Medical Image Analysis 35 (2017) 303–312

Contents lists available at ScienceDirect

Medical Image Analysis

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

Large scale deep learning for computer aided detection of mammographic lesions

Thijs Kooi[a,*], Geert Litjens[a], Bram van Ginneken[a], Albert Gubern-Mérida[a], Clara I. Sánchez[a], Ritse Mann[a], Ard den Heeten[b], Nico Karssemeijer[a]

[a] Diagnostic Image Analysis Group, Department of Radiology, Radboud University Medical Center, Nijmegen, The Netherlands
[b] Department of Radiology, University Medical Centre Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 11 February 2016
Revised 12 July 2016
Accepted 20 July 2016
Available online 2 August 2016

Keywords:
Computer aided detection
Mammography
Deep learning
Machine learning
Breast cancer
Convolutional neural networks

ABSTRACT

Recent advances in machine learning yielded new techniques to train deep neural networks, which resulted in highly successful applications in many pattern recognition tasks such as object detection and speech recognition. In this paper we provide a head-to-head comparison between a state-of-the art in mammography CAD system, relying on a manually designed feature set and a Convolutional Neural Network (CNN), aiming for a system that can ultimately read mammograms independently. Both systems are trained on a large data set of around 45,000 images and results show the CNN outperforms the traditional CAD system at low sensitivity and performs comparable at high sensitivity. We subsequently investigate to what extent features such as location and patient information and commonly used manual features can still complement the network and see improvements at high specificity over the CNN especially with location and context features, which contain information not available to the CNN. Additionally, a reader study was performed, where the network was compared to certified screening radiologists on a patch level and we found no significant difference between the network and the readers.

1. Introduction

Nearly 40 million mammographic exams are performed in the US alone on a yearly basis, arising predominantly from screening programs implemented to detect breast cancer at an early stage, which has been shown to increase chances of survival (Tabar et al., 2003; Broeders et al., 2012). Similar programs have been implemented in many western countries. All this data has to be inspected for signs of cancer by one or more experienced readers which is a time consuming, costly and most importantly error prone endeavor. Striving for optimal health care, Computer Aided Detection and Diagnosis (CAD) (Giger et al., 2001; Doi, 2007; 2005; van Ginneken et al., 2011) systems are being developed and are currently widely employed as a second reader (Rao et al., 2010; Malich et al., 2006), with numbers from the US going up to 70% of all screening studies in hospital facilities and 85% in private institutions (Rao et al., 2010). Computers do not suffer from drops in concentration, are consistent when presented with the same input data and can potentially be trained with an incredible amount of training samples, vastly more than any radiologist will experience in his lifetime.

Until recently, the effectiveness of CAD systems and many other pattern recognition applications depended on meticulously handcrafted features, topped off with a learning algorithm to map it to a decision variable. Radiologists are often consulted in the process of feature design and features such as the contrast of the lesion, spiculation patterns and the sharpness of the border are used, in the case of mammography. These feature transformations provide a platform to instill task-specific, a-priori knowledge, but cause a large bias towards how we humans think the task is performed. Since the inception of Artificial Intelligence (AI) as a scientific discipline, research has seen a shift from rule-based, problem specific solutions to increasingly generic, problem agnostic methods based on learning, of which *deep learning* (Bengio, 2009; Bengio et al., 2013; Schmidhuber, 2015; LeCun et al., 2015) is its most recent manifestation. Directly distilling information from training samples, rather than the domain expert, deep learning allows us to optimally exploit the ever increasing amounts of data and reduce human bias. For many pattern recognition tasks, this has proven to be successful to such an extent that systems are now reaching human or even superhuman performance (Cireşan et al., 2012; Mnih et al., 2015; He et al., 2015).

* Corresponding author.
E-mail address: thijs.kooi@radboudumc.nl, email@thijskooi.com (T. Kooi).

http://dx.doi.org/10.1016/j.media.2016.07.007
1361-8415/

The term *deep* typically refers to the layered non-linearities in the learning systems, which enables the model to represent a function with far less parameters and facilitates more efficient learning (Bengio et al., 2007; Bengio, 2009). These models are not new and work has been done since the late seventies (Fukushima, 1980; Lecun et al., 1998). In 2006, however, two papers (Hinton et al., 2006; Bengio et al., 2007) showing deep networks can be trained in a greedy, layer-wise fashion sparked new interest in the topic. Restricted Boltzmann Machines (RBM's), probabilistic generative models, and autoencoders (AE), one layer neural networks, were shown to be expedient pattern recognizers when stacked to form Deep Belief Networks (DBN) (Hinton et al., 2006; Bengio et al., 2007) and Stacked Autoencoders, respectively. Currently, fully supervised, Convolutional Neural Networks (CNN) dominate the leader boards (Krizhevsky et al., 2012; Zeiler and Fergus, 2014; Simonyan and Zisserman, 2014; Ioffe and Szegedy, 2015; He et al., 2015). Their performance increase with respect to the previous decades can largely be attributed to more efficient training methods, advances in hardware such as the employment of many core computing (Cireşan et al., 2011) and most importantly, sheer amounts of annotated training data (Russakovsky et al., 2014).

To the best of our knowledge, Sahiner et al. (1996) were the first to attempt a CNN setup for mammography. Instead of raw images, texture maps were fed to a simple network with two hidden layers, producing two and three feature images respectively. The method gave acceptable, but not spectacular results. Many things have changed since this publication, however, not only with regard to statistical learning, but also in the context of acquisition techniques. Screen Film Mammography (SFM) has made way for Digital Mammography (DM), enabling higher quality, raw images in which pixel values have a well-defined physical meaning and easier spread of large amounts of training data. Given the advances in learning and data, we feel a revisit of CNNs for mammography is more than worthy of exploration.

Work on CAD for mammography (Elter and Horsch, 2009; Nishikawa, 2007; Astley and Gilbert, 2004) has been done since the early nineties but unfortunately, progress has mostly stagnated in the past decade. Methods are being developed on small data sets (Mudigonda et al., 2000; Zheng et al., 2010) which are not always shared and algorithms are difficult to compare (Elter and Horsch, 2009). Breast cancer has two main manifestations in mammography, firstly the presence of malignant soft tissue or masses and secondly the presence of microcalcifications (Cheng and Huang, 2003) and separate systems are being developed for each. Microcalcifications are often small and can easily be missed by oversight. Some studies suggest CAD for microcalcifications is highly effective in reducing oversight (Malich et al., 2006) with acceptable numbers of false positives. However, the merit of CAD for masses is less clear, with research suggesting human errors do not stem from oversight but rather misinterpretation (Malich et al., 2006). Some studies show no increase in sensitivity or specificity with CAD (Taylor et al., 2005) for masses or even a decreased specificity without an improvement in detection rate or characterization of invasive cancers (Fenton et al., 2011; Lehman et al., 2015). We therefore feel motivated to improve upon the state-of-the art.

In previous work in our group (Hupse et al., 2013) we showed that a sophisticated CAD system taking into account not only local information, but also context, symmetry and the relation between the two views of the same breast can operate at the performance of a resident radiologist and of a certified radiologist at high specificity. In a different study (Karssemeijer et al., 2004) it was shown that when combining the judgment of up to twelve radiologists, reading performance improved, providing a lower bound on the maximum amount of information in the medium and suggesting ample room for improvement of the current system.

In this paper, we provide a head-to-head comparison between a CNN and a CAD system relying on an exhaustive set of manually designed features and show the CNN outperforms a state-of-the-art mammography CAD system, trained on a large dataset of around 45,000 images. We will focus on the detection of solid, malignant lesions including architectural distortions, treating benign abnormalities such as cysts or fibroadenomae as false positives. The goal of this paper is *not* to give an optimally concise set of features, but to use a complete set where all descriptors commonly applied in mammography are represented and provide a fair comparison with the deep learning method. As mentioned by Szegedy et al. (2014), success in the past two years in the context of object recognition can in part be attributed to judiciously combining CNNs with classical computational vision techniques. In this spirit, we employ a candidate detector to obtain a set of suspicious locations, which are subjected to further scrutiny, either by the classical system or the CNN. We subsequently investigate to what extent the CNN is still complementary to traditional descriptors by combining the learned representation with features such as location, contrast and patient information, part of which are not explicitly represented in the patch fed to the network. Lastly, a reader study is performed, where we compare the scores of the CNN to experienced radiologists on a patch level.

The rest of this paper is organized as follows. In the next section, we will give details regarding the candidate detection system, shared by both methods. In Section 3, the CNN will be introduced followed by a description of the reference system in Section 4. In Section 5, we will describe the experiments performed and present results, followed by a discussion in Section 6 and conclusion in Section 7.

2. Candidate detection

Before gathering evidence, every pixel is a possible center of a lesion. This approach yields few positives and an overwhelming amount of predominantly obvious negatives. The actual difficult examples could be assumed to be outliers and generalized away, hindering training. Sliding window methods, previously popular in image analysis are recently losing ground in favor of candidate detection (Hosang et al., 2015) such as selective search (Uijlings et al., 2013) to reduce the search space (Girshick et al., 2014; Szegedy et al., 2014). We therefore follow a two-stage classification procedure where in the first stage, candidates are detected and subjected to further scrutiny in a second stage, similar to the pipeline described in Hupse et al. (2013). Rather than class agnostic and potentially less accurate candidate detection methods, we use an algorithm designed for mammographic lesions (Karssemeijer and te Brake, 1996). It operates by extracting five features based on first and second order Gaussian kernels, two designed to spot the center of a focal mass and two looking for spiculation patterns, characteristic of malignant lesions. A final feature indicates the size of optimal response in scale-space.

To generate the pixel based training set, we extracted positive samples from a disk of constant size inside each annotated malignant lesion in the training set, to sample the same amount from every lesion size and prevent bias for larger areas. To obtain normal pixels for training, we randomly sampled 1 in 300 pixels from normal tissue in normal images, resulting in approximately 130 negative samples per normal image. The resulting samples were used to train a random forest (Breiman, 2001) (RF) classifier. RFs can be parallelized easily and are therefore fast to train, are less susceptible to overfitting and easily adjustable for class-imbalance and therefore suitable for this task.

To obtain lesion candidates, the RF is applied to all pixel locations in each image, both in the train and test set, generating a likelihood image, where each pixel indicates the estimated suspi-