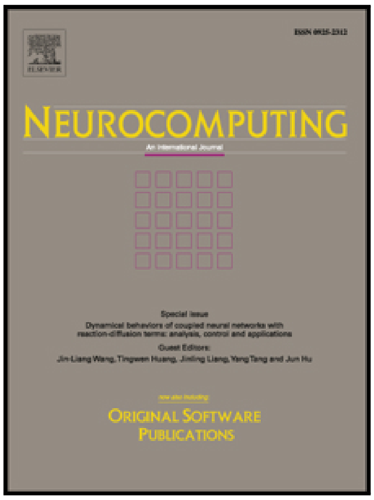
Accepted Manuscript



Content-Based Image Retrieval with Compact Deep Convolutional Features

Ahmad Alzu'bi , Abbes Amira , Naeem Ramzan

PII: S0925-2312(17)30618-5
DOI: 10.1016/j.neucom.2017.03.072
Reference: NEUCOM 18305

To appear in: *Neurocomputing*

Received date: 8 August 2016
Revised date: 15 December 2016
Accepted date: 26 March 2017

Please cite this article as: Ahmad Alzu'bi , Abbes Amira , Naeem Ramzan , Content-Based Image Retrieval with Compact Deep Convolutional Features, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.03.072

Content-Based Image Retrieval with Compact Deep Convolutional Features

Ahmad Alzu'bi[a], Abbes Amira[a,b], Naeem Ramzan[a]

[a]School of Engineering and Computing
University of the West of Scotland, Paisley, PA1 2BE, UK
[b]KINDI CenterforComputingResearch, Qatar University, Qatar
(ahmad.alzubi; abbes.amira; naeem.ramzan)@uws.ac.uk

Abstract. Convolutional neural networks (CNNs) with deep learning have recently achieved a remarkable success with a superior performance in computer vision applications. Most of CNN-based methods extract image features at the last layer using a single CNN architecture with orderless quantization approaches, which limits the utilization of intermediate convolutional layers for identifying image local patterns. As one of the first works in the context of content-based image retrieval (CBIR), this paper proposes a new bilinear CNN-based architecture using two parallel CNNs as feature extractors. The activations of convolutional layers are directly used to extract the image features at various image locations and scales. The network architecture is initialized by deep CNNs sufficiently pre-trained on large generic image dataset then fine-tuned for the CBIR task. Additionally, an efficient bilinear root pooling is proposed and applied to the low-dimensional pooling layer to reduce the dimension of image features to compact but high discriminative image descriptors. Finally, an end-to-end training with backpropagation is performed to fine-tune the final architecture and to learn its parameters for the image retrieval task. The experimental results achieved on three standard benchmarking image datasets demonstrate the outstanding performance of the proposed architecture at extracting and learning complex features for the CBIR task without prior knowledge about the semantic meta-data of images. For instance, using a very compact image vector of 16-length, we achieve retrieval accuracy 95.7%(*m*AP)on Oxford5K and88.6%on Oxford105K; which outperforms the best results reported by state-of-the-art approaches. Additionally, a noticeable reduction is attained in the required extraction time for image features and the memory size required for storage.

Keywords: CBIR**;** Deep learning; Convolutional neural networks; Bilinear compact pooling; Similarity matching

1. INTRODUCTION

In the domain of content-based image retrieval (CBIR), the retrieval accuracy is essentially based on the discrimination quality of the visual features extracted from images or small patches. Image contents (objects or scenes) may include different deformations and variations, e.g. illumination, scaling, noise, viewpoint, etc, which makes retrieving similar images one of the challenging vision tasks. The typical CBIR approaches consist of three essential steps applied on images: detection of interest points, formulation of image vector, and similarity/dissimilarity matching.

In order to extract representative image features, the most existing CBIR approaches use some hand-crafted low-level features, e.g. scale-invariant features transform (SIFT) [1] and speed-up robust features (SURF) [2] descriptors. Such features are usually encoded by general orderless quantization methods such as vector of locally aggregated descriptors (VLAD) [3]. The resulting image representations have shown a high capability on preserving the local patterns of image contents by capturing local characteristics of image objects, e.g. edges and corners. Therefore, they are suitable for the image retrieval task and widely used for matching local patterns of objects. However, convolutional neural networks (CNNs) have recently demonstrated a superior performance over hand-crafted features on image classification [4-6]. Adopting a deep learning procedure on multiple layers of convolutional filters makes CNNs able to subjectively learn even complex representations for many vision and recognition tasks.

Many recent works [5,7,8] demonstrate that the CNN-based generic features adequately trained on sufficient and diverse image datasets, e.g. ImageNet [9], can be successfully applied to other visual recognition tasks. Additionally, performing a proper fine-tuning on CNNs using domain-specific training data can achieve a noticeable performance in common vision tasks [5,10]; including object localization and instance image retrieval. Despite the promising results achieved by CNNs so far, there is no exact understanding or common agreement on how these deep learning architectures work; especially at the intermediate hidden layers. Several successful approaches [11-14] have applied CNNs to extract generic features for image retrieval tasks and obtained promising results. They mainly utilize the power of local features to generate a generic image representation based on some pre-trained CNNs. Nevertheless, many open questions and challenges need more investigation. Foremost, the effectiveness of fine-tuning the CNN models pretrained for specific task, e.g. image classification, on transfer learning to the CBIR task. Secondly, the discrimination quality of image features