Signal Processing: Image Communication 63 (2018) 30–43

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Deep convolutional image retrieval: A general framework

Maria Tzelepi [*], Anastasios Tefas

Aristotle University of Thessaloniki, Department of Informatics, Greece

ARTICLE INFO

Keywords:
Content based image retrieval
Convolutional neural networks
Deep learning
Query expansion

ABSTRACT

In this paper a Convolutional Neural Network framework for Content Based Image Retrieval is proposed. We employ a deep CNN model to obtain the feature representations from the activations of the deepest layers and we retrain the network in order to produce more efficient image descriptors, relying on the available information. Our method suggests three basic model retraining approaches. That is, the Fully Unsupervised Retraining, if no information except from the dataset itself is available, the Retraining with Relevance Information, if the labels of the dataset are available, and the Relevance Feedback based Retraining, if feedback from users is available. We propose these approaches independently or in a pipeline, where each retraining approach operates as a pretraining step to the subsequent one. We also apply a query expansion method with spatial reranking on top of these approaches in order to boost the retrieval performance. The experimental evaluation on six publicly available image retrieval datasets indicates the effectiveness of the proposed method in learning more efficient representations for the retrieval task, outperforming other CNN-based retrieval techniques, as well as conventional hand-crafted feature-based approaches.

1. Introduction

Information Retrieval (IR) refers to the process of obtaining material (text documents, images, audio etc.) that satisfies a certain information need from large databases [1]. Over the long history of IR, numerous works emerged in the field of text retrieval [2], audio [3], video [4], and image retrieval [5]. Image retrieval is a research area of IR of great scientific interest since 1970s. Earlier studies include manual annotation of images using keywords and searching by text [6]. Due to the difficulties of text-based image retrieval, deriving from the manual annotation of images, that is based on the subjective human perception, and the time and labor requirements of annotation, in 1990s Content Based Image Retrieval (CBIR) has been proposed [7].

The objective of CBIR is to retrieve images that are relevant to a query image from a large collection based on their visual content [8]. A key issue concerning CBIR is to extract meaningful information from raw data in order to eliminate the so-called semantic-gap [9]. The semantic-gap refers to the difference between the low level representations of images and their higher level concepts. While earlier works focus on primitive features that describe the image content such as color, texture, and shape, numerous more recent works have been elaborated on the direction of finding semantically richer image representations. Among the most effective are those that use the Fisher Vector descriptors [10], Vector of Locally Aggregated Descriptors (VLAD) [11] or combine bag-of-words models [12] with local descriptors such as Scale-Invariant Feature Transform (SIFT) [13].

Several recent studies introduce Deep Learning algorithms [14] against the shallow aforementioned approaches to a wide range of computer vision tasks, including image retrieval [15–18]. The main reasons behind their success are the availability of large annotated datasets, and the GPUs computational power and affordability.

Deep Convolutional Neural Networks (CNN), [19,20], are considered the more efficient Deep Learning architecture for visual information analysis. CNNs comprise of a number of convolutional and subsampling layers with non-linear neural activations, followed by fully connected layers (an overview of the utilized network is provided in Fig. 1). That is, the input image is introduced to the neural network as a three dimensional tensor with dimensions (i.e., width and height) equal to the dimensions of the image and depth equal to the number of color channels (usually three in RGB images). Three dimensional filters are learned and applied in each layer where convolution is performed and the output is passed to the neurons of the next layer for non-linear transformation using appropriate activation functions. After multiple convolution layers and subsampling the structure of the deep architecture changes to fully connected layers and single dimensional signals. These activations are usually used as deep representations for classification, clustering or retrieval.

[*] Corresponding author.
E-mail addresses: mtzelepi@csd.auth.gr (M. Tzelepi), tefas@aiia.csd.auth.gr (A. Tefas).

https://doi.org/10.1016/j.image.2018.01.007
Received 23 August 2017; Received in revised form 12 December 2017; Accepted 25 January 2018
Available online 2 February 2018

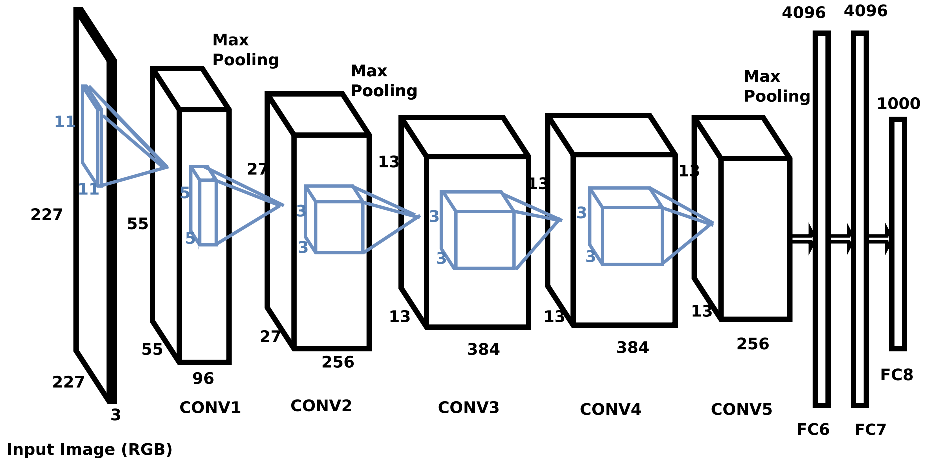

Fig. 1. Overview of the CaffeNet architecture.

Over the last few years, deep CNNs have been established as one of the most promising avenues of research in the computer vision area due to their outstanding performance in a series of vision recognition tasks, such as image classification [21,22], face recognition [23,24], digit recognition [25,26], pose estimation [27], object and pedestrian detection [28,29], and action recognition [30]. It has also been demonstrated that features extracted from the activation of a CNN trained in a fully supervised fashion on a large, fixed set of object recognition tasks can be re-purposed to novel generic recognition tasks, [31]. Inspired by these results, deep CNNs introduced in the vivid research area of CBIR. The primary approach of applying deep CNNs in the retrieval domain is to extract the feature representations from a pretrained model by feeding images in the input layer of the model and taking activation values usually drawn from the last layers, while several recent works are directed at utilizing the convolutional layers for the feature extraction. Current research also includes model retraining approaches, which are more relevant to our work, while other studies focus on the combination of the CNN descriptors with conventional descriptors like the VLAD representation. The existing related works are discussed in the following section.

Our work investigates model retraining approaches in order to enhance the deep CNN descriptors. We employ a pretrained model to derive feature representations from the activations of the deepest layers and we retrain the model, exploiting the idea that a deep neural architecture can non-linearly distort the feature space in order to modify the feature representations, with respect to the available information. This information can consist in only the dataset to be searched, the labels of the dataset or of a part of the dataset, and finally information acquired from users' feedback, that is, relevant or irrelevant images as deemed by multiple users.

In this paper we propose a general framework for CNN model retraining in the retrieval domain. The contributions of our study can be summarized as follows:

- To the best of our knowledge this is the first work that is able to exploit any kind of available information about the retrieval task. The proposed retraining approaches of our method can be categorized as follows:

  Fully Unsupervised Retraining (FU): if no information is available, except for the dataset itself.

  Retraining with Relevance Information (RRI): if the labels of the dataset or of a part of the dataset are available.

  Relevance Feedback-based Retraining (RF): if feedback from users is available.

- We deploy combinatory schemes, where all the above approaches can be employed in a pipeline. In this fashion each retraining approach operates as a pretraining step to the subsequent one.

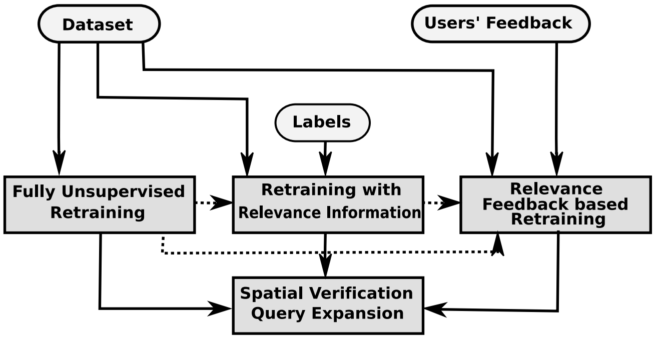
Dataset
Users' Feedback
Labels
Fully Unsupervised Retraining
Retraining with Relevance Information
Relevance Feedback based Retraining
Spatial Verification Query Expansion

Fig. 2. The proposed retraining approaches of our method based on the available information.

- We suggest a query expansion technique with a spatial verification step applicable to all the above cases.
- This is the first approach that uses retargeting for the learning phase, instead of triplet loss, allowing for single sample training which is very fast and can be easily parallelized and implemented in a distributed manner.

In Fig. 2 we schematically describe the proposed framework.

The remainder of the manuscript is structured as follows. Section 2 discusses prior work. The proposed framework is described in detail in Section 3. The proposed spatial verification and query expansion technique is presented in Section 4. The experiments are provided in Section 5. Finally, conclusions are drawn in Section 6.

2. Prior work

In this Section we present previous CNN-based works for image retrieval. Firstly, an evaluation of CNN features in various recognition tasks, including image retrieval that improve the baseline performance using spatial information is presented in [32]. In [33] an image retrieval method, where a CNN pretrained model is retrained on a different dataset with relevant image statistics and classes to the dataset considered at the test time and achieves improved performance, is proposed. From a different viewpoint, in [34,35], CNN activations at multiple scale levels are combined with the VLAD representation. In [36], a feature aggregation pipeline is presented using sum pooling. while in [37] a cross-dimensional weighting and aggregation of deep convolutional neural network layer output is proposed. An approach that produces compact feature vectors derived from the convolutional