Information Sciences 430–431 (2018) 331–348

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Large-scale semantic web image retrieval using bimodal deep learning techniques

Changqin Huang[a,b,*], Haijiao Xu[a,*], Liang Xie[c], Jia Zhu[b], Chunyan Xu[d], Yong Tang[b]

[a] *School of Information Technology in Education, South China Normal University, Guangzhou, China*
[b] *Guangdong Engineering Research Center for Smart Learning, South China Normal University, Guangzhou, China*
[c] *School of Science, Wuhan University of Technology, Wuhan, China*
[d] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China*

ARTICLE INFO

Article history:
Received 24 July 2017
Revised 18 November 2017
Accepted 20 November 2017
Available online 21 November 2017

Keywords:
Convolutional neural networks
Multi-concept scene classifiers
Concept based image retrieval
Bimodal learning

ABSTRACT

Semantic web image retrieval is useful to end-users for semantic image searches over the Internet. This paper aims to develop image retrieval techniques for large-scale web image databases. An advanced retrieval system, termed Multi-concept Retrieval using Bimodal Deep Learning (MRBDL), is proposed and implemented using Convolutional Neural Networks (CNNs) which can effectively capture semantic correlations between a visual image and its free contextual tags. Different from existing approaches using multiple and independent concepts in a query, MRBDL considers multiple concepts as a holistic scene for retrieval model learning. In particular, we first use a bimodal CNN to train a holistic scene classifier in two modalities, and then semantic correlations of the sub-concepts included in the images are leveraged to boost holistic scene recognition. The predicted semantic scores obtained from holistic scene classifier are combined with complementary information on web images to improve the retrieval performance. Experiments have been carried out over two publicly available web image databases. The results show that our proposed approach performs favorably compared with several other state-of-the-art methods.

1. Introduction

It is well-known that social media data is an indispensable part of our life. Large-scale web information retrieval becomes a very important topic in information technology. A common characteristic of social media is that users commonly add extra information directly or indirectly to the media. This metadata associates multi-concept semantics to the media that can be used for retrieval. For example, web users often assign a variety of social tags [8] or comments to web images for sharing, then retrieve and utilize them via social websites. Therefore, for large-scale retrieval over semantic image databases, an efficient approach is vital to process users' semantic queries.

The tags or textual descriptions of web images are informative but mostly noisy. Fig. 1 shows some query samples from *NUS-WIDE* dataset [5]. Obviously, the social tags associated with web images, conveying semantic information, can be taken as semantic description. Hence, a straightforward idea to implement such a retrieval system is to use the web image tags in the retrieval system design. However, these tags are often ambiguous and noisy, leading to undesirable retrieval

[*] Corresponding author.
E-mail addresses: cqhuang@scu.edu.cn (C. Huang), guesskkk@hust.edu.cn (H. Xu).

https://doi.org/10.1016/j.ins.2017.11.043
0020-0255/© 2017 Elsevier Inc. All rights reserved.

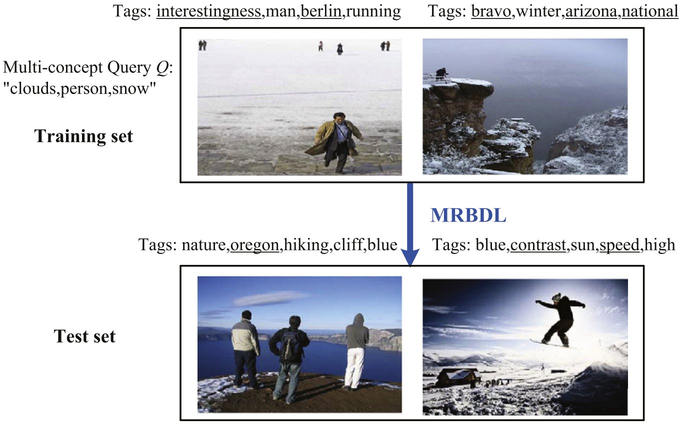
Tags: interestingness,man,berlin,running
Tags: bravo,winter,arizona,national
Multi-concept Query Q: "clouds,person,snow"
Training set
MRBDL
Tags: nature,oregon,hiking,cliff,blue
Tags: blue,contrast,sun,speed,high
Test set

Fig. 1. Query samples from the *NUS-WIDE* dataset. Given a multi-concept query $Q =$ “clouds, person, snow”, example images from the training set and the test set are shown in the top and bottom row respectively.

performance. For example, as shown in Fig. 1, the underlined tags such as “interestingness”, “bravo” and “national” in the first row do not directly reflect the visual content, and the tags such as “Oregon” and “speed” in the second row only partially depict the visual content. Although the noisy social tags with weak semantic information cannot be directly used as retrieval concepts, they can be considered as auxiliary low-level features (*i.e.* text modality) for image retrieval.

For web image retrieval, conventional unimodal approaches employ either visual modality [32] or text modality [8]. To boost web image retrieval performance, multi-modal and cross-modal retrieval approaches, exploring the correlations of these two modalities, have been proposed in [30]. Most existing work concentrates on single-concept-based image retrieval, where each query is assumed to have only one concept. This is inconsistent with real-world scenarios where a user always conducts retrieval with multiple concepts, namely the Multi-Concept-based Image Retrieval (MCIR). As shown in Fig. 1, given a multi-concept query $Q =$ “clouds, person, snow”, web images simultaneously describing the three concepts are returned from the database. To tackle the MCIR problem, traditional approaches are ineffective as a multi-concept scene may contain unique visual characteristics that are difficult to identify solely by single-concept classifiers [11]. Thus, further researches on the multi-concept-based image retrieval are significant and useful.

Recently, CNNs have been applied for single-concept-based image retrieval [23]. The achieved performance is promising, and indicates that deep descriptors learned by CNN can well capture the underlying semantic structures of images. Inspired by this work, we propose a multi-concept retrieval approach using deep learning techniques to resolve the MCIR problem. In our proposed framework, MRBDL effectively combines the single-concept classifier and the holistic scene classifier in the visual and text modalities, respectively. From our observations on experimental results, such a design schema can substantially improve the discriminative power of the classifier for multi-concept scene recognition. In our proposed MRBDL, we firstly devise a bimodal CNN where the training images and associated texts are separately fed into the corresponding convolution block layers and the Fully-Connected (FC) classifier layers. The FC classifier layer is composed of two types of classifiers, that is, the single-concept FC classifier that best suits single-concept recognition, and the multi-concept scene FC classifier contributes to holistic scene recognition. Next, a two-phase training strategy is proposed to train the bimodal CNN. Finally, the semantic correlations among concepts are utilized to estimate the semantic scores of the concepts in order to enhance the discriminative capability of FC classifiers. If a concept C_j and its related semantic concepts C_r have a high co-occurrence frequency in the image set, we boost the semantic score of predicting this concept C_j. To combine the complementary information from the visual and text modalities, we make an ensemble of these predicted semantic scores by a fusion operator. To compensate the varying frequencies of concepts derived from imbalanced image datasets [29], the gradient descent algorithm is applied for maximizing the log-likelihood of the semantic scores over the training images.

The rest of the paper is organized as follows. Section 2 briefly reviews some related work. Section 3 details the proposed MRBDL framework. Section 4 describes our experimental setup. Section 5 reports the experiments with results and analysis. Finally, Section 6 concludes this paper.

2. Related work

This section provides some background knowledge, including unimodal, multi-modal and cross-modal image retrieval techniques, and deep learning concept associated with CNNs.

2.1. Unimodal learning

Unimodal image retrieval systems can be roughly grouped into two categories: Content-Based Image Retrieval (CBIR) and COncept-based Image Retrieval (COIR). Design of CBIR systems is usually based on local and global visual descrip-