Deep Multimodal Embedding Model for Fine-grained Sketch-based Image Retrieval

Fei Huang
School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, China
15210240036@fudan.edu.cn

Yong Cheng
School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, China
13110240027@fudan.edu.cn

Cheng Jin
School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, China
jc@fudan.edu.cn

Yuejie Zhang
School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, China
yjzhang@fudan.edu.cn

Tao Zhang
School of Information Management and Engineering
Shanghai University of Finance and Economics
China
taozhang@mail.shufe.edu.cn

ABSTRACT

Fine-grained Sketch-based Image Retrieval (Fine-grained SBIR), which uses hand-drawn sketches to search the target object images, has been an emerging topic over the last few years. The difficulties of this task not only come from the ambiguous and abstract characteristics of sketches with less useful information, but also the cross-modal gap at both visual and semantic level. However, images on the web are always exhibited with multimodal contents. In this paper, we consider Fine-grained SBIR as a cross-modal retrieval problem and propose a deep multimodal embedding model that exploits all the beneficial multimodal information sources in sketches and images. In our experiment with large quantity of public data, we show that the proposed method outperforms the state-of-the-art methods for Fine-grained SBIR.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**

KEYWORDS

Fine-grained Sketch-based Image Retrieval (Fine-grained SBIR); Deep Multimodal Embedding; Multimodal Ranking Loss

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08…$15.00
http://dx.doi.org/10.1145/3077136.3080681

1 INTRODUCTION

To capture the important property of objects at the fine-grained level, such as pose, viewpoint and shape, Fine-grained Sketch-based Image Retrieval (Fine-grained SBIR) is proposed recently. It focuses on finding the correspondences between sketches and images at the instance level, since the insistent demand of better in-depth understanding for sketches. The difficulty of Fine-grained SBIR is the inherently visual and semantic gap between different modalities of sketches and images. An image is generally exhibited in a form with different modalities (i.e., visual and semantic), such as a web image with textual description. However, sketches are less informative, and their inherent abstractness and ambiguities cannot be well handled only by exploiting the visual information [1, 2].

The first work for Fine-grained SBIR was proposed by [3], which first learned the mid-level sketch representation, and then used the graph matching to discover the pose correspondence between sketches and images. However, these hand-crafted features cannot bridge the cross-domain gap in deep level. [4] formulated a cross-domain framework to learn the cross-domain feature space to conduct Fine-grained SBIR using fine-grained visual attributes and instance-level pair-wise annotations. [5] introduced a specific database of shoes and chairs and developed a deep triplet-ranking model for the instance-level SBIR. However, such methods are either based on visual contents without considering semantic attributes, or limited to small-scale datasets. Recently, [6] proposed a large-scale database, "*sketchy database*", as a benchmark dataset for Fine-grained SBIR and tested several popular cross-domain convolutional network architectures. In the real SBIR environment, the multimodal information of annotated images is usually ignored. Thus much closer attention has been given to the methods that rely on exploiting multimodal attributes.

To capture both the visual and semantic similarities between sketches and images at the fine-grained level, we introduce a deep multimodal embedding model for Fine-grained SBIR. Our model is

an end-to-end learning framework to optimize the retrieval performance by mining all the possible beneficial multimodal information in sketches and annotated images. The model first map three modalities of sketches, images and textual descriptions into a shared common space through deep neural network and then the correspondence correlations between query sketches and their images/descriptions are maximized in that space through a special combination of multimodal ranking loss and classification loss. This is a novel and meaningful way that incorporates the sketch-image visual comparisons and the sketch-description semantic comparisons to enhance the performance of Fine-grained SBIR. Our experiments on a large-scale benchmark dataset demonstrate the superiority of our proposed model over the other existing competitive methods for Fine-grained SBIR.

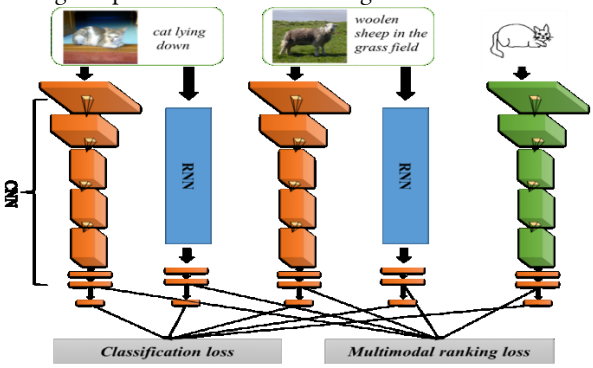

Figure 1: The basic framework of our proposed model.

2 DEEP MULTIMODAL EMBEDDING MODEL

Our architecture is a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The CNN maps sketches and images to their feature vectors and the RNN maps the textual descriptions to the feature vectors in a shared space. Then two objective functions are combined to constraint the correlations of sketch-image pairs in that space.

2.1 Convolutional Neural Network for Visual Embedding

We use the CNN for sketches and images to map them to their intermediate vectors. Comparing to traditional visual descriptors like HOG [7], such representations are closer to image semantics due to the supervised learning. The CNN takes a raw fixed-sized image as the input and produces a fixed-length vector after stacked layers of operations, such as convolution, nonlinear transformation and pooling. The feature map in the lower layer is computed and then transmitted to the higher layer to acquire the better representation. Since sketches and images belong to the visual domain but different modalities, such two CNN architectures for them are identical but they are trained without sharing parameters. Specifically, given a set of images $[p_1, p_2, p_3, \dots, p_N]$ and sketches $[s_1, s_2, s_3, \dots, s_N]$ and let $p_i \in R^{m \times n}$ and $s_i \in R^{m \times n}$ be an image and a sketch of size $m \times n$, the embedding function for these two CNNs can be formulated as:

$$[P_1, P_2, P_3, \dots, P_N] = CNN_I([p_1, p_2, p_3, \dots, p_N]|\theta_I)$$
$$[S_1, S_2, S_3, \dots, S_N] = CNN_S([s_1, s_2, s_3, \dots, s_N]|\theta_S) \quad (1)$$

where $P_i \in R^d$ and $S_i \in R^d$ are the learnt intermediate vectors for the given image and sketch; and θ_I and θ_S are the parameters for the image CNN_I and sketch CNN_S. Following the prior work of [6], we take *GoogLeNet* [8] as the embedding neural network for visual sketches/images, and the 1,024-way average pooling layer output after the last inception module is taken as the embedded intermediate feature vectors for both images and sketches.

To obtain better discriminative feature vectors that can preserve fine-grained details, we use the annotated quintuple $\{(s_i, p_i^+, t_i^+, p_i^-, t_i^-)\}_i^K$ as the supervised information to train the neural networks. Each quintuple consists of a query sketch s and two images $p^+$and p^- with their descriptions $t^+$and t^-, which are named as the positive and negative sample. For Fine-grained SBIR, the positive samples are selected from the target image set that shares both the fine-gained visual similarity and semantic similarity to the query sketch, while the negative ones are selected from the residual irrelevant sets. Thus there are two branches of CNNs for images. Since the positive and negative images are in the same modality, their CNNs share the same architecture and parameters.

2.2 Recurrent Neural Network for Semantic Embedding

RNN has been widely used in natural language processing due to its effectiveness in learning significant patterns of sequential data. Thus we use RNN to model the textual descriptions for annotated images. Given the textual descriptions $[t_1, t_2, t_3, \dots, t_N]$ of the images $[p_1, p_2, p_3, \dots, p_N]$, the goal is to learn the semantic embedding function which can be defined as:

$$[T_1, T_2, T_3, \dots, T_N] = RNN_T([t_1, t_2, t_3, \dots, t_N]|\theta_T) \quad (2)$$

where $T_i \in R^d$ is the learnt intermediate vectors for the i-th image textual description in the embedding space; and θ_T denotes the parameters of the semantic embedding function $RNN_T(\cdot)$. Thus we first extract the deep semantic representation for each image textual description from the RNN-based sentence embedding model Skip-thought [9], which is pre-trained on a large novel corpus from the *BookCorpus* dataset [10]. The Skip-thought model aims at learning deep sentence vector representations, which are good at mapping similar sentences that share 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. This 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 (S_{i-1}, S_i, S_{i+1}), let X^t be the word2vector representation of the t-th word in the sentence S_i and M be the numbers of words in the sentence. For encoder, a GRU is used and a hidden state h^t_i is produced at each time step, which can be formulated as:

$$z^t = \sigma(W_z \cdot [h^{t-1}, X^t])$$
$$r^t = \sigma(W_r \cdot [h^{t-1}, X^t])$$
$$\tilde{h}^t = \tanh(W \cdot [r^t \cdot h^{t-1}, X^t]) \quad (3)$$
$$h^t = (1 - z^t) * h^{t-1} + z^t * \tilde{h}^t$$

where z^t is the update gate vector; r^t is the reset gate vector; and $\tilde{h}^t$ is the state update vector at the time step t. Thus h^M_i can

be interpreted as the feature vector for the full sentence S_i. For the decoder, two GRUs are used, in which one is for generating the previous sentence S_{i-1} and the other for generating the next sentence S_{i+1}. These two GRUs are trained separately without sharing any parameters. Since they share the same computation pattern, we only formulate the decoding process of the next sentence S_{i+1} as follows.

$$\begin{aligned} z^t &= \sigma(W_z \cdot [h^{t-1}, X^{t-1}, h_i]) \\ r^t &= \sigma(W_r \cdot [h^{t-1}, X^{t-1}, h_i]) \\ \tilde{h}^t &= \tanh(W \cdot [r^t \cdot h^{t-1}, X^{t-1}, h_i]) \\ h_{i+1}^t &= (1 - z^t) * h^{t-1} + z^t * \tilde{h}^t \end{aligned} \quad (4)$$

where h_{i+1}^t 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 h_i of the sentence S_i. The sum of log-probabilities for the previous and next sentences is used as the objective function to guide the training for the triplet adjacent sentences (S_{i-1}, S_i, S_{i+1}).

$$\sum_t \log P(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t \log P(w_{i-1}^t | w_{i-1}^{<t}, h_i) \quad (5)$$

where $w^1{}_i, \ldots, w^M{}_i$ denotes the words in the sentence S_i. Thus the loss function of the Skip-thought model can be achieved by summing up all the training triplets. We follow the combine-skip pattern in [9], use the learnt encoder as the feature extractor, and extract the 4,800-way vector as the deep semantic feature for each description. Such deep semantic representations are then taken into three layers of fully connected neural network to learn the semantic embedding function in an end-to-end manner. The intermediate vectors from the second layer are taken as the embedding vectors in the shared space. Similar to the visual embedding, two subnets for positive and negative samples ($t^+$and t^-) in the annotated quintuple share the same architecture and parameters.

2.3 Objective Function

In the learning process of deep multimodal embedding model, we aim at learning the embedding functions that map the input sketches, images and textual descriptions into a common space, in which the images and textual descriptions that are relevant to query sketches are closer than those irrelevant ones. Specifically, given the intermediate feature vectors $\psi = \{(S_i, P_i^+, T_i^+, P_i^-, T_i^-)\}_i^K$ in that space of annotated quintuples, the goal can be formulated as:

$$\begin{aligned} DF(S_i, P_i^+, T_i^+) &< DF(S_i, P_i^-, T_i^-) \\ DF(S_i, P_i^+, T_i^+) &= ||S_i - P_i^+||^2 + ||S_i - T_i^+||^2 \\ DF(S_i, P_i^-, T_i^-) &= ||S_i - P_i^-||^2 + ||S_i - T_i^-||^2 \end{aligned} \quad (6)$$

To achieve this goal, we extend the classic ranking loss to adjust itself to three modalities that have the stronger ability to characterize cross-modal sketch-image correlations, named as multimodal ranking loss, which is formulated as:

$$L_{RANK} = \sum_{i \epsilon \psi} max\big(0, m + DF(S_i, P_i^+, T_i^+) - DF(S_i, P_i^-, T_i^-)\big) \quad (7)$$

where m is the margin to control the relative distance between the positive and negative pairs. The optimization for the objective function will adjust the parameters to obtain the desired feature embedding function that satisfies the ranking order. We also utilize the classification loss to capture the category-level semantics of sketches and images in that embedding space. The classical "*Softmax*" loss function is used for each input modality when training. Let $\phi = \{(S'_i, P'^+_i, T'^+_i, P'^-_i, T'^-_i)\}_i^K$ be the output of the last fully connected layer in the deep neural network for sketches, images and textual descriptions and $sl(\cdot)$ be the *Softmax* loss function. Thus, the overall classification loss can be formulated by combining all the predictions as follows:

$$L_{CATEGORY} = \sum_{i \epsilon \phi} sl(S'_i) + sl(P'^+_i) + sl(T'^+_i) + sl(P'^-_i) + sl(T'^-_i) \quad (8)$$

Thus the overall objective function of our deep multimodal embedding model is a combination of classification loss and multimodal ranking loss, which is formulated as follows:

$$min: \; cL_{RANK} + (1 - c)L_{CATEGORY} + \lambda ||\theta||_2^2 \quad (9)$$

where θ denotes the parameters of embedding neural networks; the last addend is a regularization term; and the parameter c is a constant to control the impact of two loss functions. After adequate training, the deep multimodal embedding model can capture both the object category-level semantics and fine-grained details, such as pose, viewpoint and shape. We take the intermediate feature vectors in the embedding space as the feature representation for sketches, images and textual descriptions. In the testing time, given the embedded intermediate features of a query sketch S_i and a database image P_i with the textual description T_i, we compute the distance value between them as:

$$D(S_i, P_i, T_i) = ||S_i - P_i||^2 + ||S_i - T_i||^2 \quad (10)$$

Thus the distance between query sketches and each annotated image in the whole database can be measured at both visual and semantic level. The ranking list of relevant images is obtained by sorting the distance values of sketch-image pairs.

3 EXPERIMENT AND ANALYSIS

To evaluate our approach, we use a public benchmark dataset, i.e., large-scale *Sketchy* database [6]. It has 12,500 photos and 75,471 sketches of 125 object categories. Each category contains 100 images and each photo has at least 5 well-drawn sketches along with the textual descriptions. 1,250 photos and their sketches are selected for testing and the rest for training. We use the same evaluation metrics as [6], i.e., *Recall@K*, which can be regarded as the percentage of sketches whose true-match photos are ranked in the Top-*K* and quantified by the cumulative matching accuracy at various ranks. The positive samples for each sketch is set as the true-matched images and textual descriptions. The negative ones are obtained from both intra-category and inter-category of images. We keep the sampling ratio between the intra- and inter-category at 5:1, and the margin m of multimodal ranking loss is set to 00.

Our deep multimodal embedding model aims at fusing beneficial multimodal information in sketches and annotated images, and analyzing their underlying correspondence correlations to further enhance Fine-grained SBIR. To investigate the contribution of each component, we introduce two evaluation patterns: 1) investigating the impact of semantic embedding for Fine-grained SBIR: we introduce a baseline of *GN Triplet*, which is similar to our approach but only uses visual comparisons; 2) comparing the effect of each component of objective function: we set the constant parameter c in the objective function to 0, 0.5 and 1. When c=0, only the classification loss is used. When c =1, we only use the multimodal ranking loss. When c = 0.5, both the

classification loss and multimodal ranking loss are utilized. The related experimental results are shown in Table 1.

It can be found from Table 1 that the best performance for Fine-grained SBIR can be obtained in the evaluation pattern of our full deep multimodal embedding model *Ours* (c=0.5). This confirms the obvious advantage of our whole framework for Fine-grained SBIR. Comparing *Ours* (c=0/0.5/1) with *GN Triplet* (c=0/0.5/1), the improved *Rank@K* value proves the positive impact of introducing the semantic comparisons between sketches and textual descriptions of images. It can be concluded that the Fine-grained SBIR performance can be further enhanced through mining all the possible beneficial multimodal information sources in annotated images, rather than only considering the unimodal visual information in images. Furthermore, comparing the different settings for the particular parameter c, the best performance can be achieved when c is set to 0.5 while the modest performance is obtained when c is set to 0. The main reason is that the classification loss can only capture the category-level semantics and cannot preserve the fine-grained details for Fine-grained SBIR. The multimodal ranking loss can do it better because it's based on the relative similarities of the input quintuple, which is usually judged by the fine-grained information. However, by combining such two kinds of loss function, we can learn better embedding function that capture both the category-level semantics and the fine-grained details to further improve Fine-grained SBIR.

To give full exhibition to the superiority of our model, we have performed a comparison between our model and several state-of-the-art approaches. These methods include: (1) deep ranking networks: *GN Triplet* (*GN Triplet* (c=0.5)), *GN Siamese*, *GN Triplet w/o Cat* (*GN Triplet* (c=1)), and *AN Siamese*; (2) recognition-based networks: *GN Cat* (*GN Triplet* (c=0)), *SN w/Label* and *SN*; (3) traditional hand-crafted feature: *GALIF* [11]; and (4) Simple baselines by chance: *Chance* and *Chance w/Label*. More detailed implementation information of these methods can be found in [6]. The recall value of different ranking positions, *Rank@K* (K=1, 2, 3, …, 10), are presented in Figure 2. Similar conclusion can be drawn as above. It's worth noting that some of these methods 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, after introducing the semantic comparisons between sketches and textual descriptions of images, the recall value can be further improved, especially at the high ranking position, i.e., Top-1. Meanwhile, the deep-learning-based approaches significantly outperform the basic methods with hand-crafted features like *GALIF*, which mainly attributes to the effort of supervised training.

4 CONCLUSIONS

In this work, we present a deep multimodal embedding model to support more precise Fine-grained SBIR for large-scale annotated images. Our model can encode multimodal information into a common space through deep neural networks, and exploit the fine-grained cross-modal correspondences among the attributes of different modalities in sketches and annotated images in that space. An interesting future direction is to extend our method to multimodal retrieval, where more semantic information will be used along with sketches to more precisely retrieve target object images from all the different modalities.

Table 1: The experimental results in different patterns.

Patterns	*Rank@1*	*Rank@5*	*Rank@10*
Human	0.5427	-	-
GN Triplet (c=0)	0.1263	0.4457	0.6630
GN Triplet (c=1)	0.2278	0.5326	0.6630
GN Triplet (c=0.5)	0.3710	0.8090	0.9451
***Ours* (c=0)**	0.1343	0.4672	0.6486
***Ours* (c=1)**	0.2710	0.5926	0.7480
***Ours* (c=0.5)**	**0.4216**	**0.8360**	**0.9728**

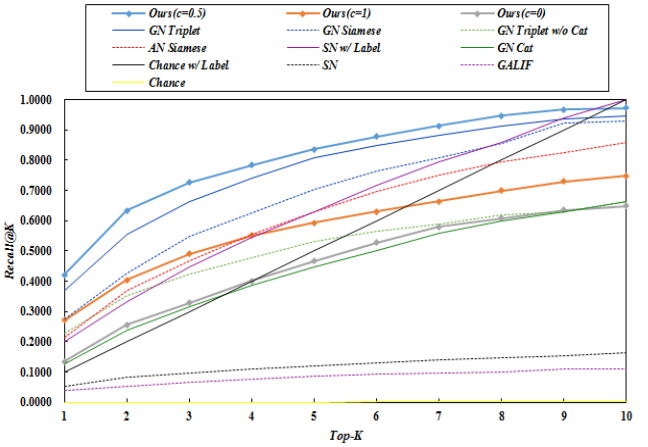

Figure 2: The experimental comparison results of *Recall@K*.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Fund of China (61572140; 61672165), Shanghai Municipal Science and Technology Commission (16JC1420401; 16511105402; 16511104704), Shanghai Municipality Program of Technology Research Leader (17XD1425000) and The Application of Big Data Computing Platform in Smart Lingang New City based BIM and GIS (#ZN2016020103). Yuejie Zhang is the corresponding author.

REFERENCES

[1] A. Chalechale, G. Naghdy, and P. Premaratne. Sketch-based shape retrieval using length and curvature of 2d digital contours. In *IWCIA*, pages 474-487, 2004.

[2] Y. Wang, M. Yu, and Q. Jia. Query by sketch: An asymmetric sketch-vs-image retrieval system. *CISP*, 3:1368-1372, 2011.

[3] Y. Li, T.M. Hospedales, Y-Z. Song, and S. Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[4] P. Xu, Q. Yin, Y. Qi, Y-Z. Song, Z. Ma and L. Wang. Instance-Level Coupled Subspace Learning for Fine-Grained Sketch-Based Image Retrieval. In *ECCV*, pages 19-34, 2016.

[5] Q. Yu, F. Liu, Y-Z Song, T. Xiang, T. M. Hospedales, and C-C Loy. Sketch me that shoe. In *CVPR*, pages 799-807, 2016

[6] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 35(4):119, 2016.

[7] N. Dalal and T. Bill. Histograms of oriented gradients for human detection. In *CVPR*, pages 886-893, 2005.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1-9, 2015.

[9] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, pages 3294-3302, 2015.

[10] Y. 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 *ICCV*, pages 19-27, 2015.

[11] M. Eitz, R. Richter, T. Boubekeur, et al. Sketch-based shape retrieval. ACM Trans. Graph. 31(4): 31:1-31:10, 2012.