g., the images in the orange dotted box are all correct matching results because they have the same category label of "*church*" with the query sketch. FG-SBIR is more challenging due to only considering the relevant images with fine-grained instance-level similarities for the query sketch, e.g., the image in the green dotted box is the only correct matching result because it is more similar to the query sketch in visual details such as *window*, *door* and so on than other images.

In addition, an image is generally exhibited in a form with different modalities (i.e., visual and semantic), such as a web image with user defined annotation tags/a narrative text description. However, free-hand drawn sketches are less discriminative, and the inherent ambiguities in sketches cannot be well handled only by exploiting the visual information. Due to the visual gap between sketches and images and the semantic gap between sketches and descriptions, there may be significant differences and independence among sketches, visual images, and semantic texts for annotated images. This leads to a huge difficulty and a high uncertainty in making full use of the relationships between the visual features (in sketches and images) and the semantic

features (in descriptions). Thus integrating the valuable multimodal information sources in sketches and annotated images to enable more refined sketch-image matching has been a key component for supporting more effective FG-SBIR.

Although SBIR has been extensively studied since recent years, FG-SBIR is still an extremely challenging task that deserves more studies for optimal solutions [8][9]. It places particular emphases on discovering the most relevant images involving more unique fine-grained visual attributes and details, which are both visually and semantically correlated with the query sketch. Due to the obviously distinct appearance across inherently heterogeneous domains of sketches, images, and descriptions, three inter-related issues should be addressed simultaneously for FG-SBIR: 1) valid characterization for formulating valuable multimodal attribute features in order to bridge the deep representational gap among sketches, images, and descriptions; 2) reasonable modeling for learning deep cross-modal correlations between sketches and images/descriptions at the instance level in order to acquire more objective pairwise sketch-image matching; and 3) appropriate optimization for determining better fine-grained correspondences between sketches and images in order to make further in-depth understanding of query sketches. To address the first issue, the appropriate attribute elements for sketches, images, and descriptions need to be explored for achieving deep multimodal feature representations. To address the second issue, a feasible cross-modal correlation modeling scheme needs to be established for mapping the fine-grained attributes of different modalities into a common embedding space, so as to acquire their instance-level statistical dependencies and correlations. To address the third issue, an efficient ranking optimization strategy needs to be developed with high accuracy but low cost, in order to maximize the inter-related cross-modal correlations for sketch-image pairs.

To meet the above challenges, a novel scheme with Deep Cascaded Cross-modal Correlation Learning is developed in this paper to facilitate more robust FG-SBIR on large-scale annotated images. Our goal focuses on constructing deep representations for sketches, images, and descriptions, and learning the optimized deep correlations across such multiple different domains. Thus for a given query sketch, its relevant images with fine-grained instance-level similarities in a specific category can be returned, and the strict requirement of the instance-level retrieval for FG-SBIR is

satisfied. Our experiments on large-scale public data sources have obtained very positive results.

The main contributions of this paper are three folds. First, we are concerned on how to integrate visual and textual cues for FG-SBIR with deep convolutional neural networks, and present a simple yet effective pipeline through combining all the beneficial multimodal cues involved in sketches and annotated images. Our results show that the textual descriptions for images are complementary to the visual information, which result in the significant performance improvement for FG-SBIR. To the best of our knowledge, such a multimodal/multi-domain scenario has not been well studied in the previous literatures. We expect it could become a new research direction that can bring fresh insights and novel applications. Second, we present a deep cascaded neural network architecture with deep representation, embedding, and ranking for implicitly revealing the multimodal visual and semantic relationships between visual and textual cues. Such a unified framework is more elegant and efficient than the existing alternatives designed for cross-modal representation and correlation learning. Both offline training and online querying are significantly different with the conventional algorithms. Our framework can serve as a novel trial towards the challenging problem of FG-SBIR, and also provides a baseline performance which is convenient to be compared with future work. Third, in order to further validate the generalization ability of our proposed scheme for the general SBIR, we collect two extensional category-level image datasets consisting of various intersection object categories. Our method also works well on the general coarse-grained SBIR for such natural color images, which demonstrates its deep potential in practical application for both SBIR and FG-SBIR. We will make these two datasets publicly available, and gradually add new image data sources to enlarge the data scale.

The rest of the paper is organized as follows. Section 2 briefly reviews some related works. In Section 3, we describe in detail our new enhanced framework on integrating multimodal visual and textual cues for FG-SBIR with deep cascaded cross-modal correlation learning. Section 4 gives our experimental results and analyses on the algorithm evaluation, and we conclude the paper in Section 5.

2. Related Works

We will first briefly review three lines of related works, i.e., FG-SBIR, Deep Learning for FG-SBIR, and Cross-modal Analysis for FG-SBIR. The former two are drawn up from the research on the target image retrieval at the fine-grained instance level using free-hand sketches. The latter one focuses on building deep cross-modal correlations among multimodal heterogeneous data for better retrieval.

2.1 Fine-grained SBIR (FG-SBIR)

Largely limited by the representation power of hand-crafted features, the general SBIR does not have the strong ability to achieve the instance-level retrieval that requires distinguishing the subtle differences among the images of the same category [10]. Because it is not realistic to obtain the strong annotations of object bounding boxes and part components for a large number of images, more SBIR methods attempt to retrieve fine-grained images using only image-level labels [11]. Thus FG-SBIR has become a relatively new popular task in the past few years, that is, given a query sketch to retrieve the target images that belong to the fine-grained meta category at the instance level. FG-SBIR is an extremely challenging problem, yet not stressed in previous studies for general SBIR.

The first work toward FG-SBIR was proposed by Li et al. [12], which established the Deformable Part-based Model (DPM) to learn the mid-level sketch representation, and then used the graph matching to discover the pose correspondences between sketches and images. However, these hand-crafted features cannot bridge the cross-modal gap in deep level. Yu et al. [13] further extended the definition of fine-grained and proposed a new dataset of sketch-photo pairs with detailed triplet annotations. They developed a deep triplet-ranking network to learn a fine-grained feature metric, but avoided addressing the cross-modal gap by converting photos to edge maps prior to training and testing. Xu et al. [14] formulated a cross-domain framework specifically designed for the task of FG-SBIR that simultaneously conducted instance-level retrieval and attribute prediction. They proposed a joint view selection and attribute subspace learning algorithm to learn domain projection matrices for photo

and sketch respectively. Xu et al. [15] further introduced and compared a series of state-of-the-art cross-modal subspace learning methods and benchmarked them on two fine-grained SBIR datasets. They demonstrated that the subspace learning could effectively model the sketch-photo domain-gap through detailed experiments. Despite such early success, the challenging task of FG-SBIR remains largely unsolved, especially how they can be extended to work cross-modal as for SBIR.

The recent work of Li et al. [16] remained one of the few works that specifically tackled the cross-modal nature of the FG-SBIR problem, where they used three-view Canonical Correlation Analysis (CCA) [17] to fuse fine-grained visual attributes and low-level features. However, they did not learn a joint feature space since CCA was only conducted independently on each domain, and required the separately trained set of attribute detectors at test time, which made it less generalizable to other datasets. Wang et al. [18] proposed a deep ranking model that learned fine-grained image similarities directly from images via learning to rank with image triplets. However, such methods are either mainly based on visual contents without considering semantic attributes for retrieval, or limited to small-scale FG-SBIR datasets.

It is worth noting that in the real SBIR environment, query sketches are always taken as the only input unimodal visual contents, while the multimodal information of annotated images is ignored. Thus much closer attention has been given to the methods that rely on exploiting multimodal attributes and grouping visually similar and semantically related sketches and images. Huang et al. [19] explored the textual contexts of images and introduced a multimodal embedding model for FG-SBIR by jointly learning the mutual correlations among sketches, images, and texts, in which the final similarity score was obtained by summing up both the sketch-image and sketch-text similarities. Song et al. [20] proposed a FG-SBIR model that exploited the semantic attributes and the deep feature learning in multi-task learning paradigm, which involved three learning tasks of retrieval by the fine-grained ranking on a learned representation, an attribute prediction, and an attribute-level ranking. Recently, Sangkloy et al. [21] proposed a large-scale database, "*Sketchy Database*", as a benchmark dataset for FG-SBIR and tested several popular cross-modal convolutional network architectures. This database contained fine-grained associations between particular photos and sketches, and could be utilized to train

cross-modal convolutional networks which embedded sketches and photographs in a common feature space.

2.2 Deep Learning for FG-SBIR

Most image retrieval approaches are based on local features and feature aggregation strategies, e.g., Vector of Locally Aggregated Descriptor (VLAD) [22] and Fisher Vector (FV) [23]. With the great progress of deep learning in the area of computer vision, especially with Convolutional Neural Networks (CNNs) leading the advances in image classification, CNN features, as a global representation, are increasingly exploited in image retrieval and exhibit the beneficial boosting for FG-SBIR [24]. CNNs have revolutionized FG-SBIR, and provide a unified framework for learning the descriptor extraction and classification that yields major performance gains over pipelines based upon prescriptive gradient features. After the success of CNN, FG-SBIR also embraces deep learning, and explores the out-of-the-box features from pre-trained deep networks to achieve state-of-the-art results [25].

Babenko et al. [26] retrained a CNN model on several datasets which were similar to queries and extracted features for retrieval. Without a doubt, they obtained excellent performance. An important reason was that the features extracted from the retrained model retained the high-level semantic information for the original image. Lin et al. [27] introduced a deep learning framework to learn binary hash codes for fast image retrieval, which was superior to several state-of-the-art hashing algorithms. Ng et al. [28] explored the features in different layers of the deep network for image retrieval and found that the deeper layers lost the local features which were important for the instance-level image retrieval. Liu et al. [29] proposed a method to directly model the relationship between texts and clipart images by the co-occurrence relationship between words and visual words, which improved traditional SBIR, provided a baseline performance, and obtained more relevant results in the condition that all images in the database did not have any text tag.

In recent years, a CNN model, "*Sketch-a-Net*" was developed for sketch recognition by Yu et al. [30], and achieved the state-of-the-art recognition performance on the *TU-Berlin* dataset [31]. Yu et al. [13] further utilized *Sketch-a-Net*

as the basic network architecture in their FG-SBIR model, and introduced two new modifications of pre-training and sketch data augmentation to improve *Sketch-a-Net*. Song et al. [32] further improved Yu et al.'s work in [13] by introducing an attention module, combining coarse and fine semantic information via a shortcut connection fusion block, and using HOLEF loss to model feature correlations between sketches and images. Lu et al. [33] proposed a new Deep Triplet Classification Siamese Network (*DeepTCNet*), which employed *DenseNet*-169 [34] as the basic feature extractor and was optimized by the triplet loss and classification loss. Although some efforts studied what deep descriptors and deep neural networks could be used and how to use them in FG-SBIR and achieved some certain beneficial results, these approaches were usually designed for general image retrieval and SBIR, which were quite different from and did not work well for FG-SBIR.

2.3 Cross-modal Analysis for FG-SBIR

In FG-SBIR, information about the same instance/object can be easily obtained from various sources, such as sketches, images, and descriptions. Generally, information from different modalities is complementary and offers useful knowledge to each other [35]. Cross-modal analysis methods have shown the superior performance over unimodal ones for image retrieval, especially for FG-SBIR [36]. Despite some early success, the problem remains largely unsolved, especially how they can be extended to work with cross-modal data in the case of SBIR [37].

Cross-modal analysis is a classic task in multimedia information retrieval, and does not impose restrictions on modality types of the queries and retrieved results. The challenge is finding a semantic feature space that can withstand the modality variation at an abstract level. Most techniques project multimodal data to a common space, and then the similarity of multimodal data can be computed in such a common space [38]. With the progressive development of deep learning, some deep-learning-based methods are adopted for cross-modal analysis in retrieval, in which different modalities are mapped into a unified representation space by the deep architecture. Frome et al. [39] proposed a deep visual-semantic embedding model (DeViSE), which connected two modalities by cross-modal mapping using the linear

transformation and hinge rank loss. Zhuang et al. [40] proposed a deep cross-modal hashing method, named as Cross-Media Neural Network Hashing (CMNNH). With incorporating the representation learning and correlation learning into a single process, Feng et al. [41] built three deep learning models, which were found to be effective in cross-media retrieval. Jiang et al. [42] utilized the deep learning model to learn a multimodal embedding while enhancing both the local and global alignment, but they only focused on the pairwise correlation.

Although there are some existing works [28, 29, 30, 35, 37, 38] for solving the cross-modal analysis problem in retrieval, most of them focus on learning correlated features and cannot achieve the high recall and high ranking at the same time, especially for FG-SBIR. Meanwhile, all aforementioned cross-modal models cannot work with instance-level annotations (e.g., sketch-image pairs), which largely limits their applicability for fine-grained retrieval. Thus it is important to establish more reasonable cross-modal analysis for mining a joint subspace where cross-modal comparisons can be done at a fine-grained level and achieving higher retrieval effectiveness in an efficient manner.

3. A New FG-SBIR Framework with Deep Cascaded Cross-modal Correlation Learning

3.1 Overview of the New Framework

We create a new FG-SBIR framework with deep multimodal representation, embedding, and ranking to match sketch-image pairs in a coarse-to-fine fashion, which can return the most similar images through mining all the beneficial multimodal attributes, as shown in Fig. 2. 1) A deep multimodal feature representation is proposed to obtain better deep representations for sketches and annotated images by formulating a deep-level representation independently in each domain. The element of "*deep feature*" is created for encoding both the visual feature in a sketch/an image (i.e., deep visual feature) and the semantic feature in a textual description (i.e., deep semantic feature). Compared to traditional features, the deep visual features that are closer to the sketch/image semantics can alleviate the problem

of visual/semantic gap to a great degree, and the deep semantic features that integrate various relationship information among textual elements can be more representative for description semantics. 2) A deep correlation modeling scheme is designed for crossing the matching barrier between query sketches and natural images with descriptions, in which the deep multimodal embedding and correlation learning are fused together to break the limitation of modality consistency. Compared to the traditional correlation learning, such cross-modal correlation learning considers the heterologous property for different modalities. A specific mapping function is utilized to map different modalities into a common embedding space for achieving the precise characterization of inter-related multimodal correlations. 3) A deep ranking optimization mechanism is introduced to further improve the fine-grained sketch-image matching by dynamically adjusting the inter-related correlations in the final ranking. A novel similarity function with the deep triplet ranking loss is specially explored to minimize the large margin objective function for obtaining better pairwise similarity. In contrast to visual-similarity-based approaches, such re-ranking can strengthen the fine-grained variation of interest for the instance-level retrieval, which ensures the retrieved images and sketch query have as similar visual appearance as possible and enables the precise matching with fine-grained details.

3.2 Deep Multimodal Representation

Since the multimodal information is a significant expression and exhibition for sparse sketches and annotated images, the optimal basic elements for multimodal contents should be detected more precisely. Thus the deep multimodal feature representation is implemented to exploit multiple content elements in deep level, i.e., deep visual feature and deep semantic feature, so as to explore the multimodal associations between them.

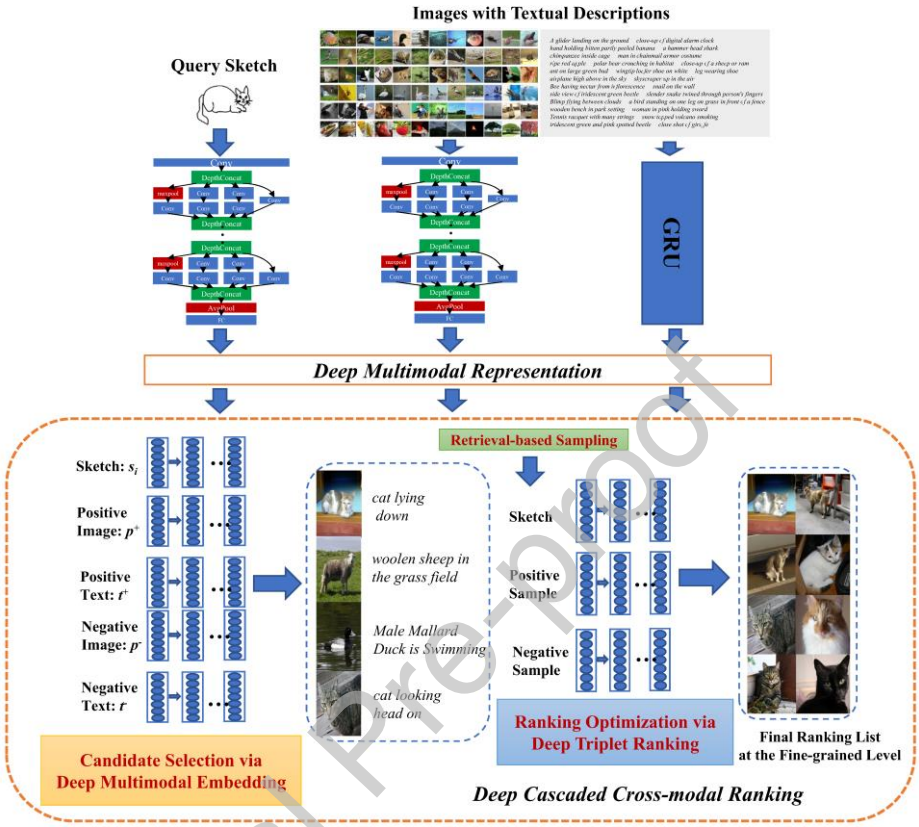


Fig. 2. An overview of our new FG-SBIR Framework with deep cascaded cross-modal correlation learning.

3.2.1 Deep Visual Feature Representation for Image and Sketch

Each image can be represented with a Convolutional Neural Network (CNN) descriptor, and we choose the classification network as our base model. Since the selection of different CNN architectures is not our main concern, we simply use *GoogLeNet* [43] to extract deep visual features for images. *GoogLeNet* is pre-trained on 1.2 million images of *ImageNet* [25]. Each image is converted to a fixed pixel size of 224×224, and then fed into the network. The 1,024-way average pool layer output after the last inception module is taken as the deep visual representation for an image.

Compared to the traditional descriptors like *HOG* [44], such deep features are closer to image semantics for visual recognition.

Since sketches and images are two domains of visual exhibitions, those CNN models trained on real images cannot be directly applied to sketches. The existing datasets for FG-SBIR are usually imbalanced with many images but without many sketch training samples. Thus following the same *GoogLeNet* architecture, a specific sketch-like generation, pre-training, and fine-tuning mechanism is conducted for expanding sketch samples and transferring the image-based CNN models to sketches.

a. Sketch-like Image Generation -- A classic *Canny* edge detector [45] is first utilized to produce the chaotic edge images, but many edge pixels cannot provide the relevant structural information and may introduce more noises. Thus we select a globalPb contour detector [31] to detect the edge pixels in an image, because it can capture the important boundary information and combine multiple local cues into a globalization framework with the spectral clustering. The globalized probability is calculated for each pixel in an image, which can be defined as a weighted sum of local and spectral signals, as shown in Eq. (1).

$$Pb(x, y, \theta) = \sum\sum \beta_{i,s} G_{\sigma(i,s)}(x, y, \theta) + \gamma . sPb(x, y, \theta) \qquad (1)$$

where s is the index scale; i denotes the index feature channel; $G_{i,\ \sigma(i,\ s)}(x,\ y,\ \theta)$ measures the histogram difference in the channel i between two halves of a disc with the radius $\sigma(i,\ s)$ centered at $(x,\ y)$ and divided by a diameter at the angle θ; $sPb(x, y, \theta)$ provides the spectral component of the globalPb contour detector; $\beta_{i,\ s}$ and γ are two weights learned by the gradient ascent on the training images from *Sketchy* [21].

It is worth noting that in the training set from *Sketchy*, each image corresponds to a unique sketch. After the above edge detection, only the important edge pixels are preserved for each image. However, the number of pixels in an edge map may be still more than those in the relevant sketch, thus we especially introduce a "*screening*" strategy. Given an edge map E and a relevant sketch S, N_E denotes the number of edge pixels in E and N_S is the number of non-zero pixels in S. The edge pixels are first sorted in descending order according to the globalized probability values, and the edge pixels with the smaller values are discarded until N_E is no more than 20% of N_S. After that, we can successfully generate sketch-like images, each of which

corresponds to its raw image with fine-grained similarity. Thus it can greatly extend our sketch training samples and help learn more robust sketch representations for *FG-SBIR*.

b. Pre-training on Sketch-like Images -- The first stage is to transfer the original CNN descriptor to the sketch domain. It is reasonable to assume that the sketch-like images obtained by our sketch-like image generation retain the main outlines of initial images, which can be approximately treated as hand-drawn sketches and utilized as the training samples. In *ImageNet*, only the images provided with bounding boxes are considered, and each bounding box is transformed to the edge map form. Thereby, the generated 1,000 categories of sketch-like images can be exploited to retrain *GoogLeNet*.

c. Fine-tuning on Real Sketches -- With 250 categories of *TU-Berlin*, the pre-trained *GoogLeNet* model is then fine-tuned on hand-drawn sketches to learn the ability for better representing real sketches. The sketch samples are expanded by performing the sketch data augmentation via the stroke removal and deformation in [13]. Finally, 30 new sketches are created per initial sketch in our training set. The fine-tuned network thus far has been optimized for the category-level sketch recognition, and is appropriate to the sketch representation. Similar to the image representation, we also extract the 1,024-dimensional feature of average pooling layer after the last inception module from the retrained *GoogLeNet* as the deep visual feature for each sketch.

3.2.2 Deep Semantic Feature Representation for Description

To extract the semantic features of image descriptions effectively, an appropriate semantic representation model should be explored and applied. It is worth noting that the selection of semantic representation model is less sophisticated, as long as the reasonable semantic features can be extracted. Thus considering the simplicity and usability of the Skip-thought model [46], we adopt it for obtaining better semantic feature representations for image descriptions. The Skip-thought model aims at learning deep sentence vector representations, which is good at mapping the sentences that share similar semantics and syntactics to similar vector representations. Its advantage is that the training is unsupervised by using the continuity of surrounding

sentences and the vocabulary of words can be easily extended online. Specifically, the Skip-thought model follows the encoder-decoder framework, in which the encoder learns the feature vectors of sentences and the decoder learns to generate the surrounding sentences. Given the triplet adjacent sentences ($\boldsymbol{S}_{i-1}$, $\boldsymbol{S}_i$, $\boldsymbol{S}_{i+1}$), let $\boldsymbol{X}^t_i$ be the *word2vector* representation of the t^{th} word in the sentence $\boldsymbol{S}_i$ and N be the number of words in the sentence. For the encoder, a GRU is used and a hidden state $\boldsymbol{h}^t_i$ is produced at each time step, which can be formulated as:

$$\boldsymbol{z}^t = \sigma(\boldsymbol{W}_z \cdot [\boldsymbol{h}^{t-1}, \boldsymbol{X}^t]) \tag{2}$$

$$\boldsymbol{r}^t = \sigma(\boldsymbol{W}_r \cdot [\boldsymbol{h}^{t-1}, \boldsymbol{X}^t]) \tag{3}$$

$$\widetilde{\boldsymbol{h}}^t = tanh(\boldsymbol{W} \cdot [\boldsymbol{r}^t \cdot \boldsymbol{h}^{t-1}, \boldsymbol{X}^t]) \tag{4}$$

$$\boldsymbol{h}^t = (1 - \boldsymbol{z}^t) * \boldsymbol{h}^{t-1} + \boldsymbol{z}^t * \widetilde{\boldsymbol{h}}^t \tag{5}$$

where $\boldsymbol{z}^t$ is the update gate vector; $\boldsymbol{r}^t$ is the reset gate vector; and $\widetilde{\boldsymbol{h}}^t$ is the state update vector at the time step *t*. Thus $\boldsymbol{h}^N_i$ can be interpreted as the feature vector for $\boldsymbol{S}_i$. For the decoder, two GRUs are used, in which one is for generating the previous sentence $\boldsymbol{S}_{i-1}$ and the other for generating the next sentence $\boldsymbol{S}_{i+1}$. These two GRUs are trained separately without sharing any parameters. Since they share the same computation pattern, we formulate the decoding process of the next sentence $\boldsymbol{S}_{i+1}$ as:

$$\boldsymbol{z}^t = \sigma(\boldsymbol{W}_z \cdot [\boldsymbol{h}^{t-1}, \boldsymbol{X}^{t-1}, \boldsymbol{h}_i]) \tag{6}$$

$$\boldsymbol{r}^t = \sigma(\boldsymbol{W}_r \cdot [\boldsymbol{h}^{t-1}, \boldsymbol{X}^{t-1}, \boldsymbol{h}_i]) \tag{7}$$

$$\widetilde{\boldsymbol{h}}^t = tanh(\boldsymbol{W} \cdot [\boldsymbol{r}^t \cdot \boldsymbol{h}^{t-1}, \boldsymbol{X}^{t-1}, \boldsymbol{h}_i]) \tag{8}$$

$$\boldsymbol{h}^t_{i+1} = (1 - \boldsymbol{z}^t) * \boldsymbol{h}^{t-1} + \boldsymbol{z}^t * \widetilde{\boldsymbol{h}}^t \tag{9}$$

where $\boldsymbol{h}^t_{i+1}$ is the hidden state of the decoder at the time step *t*, and its computation is analogous to the encoder except that the computation is conditioned on the feature vector $\boldsymbol{h}_i$ of the sentence $\boldsymbol{S}_i$. The sum of log-probabilities for the previous and next sentences is used as the corresponding objective function to guide the training for the triplet adjacent sentences ($\boldsymbol{S}_{i-1}$, $\boldsymbol{S}_i$, $\boldsymbol{S}_{i+1}$).

$$\sum_t logP(\boldsymbol{w}^t_{i+1} | \boldsymbol{w}^{<t}_{i+1}, \boldsymbol{h}_i) + \sum_t logP(\boldsymbol{w}^t_{i-1} | \boldsymbol{w}^{<t}_{i-1}, \boldsymbol{h}_i) \tag{10}$$

where $\boldsymbol{w}^1_i$, …, $\boldsymbol{w}^N_i$ denote the words in $\boldsymbol{S}_i$. Thus the loss function of the Skip-thought model can be achieved by summing up all the training triplets. A large corpus from *BookCorpus* [47] is utilized to train the Skip-thought model. We follow the combine-

skip mode in [46], use the learned encoder as the feature extractor, and extract 4,800-way vector as the deep semantic feature for each image description. With the consideration of semantic relationships among descriptions, such features can be more representative for the semantic contents in annotated images.

3.3 Deep Cascaded Cross-modal Ranking

To achieve precise correlations among multimodal contents in sketches and annotated images, a multimodal correlation modeling needs to be established for evaluating cross-modal sketch-image associations. Our Deep Cascaded Cross-modal Ranking Model (DCCRM) leverages these facts to achieve both the high recall and the top-*K* ranked effectiveness by mining all the valuable multimodal contents in a coarse-to-fine way. DCCRM learns the cross-modal correspondences in two cascaded stages. At the first stage, a Deep Multimodal Embedding Model (DMEM) learned on three modalities of sketch, image, and description is exploited to measure the inter-related sketch-image correlations, and then the top-*K* similar candidates are found at both visual and semantic levels. At the second stage, a Deep Triplet Ranking Model (DTRM) is utilized to learn for improving the top-*K* ranked retrieval effectiveness using multimodal embedded features.

3.3.1 Candidate Selection via Deep Multimodal Embedding

In order to integrate three modalities (sparse sketches, visual images, and textual descriptions) for cross-modal correlation learning, we construct a Deep Multimodal Embedding Model (DMEM) to map them into a common space, where the relevant three modalities are associated with each other, as shown in Fig. 3. DMEM learned on three modalities is exploited to measure the inter-related sketch-image correlations. This model works as a sorting model with the aim to find the top-*K* similar candidates at both visual and semantic levels.

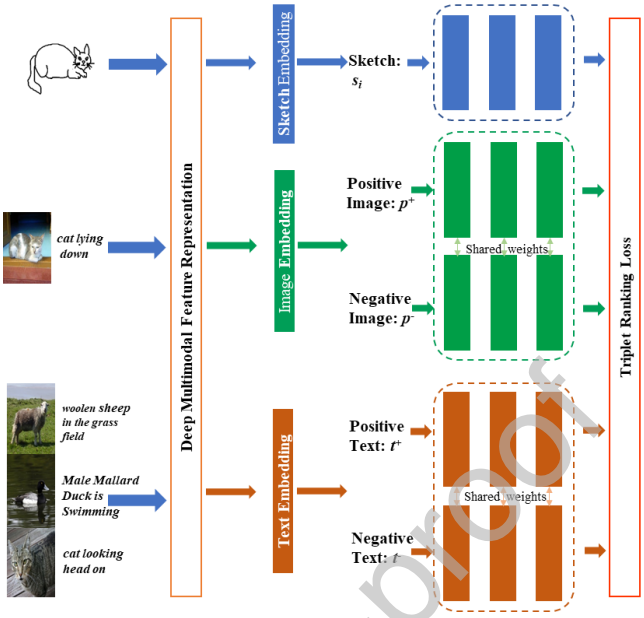


Fig. 3. The architecture of our Deep Multimodal Embedding Model (DMEM).

The mapping function of DMEM is composed of multiple stacked layers of nonlinear transformation. Each layer takes the output of the previous layer $h_{i-1} \in R^{di-1}$ to compute its output $h_i = \sigma(W_i h_{i-1} + b_i) \in R^{di}$, where $W_i \in R^{di \times di-1}$ and $b_i \in R^{di}$ are the weight matrices and biases for the i^{th} layer respectively. The outputs of the last layer are taken as the final correlated features after training. To obtain better discriminative mapping function, we use the annotated quintuple $\{(\boldsymbol{s}_i, \boldsymbol{p}^+_i, \boldsymbol{t}^+_i, \boldsymbol{p}^-_i, \boldsymbol{t}^-_i)\}^N_i$ as the supervised information. Each quintuple consists of a query sketch $\boldsymbol{s}$ and two images $\boldsymbol{p}^+$ and $\boldsymbol{p}^-$ with their descriptions $\boldsymbol{t}^+$ and $\boldsymbol{t}^-$, in which $\boldsymbol{p}^+$ with $\boldsymbol{t}^+$ and $\boldsymbol{p}^-$ with $\boldsymbol{t}^-$ are named as the positive and negative sample respectively. The positive samples are selected from the target image set that shares both the fine-gained visual and semantic similarity to the query sketch, while the negative ones are selected from the residual irrelevant sets. Thus DMEM has five subnets with a shared architecture for the quintuple input. Each subnet takes the corresponding deep visual/semantic features as the input and produces a fixed-length output feature vector. Since the positive and negative images/descriptions are in the same modality, their subnets share the same parameters. DMEM aims at learning a nonlinear function $F(\bullet|\theta)$ to map the input deep

visual/semantic features to a common space, in which the images and descriptions relevant to a query sketch are closer than those irrelevant ones, which can be formulated as:

$$DF(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+) < DF(\boldsymbol{s}_i, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-) \quad (11)$$

$$DF(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+) = \left\| F(\boldsymbol{s}_i|\boldsymbol{\theta}_s) - F(\boldsymbol{p}_i^+|\boldsymbol{\theta}_p) \right\|^2 + ||F(\boldsymbol{s}_i|\boldsymbol{\theta}_s) - F(\boldsymbol{t}_i^+|\boldsymbol{\theta}_t)||^2 \quad (12)$$

$$DF(\boldsymbol{s}_i, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-) = \left\| F(\boldsymbol{s}_i|\boldsymbol{\theta}_s) - F(\boldsymbol{p}_i^-|\boldsymbol{\theta}_p) \right\|^2 + ||F(\boldsymbol{s}_i|\boldsymbol{\theta}_s) - F(\boldsymbol{t}_i^-|\boldsymbol{\theta}_t)||^2 \quad (13)$$

where $DF(*)$ represents the distance calculation function; $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_p$, and $\boldsymbol{\theta}_t$ denote three parameters of the corresponding subnets for sketch, image, and description respectively. To achieve this goal, we extend the classic ranking loss to adjust itself to three modalities, which has the stronger ability to characterize cross-modal sketch-image correlations among deep visual/semantic features in sketches, images, and descriptions. Given a training set of quintuples $\psi=\{(\boldsymbol{s}_i, \boldsymbol{p}^+_i, \boldsymbol{t}^+_i, \boldsymbol{p}^-_i, \boldsymbol{t}^-_i)\}^N_i$, the corresponding objective function can be defined as:

$$min \sum_{i\epsilon\psi} L(i) + \lambda||\boldsymbol{\theta}||_2^2 \quad (14)$$

$$L(i) = \max(0, \boldsymbol{m} + DF(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+) - DF(\boldsymbol{s}_i, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-)) \quad (15)$$

where $\boldsymbol{m}$ is a margin to control the relative distance between the positive and negative pairs; and $\boldsymbol{\theta}$ denotes the parameter of DMEM. The optimization for the objective function will adjust the parameter $\boldsymbol{\theta}$ for each subnet, so as to obtain the desired feature mapping function that satisfies the ranking order. With adequate training, we can feed the deep visual/semantic features of three modalities to DMEM and conduct the initial retrieval on the learned common space. Given a query sketch $\boldsymbol{s}$ and a sample of annotated image $\boldsymbol{p}$ with the textual description $\boldsymbol{t}$, the distance between a query sketch and an annotated image can be computed by $DF(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{t})$. Thus the similarity between the query sketch and each annotated image in the whole dataset can be measured at both visual and semantic levels, and a rank list of candidate relevant images is produced by sorting the similarity values of sketch-image pairs. The top-K ranked images are then selected as the candidates and transmitted to the subsequent ranking optimization.

The process of Deep Multimodal Embedding Model is summarized in Algorithm 1.

Algorithm 1: Deep Multimodal Embedding Model (DMEM)

Input: Training dataset: $\boldsymbol{Z} = \{\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{t}\}_{i=1}^{n}$, DMEM function $\boldsymbol{F}\left(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-; \theta\right)$, Number of epochs: E
Output: DMEM function: $\boldsymbol{F}(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-; \boldsymbol{\theta})$.
1. Input selection:
a) From the target image set, select the positive samples: $(\boldsymbol{p}_i^+, \boldsymbol{t}_i^+)$.
b) From the residual irrelevant sets, select the negative samples: $(\boldsymbol{p}_i^-, \boldsymbol{t}_i^-)$.
c) Combine the query sketch $\boldsymbol{s}_i$, positive sample $(\boldsymbol{p}_i^+, \boldsymbol{t}_i^+)$, and negative sample $(\boldsymbol{p}_i^-, \boldsymbol{t}_i^-)$ into the
quintuple $\{\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-\}_i^N$ as the supervised information.
2. Training steps:
for e = 1 : E **do**
for i = 1 : N **do**
/* Mapping ***/**
a) Feed the quintuple $\left\{\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-\right\}_i^N$ into DMEM:
$F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s) \leftarrow \boldsymbol{s}_i$ /*sketch feature mapping*/
$F(\boldsymbol{p}_i^+\|\boldsymbol{\theta}_i), F(\boldsymbol{p}_i^-\|\theta_i) \leftarrow \boldsymbol{p}_i^+, \boldsymbol{p}_i^-$ /*positive/negative image feature mapping*/
$F(\boldsymbol{t}_i^+\|\boldsymbol{\theta}_t), F(\boldsymbol{t}_i^-\|\boldsymbol{\theta}_t) \leftarrow \boldsymbol{t}_i^+, \boldsymbol{t}_i^-$ /*positive/negative text feature mapping*/
b) $DF(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+) \leftarrow F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s), F(\boldsymbol{p}_i^+\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i^+\|\boldsymbol{\theta}_t)$
$DF(\boldsymbol{s}_i, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-) \leftarrow F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s), F(\boldsymbol{p}_i^-\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i^-\|\boldsymbol{\theta}_t)$ /*calculate $DF(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{t})$ */
c) Calculate the ranking loss according to Eq. (15), and update the model parameters with the
stochastic gradient descending.
end
end
Output DMEM function $\boldsymbol{F}(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-; \theta)$.

3.3.2 Ranking Optimization via Deep Triplet Ranking

After the first stage, the candidate top-*K* ranked images that are similar to the query sketch at both visual and semantic levels can be preserved, and most irrelevant images are filtered away. In the next stage, the goal is to optimize the ranking effectiveness of such top-*K* images. Thus we construct another Deep Triplet Ranking Model (DTRM) to map *K* images into a common space, in which the re-ranking can be further performed.

DTRM is composed of three subnets following the similar subnet layer configuration as DMEM. Three subnets correspond to the triplet input $\{(\boldsymbol{s}_i, \boldsymbol{p}^+_i, \boldsymbol{p}^-_i)\}^K_i$, where $\boldsymbol{s}_i$ represents the query sketch, and $\boldsymbol{p}^+_i$ or $\boldsymbol{p}^-_i$ denotes the positive or negative image-description sample. It is worth noting that the triplet form is different from the quintuple in the first stage. This is because it is based on the initial retrieval results of the first stage, and here we focus on the discriminative fine-grained characteristics that are usually ignored in the first stage. As a result, how to sample these training triplets becomes very important. Thus a specific image-retrieval-based sampling strategy is explored to solve this problem. For each sample in the training set with three modalities of sketch, image, and description, we extract the learned features of three modalities from DMEM and concatenate three feature vectors. The Normalized Cosine (NC) similarity [48] is utilized to conduct the retrieval on the training set. For FG-SBIR, the top-10 returned images except the query image can be sampled as the negative samples. Hence, for each sketch $\boldsymbol{s}_i$, its corresponding images are selected as the positive samples $\boldsymbol{p}^+$ and the sampled images as the negative ones $\boldsymbol{p}^-$.

In the ranking optimization stage, we simply concatenate the deep visual and semantic feature as the multimodal embedded feature for each image. Given a fine-grained triplet training set $\phi=\{(\boldsymbol{s}_i, \boldsymbol{p}^+_i, \boldsymbol{p}^-_i)\}^K_i$, the deep visual feature for the query sketch and the multimodal embedded features for annotated images are fed into DTRM. It aims to learn the embedding space that the similarity of the positive pair is larger than that of the negative pair with a large margin. Let $(F(\boldsymbol{s}_i), F(\boldsymbol{p}^+_i), F(\boldsymbol{p}^-_i))$ be the learned feature vector of $(\boldsymbol{s}_i, \boldsymbol{p}^+_i, \boldsymbol{p}^-_i)$ in the DMEM's embedding space, the triplet ranking loss for the ranking optimization can be formulated as:

$$min \sum_{i\epsilon\phi} \max(0, \boldsymbol{m} + ||\boldsymbol{F}(\boldsymbol{s}_i) - F(\boldsymbol{p}_i^+)||^2 - ||\boldsymbol{F}(\boldsymbol{s}_i) - F(\boldsymbol{p}_i^+)||^2) \quad (16)$$

where $\boldsymbol{m}$ is a parameter for the margin. In fact, our optimization function can capture the fine-grained distance-based pairwise similarity, which fuses both visual and semantic similarity through mining the relative similarities of the training data. In the real SBIR/FG-SBIR environment with high efficiency requirements, the final ranking list can also be obtained by sorting the Euclidean distances between the query sketch and the top-*K* samples from the first stage in the embedded space of DTRM. It is both fast and accurate, which can be easily scaled to large-scale annotated images.

The process of Deep Triplet Ranking Model is summarized in Algorithm 2.

Algorithm 2: Deep Triplet Ranking Model (DTRM)

Input: Training dataset: $\boldsymbol{Z} = \{\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{t}\}_{i=1}^{n}$, DMEM function $\boldsymbol{F}\left(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{t}_i^+, \boldsymbol{p}_i^-, \boldsymbol{t}_i^-; \theta\right)$, Number of epochs: E.
Output: Candidate image ranking list.
1. Input selection:
a) Extract the learned features by the DMEM function $\boldsymbol{F}(*;\ \theta)$:
$F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s), F(\boldsymbol{p}_i\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i\|\boldsymbol{\theta}_t) \ \leftarrow \boldsymbol{s}_i, \boldsymbol{p}_i, \boldsymbol{t}_i$
b) Concatenate three feature vectors in series:
$[F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s), F(\boldsymbol{p}_i\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i\|\boldsymbol{\theta}_t)] \leftarrow F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s), F(\boldsymbol{p}_i\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i\|\boldsymbol{\theta}_t)$
c) Compute the Normalized Cosine (NC) similarity, conduct the retrieval on the training set, and
return the top-*K* images (sampled images).
d) For a query sketch $\boldsymbol{s}_i$, select the corresponding images as the positive samples $\boldsymbol{p}_i^+$, and select the
sampled images as the negative ones $\boldsymbol{p}_i^-$.
e) Concatenate the deep visual and semantic features together as the multimodal embedded feature for
each image:
$[F(\boldsymbol{p}_i\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i\|\boldsymbol{\theta}_t)] \leftarrow F(\boldsymbol{p}_i\|\boldsymbol{\theta}_i), F(\boldsymbol{t}_i\|\boldsymbol{\theta}_t)$
f) Use query sketches, positive samples, and negative samples to form the triplet training set $\phi =$
$\left(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{p}_i^-\right)_i^K$.
2. Training step:
for e = $1 : E$ **do**
for i = $1 : K$ **do**
/* Mapping */

a) Feed the triplet $\left(\boldsymbol{s}_i, \boldsymbol{p}_i^+, \boldsymbol{p}_i^-\right)_i^K$ into DMEM:
$F(\boldsymbol{s}_i\|\boldsymbol{\theta}_s) \leftarrow \boldsymbol{s}_i$ /*sketch feature mapping*/
$F(\boldsymbol{p}_i^+\|\boldsymbol{\theta}_i), F(\boldsymbol{p}_i^-\|\boldsymbol{\theta}_i) \leftarrow \boldsymbol{p}_i^+, \boldsymbol{p}_i^-$ /*positive/negative multimodal embedded features mapping*/
b) Calculate the ranking loss according to Eq. (16), and update the model parameters with the
stochastic gradient descending.
end
end
3. Ranking:
a) Compute the Euclidean distances between the sketch query and top-*K* samples in testing set.
b) Sort the distances to obtain the final ranking list.

4. Experiment and Analysis

4.1 Datasets and Evaluation Metrics

Sketchy, the large-scale benchmark dataset for FG-SBIR, is utilized as the main dataset in our experiment, which contains 12,500 photos and 75,471 sketches of 125 object categories. Each category contains 100 images, and each image has at least 5 well-drawn sketches along with descriptions. 1,250 images and their sketches are selected for testing, and the rest for training. We also extend the *Sketchy* dataset. More specifically, we select the same categories of images from the *Flicrk30K* [49] and *MSCOCO* [50] datasets to form two new extended datasets, which are called the *Flickr30K-Sketchy* intersection dataset and the *MSCOCO-Sketchy* intersection dataset. The purpose of supplementing these two auxiliary datasets is to help validate the generalization ability of our FG-SBIR model. Besides exhibiting the superiority of our proposed framework for fine-grained SBIR, we also want to verify its adaptability and extendibility for general coarse-grained SBIR.

Generally, each image in *Flickr30k* and *MSCOCO* has five ground truth captions. We employ an image selection strategy to select the images that have the same category labels as those of the sketches in *Sketchy*. First, a set of tokens is built from

image captions, and if this token set has an intersection with the category set for an image, the image is put into the candidate set. Second, for each image in the candidate set, if one category appears in the image caption, the category is assigned to the image. Third, both manual and automatic methods with the help of the context information in image descriptions are used to remove the unqualified images. However, the polysemy problem of category label exists. For example, the word "*bat*" can refer to not only a kind of animal but also a tool used in the baseball game, while we only need those images that contain the animal. Therefore, we manually remove the images where the "*baseball bat*" appears. In the other case, the word like "*apple*" has different kinds of characteristics. When "*apple*" occurs with "*computer*", the associated images would show electronic products instead of fruits. We then use the context of an image description to filter the category result. If "*apple*" and "*computer*" co-occur in an image description, we remove the image from the category of "*apple*" (fruit). Fourth, some selected categories contain too few images which hardly help our experiment, and then these categories are discarded through the category filtering. Thus two extended category-level SBIR datasets of *Flickr30K-Sketchy* and *MSCOCO-Sketchy* can be obtained from *Flickr30K* and *MSCOCO* respectively, as summarized in Tables 1 and 2. *Flickr30K-Sketchy* contains a total of 8 categories, and *MSCOCO-Sketchy* contains a total of 32 categories. It can be seen that the sample numbers for different categories are obviously unbalanced, especially in *MSCOCO-Sketchy*, which makes *MSCOCO-Sketchy* more challenging for SBIR.

For FG-SBIR, we use *Recall@K* (*K*=1, 5, 10) as the evaluation metric because FG-SBIR mostly focuses on the top ranking position of a target image. It can be regarded as the percentage of sketches whose true-match images are ranked in the top-*K* positions. This corresponds to an application scenario where the goal is simply to find a relevant image as quickly as possible. For SBIR, we use both *Recall@K* and *Precision@K* (*K*=1, 5, 10) as the evaluation metrics because our model is a ranking model. Different from FG-SBIR, *Recall@K* and *Precision@K* in SBIR have different definitions of whether a retrieved image is a true-match. For a query sketch, any image among the top-*K* retrieved images whose category is the same with the query sketch is a matched result. *Recall@K* is the percentage of relevant images that have

been retrieved over all the relevant images in the top-*K* positions. *Precision@K* is the percentage of relevant images among the retrieved images in the top-*K* positions.

Table 1. The statistics results for the *Flickr30K-Sketchy* intersection dataset.

Intersection Category	Number of Images	Number of Sketches
bench	612	583
bicycle	790	614
chair	498	669
dog	1,930	692
guitar	535	528
hat	1,844	529
table	984	563
tree	450	533
In total	7,643	4,127

Table 2. The statistics results for the *MSCOCO-Sketchy* intersection dataset.

Intersection Category	Number of Images	Number of Sketches
airplane	2,441	709
apple	415	551
banana	880	635
bear	907	722
bench	2,633	583
bicycle	1,180	614
bread	995	563
car (sedan)	2,002	642
cat	3,841	692
chair	718	669
couch	544	652
cow	1,017	728
cup	933	697
dog	3,020	692
door	1,231	578
elephant	1,630	661
flower	713	518
giraffe	2,318	673
hat	1,299	529
horse	1,988	738
hotdog	726	634
knife	727	624
motorcycle	2,052	643
pizza	2,461	606
racket	2,332	549
scissors	464	558
sheep	1,257	669
table	7,673	563
tree	1,814	533
umbrella	1,515	546
window	1,597	554
zebra	1,288	608
In total	56,999	16,908

4.2 Implementation Settings

We utilize *Caffe* to implement our DCCRM. The number of layers in each subnet is set to 3. For DMEM, the number of units per layer is set to {1,024, 2,048, 512} for the sketch branch, {1,024, 2,048, 512} for the image branch, and {4,800, 2,048, 512} for the description branch. For DTRM, the number of units per layer is set to {1,024, 2,048, 512} for the sketch branch and {5,824, 2,048, 512} for the annotated image

branch. Each layer is initialized by pre-training a denoising autoencoder as in [16]. The top-50 candidate images are selected for ranking optimization. The margin *m* of the objective function for both DMEM and DTRM is set to 100.

4.3 Ablation Study

Our model is created by integrating DCCRM with deep multimodal features. To investigate the contribution of each component in our framework, we introduce two evaluation designs: 1) Utilizing different combinations of multimodal features for DCCRM, i.e., *Sketch+Image* -- *DCCRM*(*S+I*) and *Sketch+Image+Description* -- *DCCRM*(*S+I+D*), to evaluate the necessity for exploiting the semantic modality of annotated images in FG-SBIR, in which for *DCCRM*(*S+I*) we use the same DTRM for the ranking optimization but modify the number of units for the input layer; and 2) Comparing different components of DCCRM, i.e., *DMEM*(*S+I+D*), *DTRM*(*S+I*), and *DCCRM*(*S+I+D*), to prove the validity of our deep cascaded model, in which for *DTRM*(*S+I*) we ignore the candidate selection stage and use the same triplet sampling on deep visual features. We also introduce a simple baseline by conducting the retrieval on 1,024-way deep visual features for sketches and the generated sketch-like images. The experimental results on *Sketchy* are shown in Table 3.

Table 3. The contributions of different components for FG-SBIR on *Sketchy*.

Evaluation Pattern	*Recall*@1	*Recall*@5	*Recall*@10
Baseline	5.19%	11.96%	16.30%
***DMEM*(*S+I+D*)**	35.60%	78.88%	89.60%
***DTRM*(*S+I*)**	27.60%	62.88%	78.48%
***DCCRM*(*S+I*)**	40.16%	81.84%	92.00%
***DCCRM*(*S+I+D*)**	46.20%	84.64%	96.49%

It can be seen from Table 3 that the best performance is obtained for *DCCRM*(*S+I+D*), which fuses all the useful modality information with *DCCRM*. This confirms the obvious advantage of our framework for FG-SBIR. For the recall percentage at top-1, top-5, and top-10, these indicators of *DCCRM*(*S+I+T*) outperform those of *DCCRM*(*S+I*) by 6.04%, 2.80%, and 4.49% respectively, which proves the necessity of introducing the semantic modality (i.e., textual description) of annotated images. It can be concluded that the FG-SBIR performance is further enhanced

through mining all the beneficial multimodal information in annotated images, especially semantic description information, rather than only considering the unimodal visual information in images. Comparing our cascaded models of *DCCRM*(*S*+*I*) and *DCCRM*(*S*+*I*+*D*) with the single ranking models of *DMEM*(*S*+*I*+*D*) and *DTRM*(*S*+*I*), both cascaded models significantly outperform *DMEM*(*S*+*I*+*D*) and *DTRM*(*S*+*I*). It can be seen that even the ranking optimization in the second stage processes a substantially smaller set from the first stage, it further greatly improves the ranking effectiveness for retrieval, which shows the validity of our cascaded model again.

In addition to the experimental validation on *Sketchy*, we also consider transferring the same architecture to the general coarse-grained SBIR. We perform extensive generalization ability experiments on *Flickr30K-Sketchy* and *MSCOCO-Sketchy* with the same evaluation designs and implementation settings. The experimental results on these two extended datasets are shown in Tables 4 and 5, respectively.

Table 4. The contributions of different components for SBIR on *Flickr30k-Sketchy*.

Evaluation Pattern	*Recall*@1		*Recall*@5		*Recall*@10		*Precision*@1	*Precision*@5	*Precision*@10
	Raw	Normalized	Raw	Normalized	Raw	Normalized			
Baseline	1.01%	73.27%	4.57%	66.02%	8.47%	61.19%	78.69%	71.79%	67.57%
***DMEM*(*S*+*I*+*D*)**	1.29%	92.92%	5.59%	80.83%	10.84%	78.33%	94.79%	86.98%	85.81%
***DTRM*(*S*+*I*)**	1.19%	86.20%	5.18%	74.85%	9.78%	70.65%	89.57%	81.41%	77.28%
***DCCRM*(*S*+*I*)**	1.21%	87.20%	5.20%	75.11%	9.78%	70.68%	89.57%	81.90%	78.01%
***DCCRM*(*S*+*I*+*D*)**	1.29%	93.01%	5.62%	81.21%	11.09%	80.11%	94.90%	87.53%	86.77%

Table 5. The contributions of different components for SBIR on *MSCOCO-Sketchy*.

Evaluation Pattern	*Recall*@1		*Recall*@5		*Recall*@10		*Precision*@1	*Precision*@5	*Precision*@10
	Raw	Normalized	Raw	Normalized	Raw	Normalized			
Baseline	0.51%	59.14%	2.37%	54.87%	4.26%	49.28%	78.69%	71.79%	67.57%
***DMEM*(*S*+*I*+*D*)**	0.58%	67.25%	2.78%	64.29%	5.30%	61.34%	94.79%	86.98%	85.81%
***DTRM*(*S*+*I*)**	0.51%	59.03%	2.36%	54.52%	4.61%	53.33%	89.57%	81.41%	77.28%
***DCCRM*(*S*+*I*)**	0.51%	59.03%	2.37%	54.92%	4.71%	54.52%	89.57%	81.90%	78.01%
***DCCRM*(*S*+*I*+*D*)**	0.58%	67.30%	2.82%	65.21%	5.44%	62.97%	94.90%	87.53%	86.77%

It is worth noting that different from *Sketchy*, sketches and images in *Flickr*30*k-Sketchy* and *MSCOCO-Sketchy* are not in pairs. We use the recall percentage at top-1, top-5, and top-10 to evaluate the retrieval performance for SBIR. Because the number of the corresponding images for each category exceeds 10, it will lead to the theoretical maximum values of *Recall*@1, *Recall*@5, and *Recall*@10 not 1. For example, for a query sketch, if the number of relevant images in the testing image set is 100, the theoretical maximum values of *Recall*@1, *Recall*@5, and *Recall*@10 will be 1.00%, 5.00%, and 10.00% respectively. It can be observed from Tables 4 and 5 that the raw *Recall@K* (*K*=1, 5, 10) values are too small and the performance differences among different models are not obvious. Thus for these two extended datasets, we make a special normalization for the *Recall@K* values. We first calculate the theoretical maximum value for *Recall@K*, and then use the quotient of the raw value divided by the theoretical maximum value as the normalized value. The normalized *Recall@K* values can certainly display the performance differences among different models, but change the commonsense variation tendency of the recall values. More specifically, for each model, the normalized *Recall@K* values decrease with the *K* value increasing, while by common sense the *Recall@K* values should increase with the *K* value increasing. However, from the raw *Recall@K* values without any normalization for each separate model, we can still view the reasonable variation tendency of the recall values with different settings of *K*.

From Tables 4 and 5, we can see that the best performance is obtained for *DCCRM*(*S*+*I*+*D*), which proves that our framework also has a strong advantage for general coarse-grained SBIR. The results on *Flickr30k-Sketchy* are better than those on *MSCOCO-Sketchy*, because *MSCOCO-Sketchy* is larger and more difficult, and contains more unbalanced categories. In Table 5, the recall percentage at top-1 for *DCCRM*(*S*+*I*) is even slightly worse than that for *Baseline.* For the recall and precision percentage at top-1, top-5, and top-10, *DCCRM*(*S*+*I*+*D*) has a better performance than *DCCRM*(*S*+*I*), which means that introducing the semantic modality of annotated images can bring a significant performance enhancement. Mining all the beneficial multimodal information in annotated images is not only beneficial for FG-SBIR but also for SBIR. Furthermore, comparing the cascaded models of *DCCRM*(*S*+*I*) and *DCCRM*(*S*+*I*+*D*) with the single ranking models of *DMEM*(*S*+*I*+*D*)

and *DTRM*(*S*+*I*), it can be seen that the ranking optimization in the second stage processes a substantially smaller set from the first stage. This can yield a certain improvement on the retrieval performance, because in the first ranking selection stage our ranking results only focus on top-50 candidate images, most of which are in the same category with the query sketch.

4.4 Comparison with Other Approaches

To further verify the superiority of our model, we perform extensive comparisons with various baseline models. We first give a short overview of these approaches, and then show the comparisons via the quantitative results in Table 6.

- **Chance --** Given a query sketch, the target images are retrieved in a random manner. Specifically, two patterns are compared, that is, *Chance* (Retrieval in random order on the whole dataset) and *Chance w/ Label* (Retrieval in random order within the ground-truth category).
- **GALIF** [31] **--** *Gabor Local Line based Feature*, which builds on a blank of *Gabor* filter followed by a bag-of-words method and has been successfully used for SBIR.
- **RankSVM-based** [13] **--** It uses two types of features, i.e., *HOG-BoW* and *Dense-HOG*, in which *HOG-BoW* utilizes a *BoW* descriptor (500D) generated from the *HOG* features and *Dense-HOG* (200,704D) concatenates the *HOG* features over a dense grid. Based on these features, *RankSVM* is utilized for retrieval.
- **Classification Network --** "*Retrieval by categorization*", which includes *GN Cat* (*GoogLeNet* trained with the classification loss on *Sketchy*), *SN* (Retrieval based on deep features from the fine-tuned *GoogLeNet* on *TU-Berlin* sketches), and *SN w/ Label* (same as *SN* but Retrieval within the ground-truth category).
- **Sketch-a-Net** [13] **--** This network consists of three CNN branches (sketches, positive images, and negative images) with shared weights, in which each branch is based on *Sketch-a-Net*.

- **Siamese Network** [21] **--** *GN Siamese/AN Siamese*, which consists of two asymmetric sketch and image branches. Both are initialized with *GoogLeNet*/ *AlexNet*, and trained with the classification loss on *Sketchy*.
- **Triplet Network** [21] **--** This network consists of three branches of CNNs for sketch, positive image, and negative image. Two comparison patterns are introduced, that is, *GN Triplet* (*GoogLeNet* trained with the triplet and classification loss on *Sketchy*), and *GN Triplet w/o Cat* (*GoogLeNet* trained with the triplet loss on *Sketchy*).
- **Quadruplet Network** [51] **--** It is similar to Triplet Network, but uses the *ResNet*-18 architecture with the shared weights for both sketch and image branches. The training involves two steps: (i) training with the classification loss on *Sketchy*; and (ii) training a network with the triplet loss on *Sketchy*, while mining three different types of triplets.
- **DeepTCNet** [33] **--** *Deep Triplet Classification Siamese Network*, which employs *DenseNet*-169 as the basic feature extractor and does not share weights for the branches of sketches and images. It is optimized by utilizing both the triplet loss and classification loss.

Table 6. The comparison results between our model and baseline approaches.

Category	Methods	*Recall@1*
Chance	*Chance*	0.01%
	Chance w/ Label	10.00%
GALIF [31]	*GALIF*	3.92%
RankSVM-based [13]	*BoW-HOG+RankSVM*	15.22%
	Dense-HOG+RankSVM	18.58%
Classification Network	*GN Cat*	12.63%
	SN	5.19%
	SN w/ Label	20.09%
Sketch-a-Net [13]	*Sketch-a-Net*	25.87%
Siamese Network [21]	*GN Siamese*	27.36%
	AN Siamese	21.36%
Triplet Network [21]	*GN Triplet*	37.10%
	GN Triplet w/o Cat	22.78%
Quadruplet Network [51]	*Quadruplet_MT*	38.21%
	Quadruplet_MT_V2	42.16%
DeepTCNet [33]	*DeepTCNet*	40.81%
Ours with DCCRM	***DCCRM(S+I)***	**40.16%**
	DCCRM(S+I+D)	**46.20%**
--	**Human** [21]	**54.27%**

It can be observed from Table 6 that the best performance is achieved by our DCCRM under the consideration of all the available modality information, i.e., *DCCRM*(*S*+*I*+*D*). The *Recall*@1 value can reach 46.20%, which apparently outperforms those of all the other existing approaches. This demonstrates that our cascaded ranking scheme with deep multimodal feature representation can exactly play an important role in FG-SBIR. Compared to the latest *DeepTCNet* which uses *DenseNet*-169 as the feature extractor, our approach can also get very competitive results. Compared to the *Human* performance, the relatively small gap between the *Human* performance (54.27%) and ours (46.20%) still exhibits the promising potentials for developing more powerful FG-SBIR with DCCRM. Furthermore, we can find that those very deep end-to-end models, e.g., *GN Triplet*, *GN Siamese*, and *Sketch-a-Net*, cannot always obtain a better performance. We also observe that our DCCRM model acquires much higher *Recall*@1 value than those of *GN Triplet*, *GN Siamese*, and *Sketch-a-Net*. Around 9%, 19%, and 20% improvements for *Recall*@1 can be obtained over *GN Triplet*, *GN Siamese*, and *Sketch-a-Net* respectively, which is mainly due to the lack of effective training and overfitting for such models. The deep end-to-end models require both large-scale training data with great diversity and adequate training. In fact, it is usually hard to train such deep models to achieve the desired convergence of an objective function. However, DCCRM leverages two deep cascaded embedding and ranking models to mine all the beneficial multimodal information in sketches and annotated images. Thus a more appropriate ranking scheme can be well learned in a coarse-to-fine way, which is easy to train and can be applied to small-scale datasets. Meanwhile, the deep-learning-based approaches obviously outperform the basic methods with hand-crafted features like *GALIF*, which mainly attributes to the effort of supervised training. In addition, different from *DCCRM*(*S*+*I*+*D*), *DCCRM*(*S*+*I*) only exploits visual features in sketches and images without using any textual description information, which can be regarded as a part of *DCCRM*(*S*+*I*+*D*). Even compared with other existing models, the performance of *DCCRM*(*S*+*I*) is still better than most of other existing models except for one of them (i.e., *Quadruplet_MT_V*2), which demonstrates the effectiveness of combining both the visual and textual cues in FG_SBIR again.

We also plot the *Recall@K* (*K*=1 to 10) to give the exhibitions of comparisons at different ranking positions with some open source baseline models, as shown in Fig. 4. Similar conclusions can be drawn as above. It is interesting to note that a considerable part of the existing approaches listed in Fig. 4 can achieve the high recall over 90% within the top-10 retrieval results, such as *GN Triplet*, *GN Triplet w/o Cat*, and *GN Siamese*. However, their recall values for high-ranked positions are not satisfactory, which indicates the necessity of the appropriate ranking optimization for the top-*K* results. In comparison with such existing approaches, DCCRM can improve both the retrieval efficiency and the top-*K* ranked effectiveness, and demonstrates its superiority and suitability for FG-SBIR.

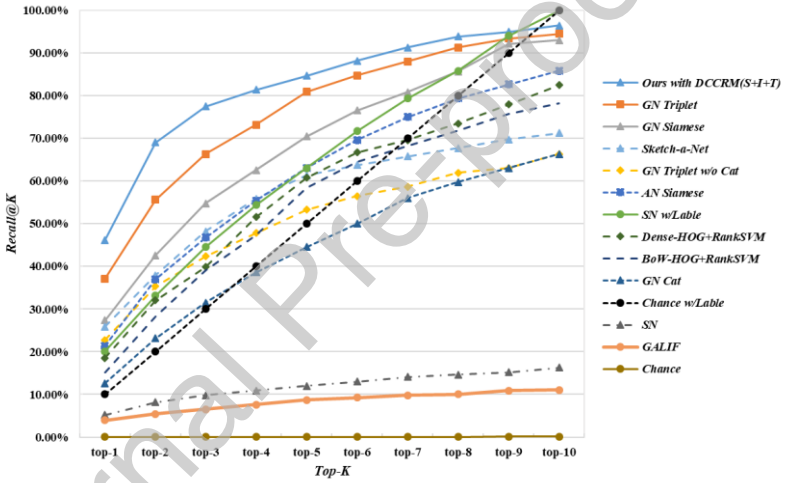
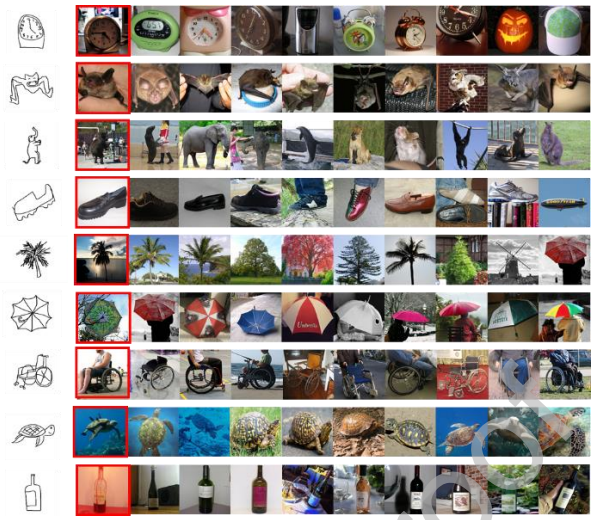
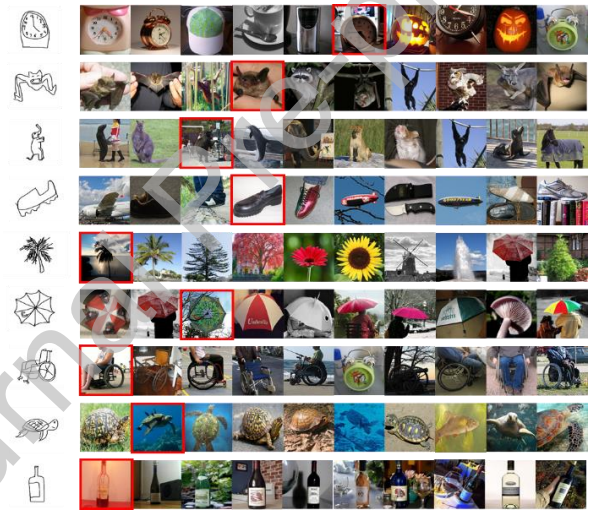


Fig. 4. The comparison of *Recall@K* within the top-*K* positions (*K*=1 to 10) between our model and open source baseline approaches.

An illustration of some sketch queries and their top-10 retrieval results is shown in Fig. 5, in which the true-match images are highlighted in the red boxes. We can see that the returned top ranking images under our scheme with *DCCRM*(*S*+*I*+*T*) correspond more closely to the query sketches than those under *DCCRM*(*S*+*I*). This indicates that our approach can effectively return both visually similar and semantically relevant images, and rule out the irrelevant ones.



(a) Under our scheme with *DCCRM*(*S*+*I*+*T*);



(b) Under our scheme with *DCCRM*(*S*+*I*).

Fig. 5. An illustration of some sketch queries and their top-10 relevant images.

Since the response time is an important factor for retrieval, the average response time of our approach is 600ms per query sketch with GPU acceleration and Nearest Neighbor Search (NNS), which exhibits its better feasibility and practicality.

4.5 Analysis and Discussion

Through the analysis for failure or error instances in the retrieval results, we observe that the fine-grained SBIR quality is highly related to the following aspects. (i) The ambiguity of sketches limited by drawing skills of users is still a stubborn problem. It may be helpful to explore a multimodal query pattern, i.e., sketch and description, to locate target object images more precisely. (ii) For real images with abundant backgrounds or multiple objects, it is novel to introduce a preprocessing step, i.e., object detection and segmentation, to reduce the negative influence of irrelevant backgrounds or objects. (iii) The underlying relevance patterns in the ranking list returned by the same query sketch should be deeply analyzed. Such information may be useful for query extension and more precise re-ranking, and then the whole retrieval performance can be further improved.

In addition, besides the *Sketchy*, *Flickr30k-Sketchy*, and *MSCOCO-Sketchy* datasets, we also want to utilize two well-known datasets for FG-SBIR in our experiments, i.e., *QMUL-Shoe* and *QMUL-Chair* [13]. *QMUL-Shoe* and *QMUL-Chair* contain 419 shoe sketch-photo pairs and 297 chair sketch-photo pairs respectively, in which the photos are real product photos collected from online shopping websites and the sketches are free-hand ones collected via crowdsourcing [32]. However, because the color images in these two datasets do not include textual descriptions for images and lack of text information, the two datasets cannot adapt well with our complete model. Hence, we consider to make limited experiments with the partial model of *DCCRM*(*S+I*) on *QMUL-Shoe* and *QMUL-Chair*. The *Recall@*1 values for *DCCRM*(*S+I*) are 29.57% and 63.92% respectively, which are lower than those of 39.13% and 69.07% for the origin model in [13]. It is noteworthy that our model uses *GoogleNet* as the backbone while the model in [13] uses *Sketch-A-Net* as the backbone. *GoogleNet* is more complex and has a stronger fit ability. Following the same dataset split for the origin model, our model uses 304 shoe sketch-photo pairs and 200 chair sketch-photo pairs for training respectively. Obviously, due to the relatively restricted scale of such two datasets, our model is easy to fall into overfitting. Thus, taking the above factors into account, *QMUL-Shoe* and *QMUL-Chair* are not exploited in our experiments with the

DCCRM model that aims to explore all the beneficial multimodal information in sketches and annotated images for FG-SBIR.

5. Conclusions and Future Work

In this work, a new framework is introduced to support more precise FG-SBIR for large-scale annotated images. A novel deep cascaded cross-modal ranking model is created to fuse deep multimodal features and exploit the cross-modal correlations among the attributes of different modalities in sketches and annotated images. The cascaded ranking scheme is established to refine the pairwise correlations of sketch-image pairs in a coarse-to-fine way. Our cascading ranking model exhibits the rationality of the cascading optimization model, which has not been paid more attention to in previous studies. It can not only be applicable to FG-SBIR, but also provide a referential framework for other retrieval tasks such as cross-modal image-text retrieval. Our future work can be summarized in two aspects. One is that since currently our approach is a multi-stage ranking model, we will focus on an end-to-end implementation with the integration of the cascaded neural network and the region-based matching scheme for FG-SBIR, and a further generalization of the same framework with text information in the general SBIR. The other is that our approach adopts the textual description information for images, which increases the burden of data collecting in practice to a certain degree. Thus we will put more emphases on establishing a more adaptable architecture that can be appropriate for different image collections with different granularities of textual descriptions, so as to achieve better retrieval effects in different application scenarios.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61976057, No. 61572140), Shanghai Municipal R&D Foundation (No. 17DZ1100504, No. 16JC1420401), Shanghai Natural Science Foundation (No. 19ZR1417200), and Humanities and Social Sciences Planning Fund of Ministry of Education of China (No. 19YJA630116). Weiguo Fan is supported by the Henry

Tippie Endowed Chair Fund from the University of Iowa. Yuejie Zhang and Tao Zhang are corresponding authors.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Chatbri, K. Kameyama, "Using scale space filtering to make thinning algorithms robust against noise in sketch images," in Pattern Recognition Lett., vol. 42, no. 1, 2014, pp. 1-10.

[2] C. Peng, X. Gao, N. Wang, J. Li, "Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation," in Pattern Recognition, vol. 84, 2018, pp. 262-272.

[3] S. Liang, Z. Sun, "Sketch retrieval and relevance feedback with biased SVM classification," in Pattern Recognition Lett., vol. 29, no. 12, 2008, pp. 1733-1741.

[4] C. Li, Y. Huang, L. Zhu, "Color texture image retrieval based on Gaussian copula models of Gabor wavelets," in Pattern Recognition, vol. 64, 2017, pp. 118-129.

[5] X. Qian, X. Tan, Y. Zhang, R. Hong, M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," in IEEE Trans. Image Processing, vol. 25, no. 1, 2016, pp. 195-208.

[6] C.H. Liu, Y.L. Lin, W.F. Cheng, W.H. Hsu, "Exploiting word and visual word co-occurrence for sketch-based clipart image retrieval," in Proc. ACM MM, 2015, pp.867-870.

[7] Z.H. Wang, S.J. Wang, P.B. Zhang, H.J. Li, and B. Liu, "Accurate and fast fine-grained image classification via discriminative learning," in Proc. ICME, 2019, pp. 634-639.

[8] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, W. Fan, "Sketch-based image retrieval with deep visual semantic descriptor," in Pattern Recognition, vol. 76, 2018, pp. 537-548.

[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in Proc. CVPR, 2014, pp. 1386-1393.

[10] S. Wang, J. Zhang, T.X. Han, and Z.J. Miao, "Sketch-based image retrieval through hypothesis-driven object boundary selection with HLR descriptor," in IEEE Trans. Multimedia, vol. 17, no. 7, 2015, pp. 1045-1057.

[11] X. Zhang and X.J. Chen, "Robust sketch-Based image retrieval by saliency detection," in Proc. MMM, 2016, pp. 515-526.

[12] Y. Li, T.M. Hospedales, Y.Z. Song, and S. Gong, "Fine-grained sketch-based image retrieval by matching deformable part models," in Proc. BMVC, 2014, pp. 1-12.

[13] Q. Yu, F. Liu, Y.Z. Song, T. Xiang, T.M. Hospedales, and C.C. Loy, "Sketch me that shoe," in Proc. CVPR, 2016, pp. 799-807.

[14] P. Xu, Q.Y. Yin, Y.G. Qi, Y.Z. Song, Z.Y. Ma, and L. Wang, "Instance-level coupled subspace learning for fine-grained sketch-based image retrieval," in Proc. ECCV, 2016, pp. 19-34.

[15] P. Xu, Q.Y. Yin, Y.Y. Huang, Y.Z. Song, Z.Y. Ma, L. Wang, T. Xiang, W.B. Kleijn, and J. Guo, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," in Neurocomputing, vol. 278, 2018, pp.75-86.

[16] K. Li, K. Pang, Y. Song, T. Hospedales, H. Zhang, and Y. Hu, "Fine-grained sketch-based image retrieval: the role of part-aware attributes," in Proc. WACV, 2016, pp.1-9.

[17] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in Proc. ICML, 2013, pp.1247-1255.

[18] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in Proc. CVPR, 2014, pp. 1386-1393.

[19] F. Huang, Y. Cheng, C. Jin, Y.J. Zhang, and T. Zhang, "Deep multimodal embedding model for fine-grained sketch-based image retrieval," in Proc. SIGIR, 2017, pp. 929-932.

[20] J. Song, Y.Z. Song, T. Xiang, T. Hospedales, and R. Xiang. "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval," in Proc. BMVC, 2017, pp. 132.1-132.11.

[21] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," in ACM Trans. Graph, vol. 35, no. 4, 2016, pp. 119-130.

[22] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in Proc. CVPR, 2010, pp. 3304-3311.

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," in Int'l J. Computer Vision, vol. 105, no. 3, 2013, pp. 222-245.

[24] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in Proc. ECCV, 2014, pp. 392-407.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. CVPR, 2015, pp. 1-9.

[26] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in Proc. ECCV, 2014, pp. 584-599.

[27] K. Lin, H.F Yang, J.H. Hsiao, C.S Chen, "Deep learning of binary hash codes for fast image retrieval," in Proc. CVPR, 2015, pp. 27-35.

[28] J.Y. Ng, F. Yang, L.S. Davis, "Exploiting local features from deep networks for image retrieval," in Proc. CVPR Workshops, 2015, pp. 53-61.

[29] C.H. Liu, Y.L. Lin, W.F. Cheng, W.H. Hsu, "Exploiting word and visual word co-occurrence for sketch-based clipart image retrieval," in Proc. MM, 2015, pp.867-870.

[30] Q. Yu, Y. Yang, Y. Song, T. Xiang, and T. Hospedales, "Sketch-a-net that beats humans," in Proc. BMVC, 2015, pp. 1-11.

[31] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?," in ACM Trans. Graph, vol. 31, no. 4, 2012, Article 44.

[32] J.F. Song, Q. Yu, Y.Z. Song, T. Xiang, and T.M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in Proc. ICCV, 2017, pp. 5552-5561.

[33] P. Lu, H.Y. Lin, Y.W. Fu, S.G. Gong, Y.G. Jiang, and X.Y. Xue, "Instance-level sketch-based retrieval by deep triplet classification Siamese network," in arXiv preprint arXiv: 1811.11375, 2018.

[34] G. Huang, K.Q. Weinberger, and L. Maaten, "Densely connected convolutional networks," in Proc. CVPR, 2017, pp. 2261-2269.

[35] X. Bai, M.K. Yang, P.Y, Lyu, and Y.C. Xu, "Integrating scene text and visual appearance for fine-grained image classification with convolutional neural networks," in arXiv preprint arXiv: 1704.04613, 2017.

[36] L. Huang and Y.X. Peng, "Cross-media retrieval by exploiting fine-grained correlation at entity level," in Neurocomputing, vol. 236, 2017, pp. 123-133.

[37] J.F. Song, Y.Z. Song, T. Xiang, T. Hospedales, X. Ruan, "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval," in Proc. BMVC, 2016, pp. 1-11.

[38] D.M. Blei and M.I. Jordan, "Modeling annotated data," in Proc. SIGIR, 2003, pp.127-134.

[39] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in Proc. NIPS, 2013, pp. 2121-2129.

[40] Y.T. Zhuang, Z. Yu, W. Wang, F. Wu, S.L. Tang, and J. Shao, "Cross-media hashing with neural networks," in Proc. ACM MM, 2014, pp. 901-904.

[41] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in Proc. ACM MM, 2014, pp. 7-16.

[42] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S.L. Tang, and Y.T. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment," in Proc. ACM MM, 2015, pp. 69-78.

[43] C. Szegedy, W. Liu, Y.Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions," in Proc. CVPR, 2015, pp. 1-9.

[44] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection," in Proc. CVPR, 2005, pp. 886-893.

[45] J. Canny, "A computational approach to edge detection," in IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6, 1986, pp. 679-698.

[46] R. Kiros, Y.K. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in Proc. NIPS, 2015, pp. 3294-3302.

[47] Y.K. Zhu, R. Kiros, R.S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in Proc. ICCV, 2015, pp. 19-27.

[48] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny. "Cosine similarity scoring without score normalization techniques," in Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop, 2010.

[49] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in TACL, vol. 2, 2014, pp. 67-78.

[50] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C.L. Zitnick, "Microsoft coco: Common objects in context," In Proc. ECCV, 2014, pp. 740-755.

[51] O. Seddati, S. Dupont, and S. Mahmoudi, “Quadruplet networks for sketch-based image retrieval,” in Proc. MM, 2017, pp. 184-191.

AuthorBiography

Yanfei Wang received the B.S. degree in Computer Science from Sun Yat-sen University, Guangzhou, China, in 2017. He is currently a master student in School of Computer Science, Fudan University, Shanghai, China. He is a member of Institution of Media Computing in School of Computer Science. His research interest is cross-media retrieval and image synthesis/translation, including sketch-based image retrieval, multi-view/multimodal correlation learning, and sketch synthesis.

Fei Huang received the B.S. degree in Computer Science from Hefei University of Technology, Hefei, China, in 2015, and the M.S. degree in Computer Science from Fudan University, Shanghai, China, in 2018. His research interest is cross-media retrieval, including sketch-based image retrieval and multi-view/multimodal correlation learning.

Yuejie Zhang received the B.S. degree in Computer Software, the M.S. degree in Computer Application, and the Ph.D. degree in Computer Software and Theory from Northeastern University, Shenyang, China, in 1994, 1997 and 1999, respectively. She was a Postdoctoral Researcher at Fudan University, Shanghai, China, from 1999 to 2001. In 2001, she joined Department of Computer Science and Engineering (now School of Computer Science), Fudan University as an Assistant Professor, and then become Associate Professor and Full Professor. Her research interests include multimedia/cross-media information analysis, processing, and retrieval, and machine learning.

Rui Feng received the B.S. degree in Industrial Automatic from Harbin Engineering University, Haerbin, China, in 1994, the M.S. degree in Industrial Automatic from Northeastern University, Shenyang, China, in 1997, and the Ph.D. degree in Control Theory and Engineering from Shanghai Jiaotong University, Shanghai, China, in 2003. In 2003, He joined Department of Computer Science and Engineering (now School of Computer Science), Fudan University as an Assistant Professor, and then become Associate Professor and Full Professor. His research interests include multimedia information analysis and processing, and machine learning.

Tao Zhang received the B.S. and M.S. degree in Automation Control, and the Ph.D. degree in System Engineering from Northeastern University, Shenyang, China, in 1992, 1997 and 2000, respectively. He was a Postdoctoral Researcher at Fudan University, Shanghai, China, from 2001 to 2003. In 2003, he joined School of Information Management and Engineering, Shanghai University of Finance and Economics as an Associate Professor and then become Full Professor. His research interests include big data analysis and mining, system modeling and optimization.

Weiguo Fan received the B.S. degree in information and control engineering from the Xi'an Jiaotong University, Xian, China, in 1995, the M.S. degree in computer science from the National University of Singapore in 1997, and the Ph.D. degree in AI and Information Systems from the University of Michigan, Ann Arbor, in 2002.

He is currently Henry Tippie Chaired professor of business analytics at the University of Iowa. He has published more than 200 refereed articles in many premier IT/IS journals and conferences such as TKDE, PR, TOIT, WWW, SIGIR, CIKM, AAAI, and KDD. His research interests include information retrieval, data mining, text mining, Web mining, and pattern recognition.