ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1626

Enhancing Sketch-Based Image Retrieval via Deep Discriminative Representation

Fei Huang, Yong Cheng, Cheng Jin, Yuejie Zhang[1] and **Tao Zhang**[2]

Abstract. In this paper we aim to employ deep learning to enhance SBIR via deep discriminative representation. Our main contributions focus on: 1) The deep discriminative representation is established to bridge both the visual appearance gap and the semantic gap between sketches and images; 2) The deep learning pattern is applied to our SBIR model through training on our transformed sketch-like images to overcome the rarity of training sketches. Our experiments on a large number of public sketch and image data have obtained very positive results.

1 INTRODUCTION

Recently, Sketch-based Image Retrieval (SBIR) is an important research topic with the popularity of touch screen devices. Sketches only contain main strokes or lines of target images and lack rich texture attributes and luminance, which results in the ambiguity for the sketch query among different users. To achieve more robust SBIR, it is significant to bridge both the visual appearance gap and the semantic gap between sparse sketches and colorful images. In this paper we aim to employ deep learning to enhance SBIR via deep discriminative representation. We adopt the low-level visual feature and Convolutional Kernel Network (CKN) to capture the local visual attributes, the Convolutional Neural Network (CNN) model to capture semantic information of sketches and images, and finally the deep discriminative representation is obtained by multimodal feature fusion. Since the existing CNN models trained on the benchmark dataset of ImageNet ineffective to the sketch representation, the major barrier for SBIR is the lack of training samples. We solve this issue by implementing the training on sketch-like images transformed from real natural images. Our experiments on a large number of public data have obtained very positive results.

2 DEEP DISCRIMINATIVE REPRESENTATION

We aim at constructing the deep discriminative representation based on the specific consideration of both low-level visual features and high-level semantic features, which can be achieved by the multimodal fusion of CNN features and local features. Before that, the sketch-like images are generated by the SE detector [1] for original images to bridge the visual domain gap.

Local Feature Descriptor

To construct the local feature descriptor with the stronger descriptive ability, we choose the BoVW framework but adopt Convolutional Kernel Networks (CKNs) [2] to learn the local convolutional features as visual words, called as BCKN.

CKN takes an image or a patch as an input. The feature representation is based on a positive-definite kernel function which can be approximated as $<\xi, \xi'>$, where $<,>$ represents the Euclidean inner-product function. Let Ω be the coordinates from the input and $z \in \Omega$, Ω_1 be a subset of Ω and $|\Omega_1|<|\Omega|$, and ξ is computed for all $u \in \Omega_1$, shown as follows:

$$\xi = \sqrt{\frac{2}{\pi}} \sum_{z \in \Omega} e^{-\frac{\|u-z\|^2}{\beta^2}} g(z) \quad (1)$$

$$g(z) = \| \varphi(z) \| \left[\sqrt{\eta_l} e^{-\frac{1}{\alpha^2}\|\tilde{\varphi}(z)-w_l\|^2} \right]_{l=1}^{p} \quad (2)$$

where $\varphi(z)$ represents a fixed sub-patch centered at z ; $\tilde{\varphi}(z)$ is a normalized version of $\varphi(z)$; α and β are two smoothing parameters of the Gaussian kernel; and p is the number of filters. ξ' is computed in the same way as ξ. The stochastic gradient descent is used to optimize the parameters $[w_l]_{l=1}^{p}$ and $[\eta_l]_{l=1}^{p}$ as follows:

$$\min_{w,\eta} \frac{1}{n} \sum_{i=1}^{n} \left[e^{-\frac{1}{2\alpha^2}\|x_i-y_i\|^2} - \sum_{j=1}^{p} \eta_j e^{-\frac{1}{\alpha^2}\|x_i-w_j\|^2} e^{-\frac{1}{\alpha^2}\|y_i-w_j\|^2} \right]^2 \quad (3)$$

where $\{(x_i, y_i)\}_{i=1, \ldots, n}$ is n pairs of training patches.

In practice, we construct the single-layer CKN using Formula (1) and (2). Such the single-layer CKN can be superimposed on each other as the multi-layer CKN to acquire the better representation. The above strategy can provide a robust local representation for our low-level features. The BoVW framework is applied to encode the local features to the global representation. First, with the sketch or sketch-like images, a set of CKN features are extracted with the (16 ×16) windows centered at interest (contour) points to learn a visual codebook using k-means. Then, each sketch or sketch-like image can be quantized to a feature vector representation with BCKN.

CNN-based Feature Descriptor

Semantic information is very important for SBIR, especially when the distortion of freehand sketches is difficult to overcome only by using low-level local features. To fully capture the discriminative semantic information, we utilize the famous Alexnet [3] and GoogLeNet [4], which recently achieve the impressive performance for CBIR, and are extended to learn the deep semantic representation for both sketches and sketch-like images in SBIR.

Since sketches and images are two domains of visual exhibitions, the *Alexnet*/*GoogLeNet* model trained on images cannot be directly applied to SBIR. In particular, we need to solve the following two crucial issues: 1) The existing SBIR datasets are especially uneven with a majority of color images, thus there is a lack of sketch training

[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China, email: {15210240036;13110240027; jc; yjzhang}@fudan.edu.cn

[2] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China, email: taozhang@mail.shufe.edu.cn

samples; and 2) With the absence of training samples, how to extend the classification-based CNN models to SBIR without losing the discrimination? For the first issue, we can rationally assume that the obtained sketch-like images retain the main outlines of initial images, which can be approximately regarded as hand-drawn sketches. Thus such sketch-like images can be used as training samples to retrain a CNN model for SBIR. Furthermore, a lot of images on the sharing websites are annotated with tags, such as *Flickr*, which can provide relatively affluent labels for training images. Here, we only consider an image with only one category label for simplicity. For the second issue, we extract the prediction scores of the last layer in both the retrained *Alexnet* model and *GoogLeNet* model as high-level semantic features, since they can capture the discriminative semantic information from both sketches and sketch-like images.

Multimodal Feature Fusion

We combine the two cosine distances of the visual feature D_{vis} and semantic feature D_{sem} through the multimodal feature fusion to form the final ranking list as follows.

$$D_{final} = \alpha \cdot D_{vis} + (1-\alpha) \cdot D_{sem} \quad (4)$$

where $\alpha \in [0,1]$, and we usually set $\alpha = 0.5$ to balance the effect of two kinds of features in practice.

3 EXPERIMENT AND ANALYSIS

To evaluate the effectiveness of our approach, we use one benchmark dataset *Flickr15k* created by Hu *et al*. [5] (including 330 sketch queries and 14,660 images), and another dataset for a specific application scenario *MECD* (including 30 sketch queries and 900 images) from a large Chinese museum collection image set. The official criteria of Mean Rank Precision (MRP) and Average Precision (AP) are introduced to evaluate the whole retrieval performance. *MRP* aims at measuring the precision of the top-*k* returned relevant images, which is defined as:

$$MRP(k) = \sum_{i=1}^{N} Precision_i(k) / N$$
$$Precision_i(k) = n(T) / k \quad (5)$$

where $n(T)$ is the number of relevant images in the top-*k* returned results; and N is the total number of sketch queries. *AP* considers the order that the returned images are presented, which is defined as:

$$AP(k) = \frac{1}{k}\sum_{r=1}^{k} rel(r) \times MRP(r) \quad (6)$$

where $rel(r) \in \{0, 1\}$ is an indicator function, if the image at Rank r is relevant to the sketch query $rel(r)=1$, otherwise $rel(r)=0$.

Our approach is a new exploration for taking full advantage of semantic information for images, and a new deep discriminative representation framework is proposed. We introduce three popular *CNN* models, that is, *Alexnet*, *GoogLeNet* and *Siamese Network* [6], and retrain them on our sketch-like images. For *Siamese Network*, we adopt the same architecture in [6]. The *Baseline*[*BCKN*] only uses our local *BCKN* features, while *DDR*[*GoogLeNet*] and *DDR*[*Alexnet*] implement the complete framework. The related comparison results are presented in Table 1. An instantiation of four sketch queries and their relevant images is shown in Figure 1.

It can be found from Table 1 that the supervised CNN outperform the unsupervised BCKN with absolute advantage on both datasets. The best performance can be acquired by our complete approach on *Flickr15k* and *MECD*. Comparing the results based on *Alexnet*, *GoogLeNet* and *Siamese Network*, *Siamese Network* obtains the worst results because it exploits the pair-wise image similarity or dissimilarity as the supervised information, which is much weaker than the direct labeling information. By comparing the performance of our *DDR*-based approach with *CNN* models and *Baseline*[*BCKN*], there is something indicating that it is effective to jointly use both low-level visual feature and high-level semantic feature for SBIR.

Table 1. The comparison results between our and the other approaches.

Dataset	Approach		Feature Dimension	Evaluation Metric	
				MRP(20)	*AP*(20)
Flickr15k	Other Approach	*GoogLeNet*	1,000	0.701	0.694
		Alexnet	1,000	0.646	0.605
		Siamese	64	0.492	0.485
	Our Approach	*Baseline*[*BCKN*]	300	0.285	0.274
		DDR[*GoogLeNet*]	-	***0.717***	***0.715***
		DDR[*Alexnet*]	-	***0.686***	***0.671***
MECD	Other Approach	*GoogLeNet*	1,000	0.715	0.707
		Alexnet	1,000	0.662	0.642
		Siamese	64	0.514	0.493
	Our Approach	*Baseline*[*BCKN*]	300	0.335	0.305
		DDR[*GoogLeNet*]	-	***0.730***	***0.738***
		DDR[*Alexnet*]	-	***0.705***	***0.684***

Figure 1. An instantiation of some retrieval results with our approach.

4 CONCLUSIONS AND FUTURE WORK

In this work, we present a novel scheme with deep discriminative representation for SBIR. We apply the deep learning pattern and propose an effective deep discriminative representation to encode both low-level and high-level features for sketches and sketch-like. We believe this is just the beginning to extend deep learning to SBIR, and in the future we will further explore the unsupervised SBIR-based CNN framework to further boost the retrieval performance.

5 ACKNOWLEDGMENTS

This work is supported by National Natural Science Fund of China (61572140), Shanghai Municipal R&D Foundation (16511105402& 16511104704), Shanghai Philosophy Social Sciences Planning Project (2014BYY009), and Zhuoxue Program of Fudan University. Yuejie Zhang is the corresponding author.

REFERENCES

[1] P. Dollár and C.L. Zitnick. 'Structured forests for fast edge detection'. *In Proceedings of ICCV 2013*, 1841-1848, (2013).

[2] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid. 'Convolutional kernel networks'. *In Proceedings of NIPS 2014*, 2627-2635, (2014).

[3] A. Krizhevsky, I. Sutskever, and G.E. Hinton. 'Imagenet classification with deep convolutional neural networks'. *In Proceedings of NIPS 2012*, 1097-1105, (2012).

[4] C. Szegedy, W.Liu, Y.Jia, P.Sermanet, S.Reed, D.Anguelov, D.Erhan, V.Vanhoucke, A.Rabinovich. 'Going deeper with convolutions'. In *Proceedings of CVPR 2015*, 1-9, (2015).

[5] R. Hu and J. Collomosse. 'A performance evaluation of gradient field hog descriptor for sketch based image retrieval'. *Computer Vision and Image Understanding*, 117(7):790-806, (2013).

[6] F.Wang, L.Kang, Y.Li. 'Sketch-based 3d shape retrieval using convolutional neural networks'. In *Proceedings of CVPR 2015,* 1875-1883, (2015).