Applied Soft Computing 63 (2018) 197–205

Contents lists available at ScienceDirect

Applied Soft Computing

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

Parallel deep solutions for image retrieval from imbalanced medical imaging archives

Amin Khatami[a,*], Morteza Babaie[b,c], Abbas Khosravi[a], H.R. Tizhoosh[c], Saeid Nahavandi[a]

[a] Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia
[b] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran
[c] KIMIA Lab at the University of Waterloo, Canada

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form
15 November 2017
Accepted 17 November 2017
Available online 26 November 2017

Keywords:
Content-based image retrieval
CBIR
Medical imaging
Deep learning
LBP
HOG
Radon

ABSTRACT

Learning and extracting representative features along with similarity measurements in high dimensional feature spaces is a critical task. Moreover, the problem of how to bridge the semantic gap, between the low-level information captured by a machine learning model and the high-level one interpreted by a human operator, is still a practical challenge, especially in medicine. In medical applications, retrieving similar images from archives of past cases can be immensely beneficial in diagnostic imaging. However, large and balanced datasets may not be available for many reasons. Exploring the ways of using deep networks, for classification to retrieval, to fill this semantic gap was a key question for this research. In this work, we propose a parallel deep solution approach based on convolutional neural networks followed by a local search using LBP, HOG and Radon features. The IRMA dataset, from ImageCLEF initiative, containing 14,400 X-ray images, is employed to validate the proposed scheme. With a total IRMA error of 165.55, the performance of our scheme surpasses the dictionary approach and many other learning methods applied on the same dataset.

1. Introduction

1.1. Background

In medical image analysis, searching for similar images (in terms of similar anatomy) can serve as "virtual peer review" for diagnostic purposes. Retrieving similar images (along with associated reports and other metadata) from the archive can establish a new level of comparative diagnosis. This is absent at the present time. Properly addressing this issue can immensely contribute to more accurate image-based diagnosis [67,70,5]. This highlights the importance of model generalisation in medical applications. The role of computerized mechanisms to aid radiologists in diagnosis is becoming important because of the massive growth of medical data. To make content-based search feasible, a robust retrieval system which can find similar cases in a large archive is required. This challenging task results in expeditiously extending the domain of content-based medical image retrieval (CBIR) [3,29,15,48,65]. CBIR deals with searching for similar images in large archives when the search query is an image and not a textual description. Hence, CBIR solutions generally operate based on some notion of pixel similarity search. However, as two-dimensional data, images cannot be easily compared with each other; rotation, scale, translation and illumination variability would hinder simple one-to-one comparisons. Many studies have been conducted on CBIR systems [70,69,54,17,20]. Based on a wealth of research reports, feature embedding and binary features have proven to be a much more reliable and efficient approaches to image search [43,64,25,10].

There seem to be two distinct trends in terms of retrieval in the CBIR literature. One class of algorithms attempts to retrieve specific organs in specific modalities such as retrieving malignant lung nodules [44] and liver lesions in CT Images [39], and chest structures from X-ray images [52]. The second class of CBIR framework focuses on global similarity search in heterogeneous PACS-like archives to categorize and retrieve similar images [15,3]. The latter is followed in this study, with the use of deep learning models to reduce the search space. Note that using deep learning for CBIR tasks in medical imaging has two main challenges: (1) generally, there are not many ready-to-use "labelled" medical image data sets large enough for training deep learning solutions. Therefore, augmentation becomes a crucial part of pre-processing to avoid overfitting;

* Corresponding author.
E-mail addresses: amin.khatami@deakin.edu.au (A. Khatami), mbabaie@uwaterloo.ca (M. Babaie), abbas.khosravi@deakin.edu.au (A. Khosravi), tizhoosh@uwaterloo.ca (H.R. Tizhoosh), saeid.nahavandi@deakin.edu.au (S. Nahavandi).

https://doi.org/10.1016/j.asoc.2017.11.024
1568-4946/

(2) medical image datasets may suffer from the "imbalance" problem for different reasons, e.g., diverse incidence rates of different malignancies. This is in contrast to non-medical cases where generally perfectly balanced data sets can be assembled, e.g., for face recognition [45,28]. Indeed, a balanced data set would be quite rare in medical domain. As a result, researchers exploit the capabilities of deep learning solutions to address these two challenges.

CBIR systems generally work based on classification or comparison of binary or real-valued "tags" attached to each image. While many methods can be used for tag generation, one can employ discriminative or nearest-neighbour methods for the retrieval process. Image texture descriptors [62,37] are widely used in medical CBIR domain. As shown in [21,50], it seems that keypoint-based descriptors such as scale-invariant feature transform (SIFT), speeded up robust features (SURF), and oriented fast and rotated brief (ORB) are not able to generate reliable feature points for some types of medical images. However, dense sampling methods such as local binary pattern (LBP) and histogram of oriented gradients (HOG) appear to be more efficient [46].

Local Binary Patterns (LBPs) as local descriptors and texture histogram, and MPEG-7 edge histogram as a global histogram were successfully used in ImageCLEFmed 2007 [56,4]. In the image retrieval competition of imageCLEF 2012, LBP combined with other features such as global SIFT and GIST, were the best among various types of feature detectors [35]. Unay et al. [60] compared LBPs and Kanade–Lucas–Tomasi feature points, and showed that LBP-based retrieval dominated the CBIR research for MR brain images. On the other hand, combinations of LBP and HOG were applied to several successful studies, such as [68] in the object recognition domain, while [63] this combination for human detection [53].

Global features are also widely used in medical image retrieval [29]. One of the recently proposed ideas for global descriptors is "Radon barcodes" [57] whereas binary vectors are extracted for the entire image (not for local neighbourhoods). A small number of Radon projections are generated, i.e., equidistant ones, and subsequently threshold to construct a *barcode*. These descriptors can serve as a first stage of retrieval for certain class of images (e.g., medical images with negligible global rotation and scale variations).

Deep representations of digital images through artificial neural networks with many hidden layers have proven to be a very successful method for learning the content of an image for purposes such as face and object recognition [13,28,27,26]. Among different architectures, convolutional neural networks, short CNNs, have been distinctly dominant in accurately learning the image structures and *embed* them in their hidden layers, most commonly in *fully connected layers* (FC layers) [13], especially in medical imaging [14].

1.2. Related work

Image retrieval in medical application (IRMA) dataset [32,31], described in experimental results section, is a well-known X-ray image dataset used for classification [15,23,24] and retrieval [16,35]. IRMA images have been put together with "semantic gap" in mind, the disparity between human perception of images and their similarity on one side, and the quantitative image representation by computer algorithms on the other side [40,70]. Visual inspection of cases in IRMA dataset can show that the embedded IRMA code (for benchmarking) does in fact reflect the semantic gap, an attribute that makes this dataset quite interesting. However, IRMA classes are heavily imbalanced, which creates a barrier for deep learning methods which generally expect large and balanced classes.

Camlica et al. [8] reported an IRMA error (Eq. (6), Section 3.2.1) of 146.55, which is the lowest reported error so far. However, their saliency method is extremely sluggish. They neglect the overhead for saliency calculations and have employed offline-generated maps for use during testing. Of course, this is not practical.

Avni et al. [2] reported an IRMA error of 169.5 by applying a dictionary approach on the IRMA data set. More specifically, they proposed a multi-resolution patch-based dictionary approach by utilizing principle component analysis (PCA) on the densely sampled patches. This was followed by a support vector machine (SVM) classifier, training on the bag-of-words.

As reported in [36], an IRMA error of 178.93 was obtained by Idiap research team. They utilised different classification approaches for SVMs, coupling two different image descriptors, i.e. LBP and modSIFT [59].

Regarding deep networks, Liu et al. [33] utilized CNN codes (the features in the last FC layer) for a local search procedure, created by LBP and Radon transform to achieve an IRMA error of 224.13. Sze et al. [55] achieved an IRMA error of 344.08 by using deep autoencoders and Radon barcodes. In another research conducted by Sharma et al. [51], KNN search was utilized to extract the features from stacked autoencoders and achieved an IRMA error of 376.

Therefore, to close the semantic gap, an ensemble of parallel deep learning solutions is investigated in this study by using diverse inputs. The proposed structure can deal with imbalanced classes in a robust way than a regular deep neural network, which generally expects nicely balanced classes of a large number of instances. The ensemble-based technique generalises learning model in order to achieve a robust detection system.

1.3. Motivations and contributions

A significant achievement of deep learning models in classification and recognition tasks motivates this research to investigate the relevant models in medical image retrieval applications. Several challenges exist in this research: (1) there are only a few studies in this area, which warrant a thorough analysis as conducted in this research, (2) the lack of access to a large-scale benchmark data set in medicine is inevitable, (3) even by accessing a large medical data set, the imbalanced problem pertaining to the data distribution could exist. Therefore, how a deep learning model could be designed to address the above-mentioned problems is important. In this research, methods to utilise deep neural networks to bridge the semantic gap between a machine learning model and a human operator are proposed. To achieve this aim, different network representations are trained for different resolutions of the inputs which have been effectively augmented to help compensate the effects of the imbalanced data distribution problem.

In this paper, we propose a generic scheme for parallel deep networks that are differently trained. A proper combination of the networks results in a shrunken search space which enables a robust local similarity-based search phase. In other words, the retrieval results of multiple networks are then subject to refined search using the LBP, the HOG, and Radon transforms features. We apply our solution on IRMA 2009 image dataset with 14,400 X-ray images, a dataset that is both rather small (in a deep-learning context) and extremely imbalanced. We compare the achieved performance with other results reported in the literature.

The main contributions of proposing a deep-based parallel solution are three-fold: (1) introduce a robust retrieval system on a strongly imbalanced medical benchmark data set which not only are efficient, but also outperform the best accuracy and performance in the literature. (2) To the best of the author's knowledge, this is the first research work which creates a feature vector on an ensemble-based shrunk search space. This contribution considerably improves the performance. (3) Propose a shrinking search space, using an ensemble model with three convolutional networks, followed by three well-known practicable transformations