Expert Systems With Applications 109 (2018) 35–48

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Reverse image search for scientific data within and beyond the visible spectrum

Flavio H.D. Araujo[a,b,c,d,1], Romuere R.V. Silva[a,b,c,d,1], Fatima N.S. Medeiros[c], Dilworth D. Parkinson[b], Alexander Hexemer[b], Claudia M. Carneiro[e], Daniela M. Ushizima[a,b,*]

[a] University of California, Berkeley, California 94720, USA
[b] Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, California 94720, USA
[c] Federal University of Ceará, Avenida Humberto Montes, s/n, PICI, Fortaleza, Ceará 60455-900, Brazil
[d] Federal University of Piauí, Rua Cícero Duarte, 912, Junco, Picos, Piauí 64607-670, Brazil
[e] Federal University of Ouro Preto, Rua Costa Sena, 171, Ouro Preto, Minas Gerais 35400-000, Brazil

ARTICLE INFO

Article history:
Received 21 December 2017
Revised 2 April 2018
Accepted 11 May 2018

Keywords:
Reverse image search
Content-based image retrieval
Scientific image recommendation
Convolutional neural network

ABSTRACT

The explosion in the rate, quality and diversity of image acquisition instruments has propelled the development of expert systems to organize and query image collections more efficiently. Recommendation systems that handle scientific images are rare, particularly if records lack metadata. This paper introduces new strategies to enable fast searches and image ranking from large pictorial datasets with or without labels. The main contribution is the development of `pyCBIR`, a deep neural network software to search scientific images by content. This tool exploits convolutional layers with locality sensitivity hashing for querying images across domains through a user-friendly interface. Our results report image searches over databases ranging from thousands to millions of samples. We test `pyCBIR` search capabilities using three convNets against four scientific datasets, including samples from cell microscopy, microtomography, atomic diffraction patterns, and materials photographs to demonstrate 95% accurate recommendations in most cases. Furthermore, all scientific data collections are released.

1. Introduction

With the increased availability of large data repositories, a substantial amount of time is spent searching for pictures, which is seldom an efficient procedure. Recent reports (Evans, 2016; GE Digital, 2016) point out that the growth in data size, rates and variety is significant; they also suggest that scientific data will grow twice as quickly as any other sector, yet less than 3% of that data will be tagged in a meaningful way. Several imaging facilities will soon be generating 1 to 50 petabytes of data per year, which poses several challenges: (a) inadequate or insufficient meta-data describing experimental records; (b) the impracticality of manual curation of massive datasets; and (c) the lack of tools adapted to the new data acquisition modes.

Photo organizers that rely on curated data have improved to include operations such as sorting and categorization by dates or media types, metadata and other user annotations. However, manual insertion of metadata is seldom achievable at scale, and even impossible in some scenarios, such as with high-throughput imaging instruments. There is a critical need to create automated and accurate methods to organize, query and retrieve unlabeled images, since "everyone searches all the time" (Eckstein, 2011). Building upon recent works on deep learning (Araque, Corcuera-Platas, Sánchez-Rada, & Iglesias, 2017; Gonçalves, Guilherme, & Pedronette, 2018; Goodfellow, Bengio, & Courville, 2016) and the ability to create annotations for unlabeled datasets from curated ones (Ushizima et al., 2016), new expert systems promise to change the human experience from hardly relevant retrievals to broadly useful (over 85% accuracy) results. As an example, Google Photos has provided automated and custom-labeling features since 2015, so that users can quickly organize large collections (JR Raphael, 2015), with high retrieval accuracies for face detection. **One of the main challenges in image recognition is to perform tasks that are easy for humans to do intuitively, but hard to describe formally** (Goodfellow et al., 2016). **For exam-**

* Corresponding author.
E-mail addresses: flavio86@ufpi.edu.br (F.H.D. Araujo), romuere@ufpi.edu.br (R.R.V. Silva), fsombra@ufc.br (F.N.S. Medeiros), DYParkinson@lbl.gov (D.D. Parkinson), ahexemer@lbl.gov (A. Hexemer), carneirocm@gmail.com (C.M. Carneiro), dushizima@lbl.gov (D.M. Ushizima).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.eswa.2018.05.015

ple, domain scientists (Donatelli et al., 2015), **who are visually trained to identify complex patterns from their experimental data, but many times are unable to describe mathematically the primitives that construct the motif.**

By using methods to identify patterns from pictures, recommendation systems, also known as reverse image search tools, represent an excellent opportunity for data reduction, in which the imaging acquisition, data collection and storage strategies are integrated and tailored toward a desired pattern. Such systems could support scientists in adjusting experimental parameters fast enough for optimal data collection, **combating a major problem at imaging facilities where overwhelming amounts of irrelevant data are collected daily.**

The main contributions of this work are as follows: **(a) Development of a new recommendation system for visual image search with an inferential engine that learns compact signatures for recovering images within massive datasets; it uses approximate ranking, and includes 10 schemes to measure distance based on different sets of features from labeled and/or unlabeled images; (b) Deployment of pyCBIR as an interactive system to enable a generic user to search for image collections from diverse domains through an intuitive graphical user interface in Qt©as illustrated in** Fig. 6; **(c) Evaluation of three Convolutional Neural Networks (CNN) implementations available within pyCBIR, describing scientific problems that rely upon deep and complex networks, but also shallow architectures. Because pyCBIR allows generalization of processing pipelines, it is quickly extensible to different training datasets; here we test data collections from four science domains, and report on accuracy and time consumption given different architectural choices; (d) Establishment of reproducible work, containing both codes for benchmarks and tests using image repositories publicly available, and software based on open-source tools.**[2]

First, we discuss previous work on CBIR in the context of the proposed expert system pyCBIR in Section 2, **including capabilities and benefits of pyCBIR to a generic user.** Section 3 **explains the four data collections of images across domains.** Section 4 **presents detailed information on computational methods tested in** Section 5, **which focuses on results of applying different learning strategies to the different datasets. Finally,** Section 6 **evaluates the impact of pyCBIR, including perspectives on future requirements.**

2. Related work

The term content-based image retrieval (CBIR) was introduced in 1992 by Kato (Hirata & Kato, 1992; Kato, 1992), and it has been associated with systems that provide image matching and retrieval for queries performed by visual example. A quarter of century later, most image retrieval systems available for scientific image search still rely on keyword-based image retrieval (van den Broek, van Rikxoort, & Schouten, 2005), although most of the image collections generated by humans lack proper annotations (Bethel, Greenwald, & Nowell, 2015). Efforts to optimize image search employing CBIR systems (Khatami et al., 2018; Tzelepi & Tefas, 2018; Yu, Yang, Yao, Sun, & Xu, 2017) exploit computer vision and machine learning algorithms to represent images in terms of more compact primitives. Given an image as an input query, instead of keywords or metadata, such an approach allows matching samples by similarity.

Since 2003, the Bag-of-Words (BoW) model has been predominantly viewed as the state-of-the-art in CBIR, augmented by 13 years of Scale Invariant Feature Transform (SIFT) methods (Zheng, Yang, & Tian, 2017). **In addition, CBIRs combining textural descriptors, such as Gabor and Fourier** (Hannan, Arebey, Begum, Basri, & Mamun, 2016; Shrivastava & Tyagi, 2015), **GLCM** (Hannan et al., 2016), **continue to be broadly used to organize natural images, and made use of similarity search based on the Euclidean distance**. Tsochatzidis et al. (Tsochatzidis et al., 2017) **extend previous ideas and include exploration of Histogram of Oriented Gradients (HOG) descriptors combined with the Earth Mover's Distance to recover mammograms based on similarity. Moreover, image retrieval methods have advanced in two main directions: SIFT-based and CNN-based, with promising improved accuracy when combining CNN and SIFT** features (Zheng et al., 2017). **To the best of our knowledge, there are no software tools that allow combining both strategies for CBIR tasks.** Several free engines for CBIR are thriving at natural and biomedical image organization, and e-commerce tasks (Shamoi, Inoue, & Kawanaka, 2015; Yu et al., 2016), but underlying codes for end-to-end workflow remain closed, and are seldom generalizable to other scientific image collections.

2.1. CNN And visual recognition

These are significant strides toward automating image catalogs, and motivates our efforts to construct convolutional neural networks (CNN) using Google TensorFlow to organize scientific data. TensorFlow (Abadi et al., 2015) is an open-source software library for Machine Intelligence that presents advantages regarding flexibility, portability, performance, and compatibility to GPU. In order to deliver high-performance C++ code, TensorFlow uses the Eigen linear algebra library in addition to CUDA numerical libraries, such as cuDNN to accelerate core computations and scale to large datasets.

A typical CNN pipeline is shown in Fig. 1, consisting of three main "neural" layers: convolutional layers, pooling layers, and fully connected layers. This algorithm requires two stages for training the network: (a) a forward stage, which represents the input image in each layer and outputs a prediction used to compute the loss cost based on the curated data (labeled samples), and (b) a backward stage, which computes the gradients of layer parameters to drive the cost function to very low values (Goodfellow et al., 2016; Guo et al., 2016).

By exploring CNN algorithms that automatically extract features at multiple levels of abstraction from large datasets, CBIR systems can benefit from complex non-linear functions that map unprocessed input data to the results, bypassing human-designed characterization reliant on domain knowledge. Wan et al. (Wan et al., 2014) investigated different deep learning frameworks for CBIR when applied to natural images, such as the ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge) dataset. That paper reported mean average precision of 0.4711 using a massive image collection with 10,000,000 hand-labeled images depicting 10,000+ object categories as training.

2.2. Expert systems for material recognition

Apart from natural scenes, recent works on recognizing material categories from images (Bell, Upchurch, Snavely, & Bala, 2014; Liu, Sharan, Adelson, & Rosenholtz, 2010; Sharan, Rosenholtz, & Adelson, 2014; Zhang, Ozay, Liu, & Okatani, 2015) include experiments using the Flickr Material Dataset (FMD) and/or the Materials in Context Database (MINC). Sharan et al. (Sharan et al., 2014) explored low and mid-level features, such as color, SIFT, HOG, combined with an augmented Latent Dirichlet Allocation model under a Bayesian generative perspective, achieving 44.6% accurate recognition rate on FMD. Using a CNN-based feature extraction mecha-

[2] Source codes/data to be published upon paper acceptance at camera.lbl.gov