Neurocomputing 303 (2018) 60–67

Contents lists available at ScienceDirect

Neurocomputing

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

Optimization of deep convolutional neural network for large scale image retrieval

Cong Bai, Ling Huang, Xiang Pan, Jianwei Zheng, Shengyong Chen*

College of Computer Science and Technology, Zhejiang University of Technology, No.288 Road Liuhe, Hangzhou 310023, China

ARTICLE INFO

Article history:
Received 27 December 2017
Revised 3 April 2018
Accepted 19 April 2018
Available online 4 May 2018

Communicated by Yongdong Zhang

Keywords:
Large-scale image retrieval
Deep learning
Pooling
Activation function
Hidden layer

ABSTRACT

Feature extraction and similarity measurement are two key steps in image retrieval. AlexNet is a classical deep convolutional neural network for image classification, but using it directly for large scale image retrieval is not efficient. To address this issue, we propose a novel framework to improve its ability for feature extraction and its efficiency for similarity measurement. The proposal optimizes AlexNet in three aspects: pooling layer, fully connected layer and hidden layer. In particular, average pooling is replaced by max-ave pooling for better local feature extraction; the non-linear activation function Maxout is used in fully connected layers for better global information extraction and hidden layer is added for mapping high-dimensional feature into binary codes. The proposed framework outperforms state-of-the-art methods on public databases for image retrieval, including large scale database.

1. Introduction

Large scale multimedia information processing has attracted great attentions in recent years due to the rapid growth of data, such as images and videos. Among them, content-based image retrieval (CBIR) is one of important problem for many applications, such as medical image analysis, video surveillance, remote sensing and so on [1,2]. CBIR aims to retrieve the images that have the most relevant visual contents from large databases. To achieve the 'content-based', the analysis of the image contents is essential. Thus feature representation and similarity measurement are crucial fundamentals in CBIR. There is a famous challenging problem named 'semantic gap' in CBIR [3]. The reason of this gap is the different manners of seeing image between human and computer. Human is accustomed to use high-level concepts to describe the visual contents and to measure the similarity. Different from human, computers extract low-level features from image pixels. As there are no direct links between the high-level concepts and the low-level features, the 'semantic gap' exists. To reduce the semantic gap mentioned above, many researches have been conducted. Most of the approaches use different hand-craft features to represent the visual contents for images and try to seek appropriate similarity measurement to make the similarity of low-level features to be more close to the similarity of the high-level concepts [4–6]. Some comprehensive surveys could be found in [7–9].

Artificial intelligence (AI), especially the machine learning technology, has attracted great attentions in the past few years [10–12]. The objective of AI is to allow computers to simulate the human intelligence and to handle the tasks in the real world. It is similar in essence with reducing semantic gaps in CBIR. There are some efforts trying to reduce the semantic gap with machine learning technologies. Especially, deep learning has achieved great progresses in recent years [13], such as deep neural network [14], deep Boltzmann machine [15], deep brief network [16] and so on. Among them, deep convolutional neural network (DCNN) has already got many significant achievements in computer vision, such as image classification [17], image segmentation [18] and object recognition [19].

Using deep learning technology to reduce the semantic gap in CBIR started from last few years [20–23]. In this paper, we make efforts to design a novel DCNN framework for semantic image retrieval in large scale database. The proposed framework optimizes the internal structure of the classical deep convolutional network AlexNet. It improves the feature representation ability of the network. Extensive experiments have been done in four representative databases and the results showed that the proposal outperformed the AlexNet and other state-of-the-art methods. It contains three main contributions:

1) Max-ave pooling for better local feature extraction and representation;
2) Non-linear activation Maxout for global information extraction and representation;

* Corresponding author.
E-mail address: sy@ieee.org (S. Chen).

https://doi.org/10.1016/j.neucom.2018.04.034
0925-2312/© 2018 Elsevier B.V. All rights reserved.

3) Hidden layer for mapping high-dimensional feature vectors into binary codes.

The reminder of this paper is organized as follows: related works are briefly reviewed in Section 2, the proposal is detailed in Section 3, followed by the experiments in Section 4. Finally, the conclusion is given in Section 5.

2. Related works

The contributions of deep learning technology to solve the semantic image retrieval problems could be divided into two categories: using them in the phase of feature extraction and representation and in the phase of similarity measurement.

Using DCNN to extract image features has been proven that it could get better semantic information than hand-craft features. The basic idea is that images are input directly into the DCNN. Features generated by the convolutional layers and pooling layers are used as low-level features and the features extracted from fully-connected layers contained rich semantic information [24]. These features could be used directly in image retrieval [20]. Cosine distance or Euclidean distance could be used to measure the similarity to complete the image retrieval [25]. Meanwhile, using the compact global descriptors learned from the image classification [26] or using the aggregating local descriptors [27] as the feature representation show better performance in image retrieval. Additionally, learning the structure information and color information from the DCNN model pre-trained on ImageNet for retrieval images is proposed [28]. However, the aforementioned methods need large scale training data with labels. The training time will increase along with the deeper of the network. To solve this problem, an unsupervised way for feature extraction on DCNN is proposed [29].

In the aspect of similarity measurement, the core idea is to learn a suitable distance metric that could make the distance of similar images to be minimum and the distance of dissimilarity image to be maximum. For example, Wu et al. [30] proposed an online multimodal deep similarity learning framework to learn both the optimal metrics and the optimal combination of multiple modalities. While Yan et al. [31], Norouzi et al. [32], and Lu et al. [33] used the DCNN to learn hashing functions to measure the similarity. The common idea of these proposals is to learn the hashing functions for mapping the high-dimensional feature vectors into binary codes. Hamming distance is then used to measure the similarity.

This paper aims to extract the feature efficiently to meet the real-time requirement of CBIR. Furthermore, it also maps the high-dimensional feature vectors to low-dimensional hash codes by introducing hidden layer in the DCNN to improve the time efficiency of similarity measurement.

3. Optimization of AlexNet for image retrieval

3.1. Baseline

The proposed method in this paper is based on AlexNet [34], which is a classical deep convolution neural network. AlexNet consists of five convolution layers, three pooling layers and three fully connected layers. The convolution layers and the pooling layers are used to extract image features. The fully connected layers follow the convolution layers and the pooling layers, which map two-dimensional feature vectors into one-dimensional feature vectors. Although the semantic gap could be reduced by adding the depth of the network [35], it also increases the computation time at the same time. We aim at reducing the semantic gap by applying some optimizations on the architectures of AlexNet for accurate and compact image representation.

3.2. Overview of the proposal

The proposed framework based on AlexNet is shown in Fig. 1. AlexNet is optimized in convolutional layers and fully connected layers to get more specific middle-level feature descriptors. (1) The max-ave pooling strategy is adopted in pooling layers for local feature representation. (2) Maxout activation is used in fully connected layers to fit global feature. (3) A hidden layer is added in fully connected layer to convert the global feature vectors into binary codes. The binary codes of the query images and that of the database images are measured by Hamming distance. Top K images are ranked by Hamming distance as the retrieval results. We name this framework as Optimized AlexNet for Image Retrieval (OANIR).

3.3. Feature extraction and representation

In semantic image retrieval, finding good feature extraction and representation is a critical step. Although the classical AlexNet has good performance on feature extraction, it still fails in some challenging tasks with the limitation of network depth. In the OANIR network, some optimizations are made without increasing the depth of the network.

3.3.1. Max-Ave pooling for local features

The key point for pooling is to get core features from joint features while keeping the important features and discarding irrelevant features. Pooling has the advantage for better feature representation. It is a compact representation that is invariant to image transformation and robust to noise and clutter. In order to get good image features, state-of-the-art deep learning methods always include pooling modules, such as max pooling [36], spatial pyramid model [37] and average pooling [38].

For the convolved matrix image features with the size of $p \times k$, for each p-dimensional feature vector v_i, we could define two pooling types, i.e., max pooling and average pooling as shown in formulas 1 and 2

$$f_m(v) = \max v_i \tag{1}$$

$$f_a(v) = \frac{1}{p} \Sigma_{i=1}^{p} v_i \tag{2}$$

After convolutional operation, the distribution of the image features for each patch can be regarded as exponential distribution with mean $E(X) = 1/\lambda$ and variance $D(X) = Var(x) = 1/\lambda^2$. The corresponding cumulative distribution function is $F(x; \lambda) = 1 - e^{(-\lambda x)}$. The high kurtosis in the given exponential distribution could model visual feature response suitably. As when people look at a image, their attentions are mostly drawn by its salient region that is corresponding to the high kurtosis in data distribution.

Let P denote cardinality of the pooling. The cumulative distribution function of max-pooled feature is

$$F(P) = (1 - e^{(-\lambda x)})^P \tag{3}$$

The means separation is

$$\mu_m = (H(P))/\lambda \tag{4}$$

and the variance is

$$\sigma_m^2 = \frac{1}{\lambda^2} \Sigma_{l=1}^{p} \frac{1}{l} (2H(l) - H(P)), \tag{5}$$

where $H(k) = \Sigma_{i=1}^{k} \frac{1}{i}$ denotes the harmonic series.

Thus, for all P,

$$\frac{\mu_1}{\mu_2} = \frac{\delta_1}{\delta_2} = \frac{\lambda_1}{\lambda_2} \tag{6}$$