Convolutional neural networks for mammography mass lesion classification

John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira and Miguel Angel Guevara Lopez

Abstract— **Feature extraction is a fundamental step when mammography image analysis is addressed using learning based approaches. Traditionally, problem dependent handcrafted features are used to represent the content of images. An alternative approach successfully applied in other domains is the use of neural networks to automatically discover good features. This work presents an evaluation of convolutional neural networks to learn features for mammography mass lesions before feeding them to a classification stage. Experimental results showed that this approach is a suitable strategy outperforming the state-of-the-art representation from 79.9% to 86% in terms of area under the ROC curve.**

I. Introduction

Breast cancer is the most common form of cancer in the world in women population with over 1.5 million predicted diagnoses in 2010 and causing more than half a million deaths per year [1]. Breast cancer has a known asymptomatic phase that can be detected with mammography, and therefore, mammography is the primary imaging modality for screening. Double-reading (two radiologists independently read the same mammograms) has been advocated to reduce the proportion of missed cancers and it is currently included in most of the screening programs [2]. However, double-reading incurs in additional workload and costs. Alternatively, computer-aided diagnosis (CADx) systems may assist a single radiologist reading mammograms providing support to her/his decisions. Such systems aim at classifying lesions identified by the radiologist.

CADx systems typically rely on machine learning classifiers (MLC) to provide diagnosis. In order to train a MLC for breast cancer diagnosis, a set of features describing the image is required. Ideally, features should have high discriminant power that allows inferring if a given image is from a malignant finding or not. This is, however, a challenging topic that has gathered the focus of researchers from different fields, from medicine to computer vision. Thus, several types of features may be used for inferring the diagnosis. Many CADx systems uses handcrafted features based on prior knowledge and expert guidance. As an alternative, the use of machine learning strategies to learn good features directly from the data is a new paradigm that has shown successful results in the computer vision area. Such paradigm is called Deep learning.

Deep learning methods have been widely applied in late years to address several computer perception tasks [3]. Their main advantage relies on avoiding the design of specific feature detectors. On the contrary, deep learning models automatically learns rerpresentative features directly from the data. Deep learning has had remarkable results, particularly in computer vision problems such as natural scene classification and object detection [3]. Deep learning models also has been adapted to different medical tasks such as tissue classification in histology and histopathology images [4], alzheimer disease diagnosis and sclerosis lesion segmentation among others [5].

However, only few works have explored deep learning methods to address the automatic mammography image analysis task [1]. In [6] stacked deep auto-encoders were used to estimate breast density score and to segment breast tissue using multiscale features and convolutional neural network (CNN) models. In [7] CNNs are used to characterize microcalcifications as a representation strategy. The closest work developed to learn the representation in order to classify malign/benign breast lesions was done in [8], which uses an adaptive deconvolutional network. In contrast, our work presents an evaluation of convolutional architectures to learn the image representation. Its main difference with respect to previous works relays on the way features are learned, since we use supervised information during CNN training. This approach takes advantage of expert knowledge represented in the lesion annotation manually made by expert radiologists.

The rest of the paper is organized as follows: Section II describes the proposed approach to perform automatic mammography image analysis. Section III details the experimental setup used to evaluated the proposed approach. Finally, Sections IV and V shows results and discusses the main conclusions of this work.

II. Material and methods

A. Breast cancer digital repository

The benchmarking dataset used in this study is part of the breast cancer digital repository (BCDR)[1]. BCDR is a wide-ranging annotated public repository composed of Breast Cancer patients' cases of the northern region of Portugal. BCDR provides normal and annotated patients cases of breast cancer including mammography lesions outlines, anomalies observed by radiologists, pre-computed image-based descriptors as well as related clinical data. In this work, a new "Film Mammography-based dataset" was used (BCDR-F03). The BCDR-F03, extracted from BCDR, is composed of 344 patients with 736 film images containing 426 benign mass lesions and 310 malign mass lesions, including clinical data and image-based descriptors. Figure 1 shows examples of of lesions (benign and malignant), with their respective segmentations.

[1] http://bcdr.inegi.up.pt, accessed on February 17, 2015

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

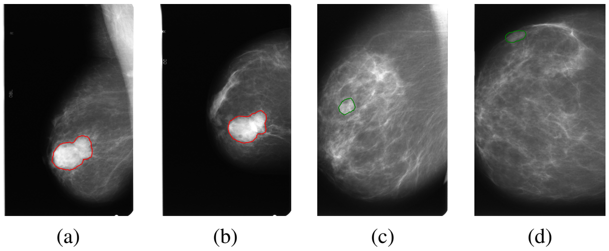

Fig. 1: Sample images from the dataset. Malignant lesion in (a) oblique and (b) craneo-caudal view. Benign lesion in (c) oblique and (d) craneo-caudal view.

Type	Features
Intensities	Mean, median, maximum, minimum, standard deviation, skewness, kurtosis
Shape	Area, perimeter, circularity, elongation, y_center_mass, x_center_mass, form
Textures	Contrast, correlation, entropy

TABLE I: Handcrafted features set. For details see [2]

1) Baseline descriptors: Based on the systematic evaluation presented by [2], histogram of gradients (HOG) descriptor and histogram of gradient divergence (HGD) were selected as our baseline since it showed the best performance against other traditional descriptors. Additionally, a set of 17 handcrafted features extracted from segmented lesions related to texture and intensities of the images are used for comparative purposes.

a) Handcrafted features (HCfeats): A two-step procedure was used for selecting handcrafted features: (1) the feature selection method proposed by Perez et al. [9] was applied to the original set of features [2], in order to ranking features; and (2) a set of 17 texture, shape and intensities features with high performance to describe masses were heuristically selected. Table I groups features according to their type. It is composed by intensity descriptors computed directly from the grey-levels of the pixels inside the lesion's contour identified by the radiologists; texture descriptors computed from the Grey-level co-occurrence matrix related to the bounding box of lesion's contour identified by the radiologists; and shape descriptors computed from the lesion's contour identified by the radiologists. Notice that this set of features requires not only the region of interest (ROI) detection, but also the manual segmentation provided by the expert.

b) Histogram of oriented gradients (HOG): HOG describes images through the distribution of the gradient. Images are divided into a grid of blocks (e.g. 3×3), and each block is described by a histogram of the orientation of the gradient. Each histogram has a predefined number of bins dividing the range of possible orientations (from 0 to 2π radians, or from 0 to π radians), and the value of each bin is calculated by summing the magnitude of the gradient of the pixels which have gradient direction within the limits of the bin.

c) Histogram of gradient divergence (HGD): Gradient divergence in a point i, j is measured as the angle between the vector of the intensity gradient on i, j and a vector pointing to the center of the image with origin in i, j. HGD describes images through the distribution of the gradient divergence [2]. Images are divided into concentric regions, and each region is described by a histogram of the gradient divergence.

B. Proposed method

Image representation is fundamental for automatic analysis of mammography images. Its goal is to describe the content of the image in a compact and discriminative way. Conventional CADx systems for mammography image analysis represent images with a carefully selection of a set of mathematical and heuristic features aiming to describe the lesion. Recent works have replaced this hand-engineering process by a learning-based approach where a model is trained in an unsupervised way using deep learning to transform the raw pixels in a set of features that feeds a classifier algorithm [8], [6]. In contrast to previous work, herein we applied a hybrid approach in which CNNs are used to learn the representation in a supervised way. That is, we used the annotations to guide the feature learning process.

Our approach comprises three main stages: preprocessing, feature learning and classification training. Preprocessing, detailed in section II-B.1, aims to prepare the data using a set of transformations such that next stage takes advantage of relevant characteristics. Feature learning, detailed in section II-B.2, is performed by training a CNN with annotated samples. The final stage, detailed in section II-B.3, is the SVM classifier training using the penultimate layer of the CNN as features.

1) Preprocessing: Preprocessing is a common stage in CADx systems. Its main goal is to enhance the characteristics of the image by applying a set of transformations that could help to improve performance in next stages. The first step in this work is to extract the ROI from the image. Secondly, an oversampling strategy is used to both, artificially get more samples and help to prevent overfitting during training. Finally, a normalization process is carried out to prepare data for learning algorithms. Such steps are detailed below:

a) Cropping: For convenience, we fixed the input size to $r \times r$ pixels (e.g. 150×150). As BCDR dataset provides manual segmentation, images were cropped to the bounding box of the segmented lesion and rescaled to $r \times r$ pixels preserving the aspect ratio only when necessary, otherwise the lesion is centered preserving surrounding region.

b) Data augmentation: For each training image, we have artificially generated 7 new label-preserving samples using combination of flipping and $90, 180$ and 270 degrees rotation transformations. In lesion classification problem, data augmentation makes sense because a lesion can be presented in any particular orientation. Thus, the model also should be able to learn from such transformations.

c) Normalization: Due to the digitalization process, the lighting conditions between different film images would be different affecting all pixel values of the image. A common way to overcome this effect is to perform a global contrast

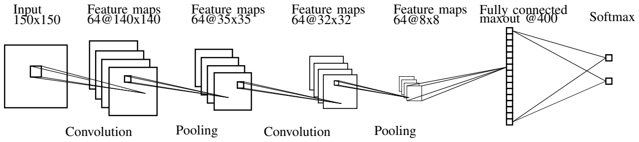

Fig. 2: Best CNN evaluated on mass classification

normalization substracting the mean of the intensities in the image from each pixel. Local contrast normalization (LCN) was also performed. Its main idea is to mimic the behavior the V1 visual cortex by performing normalization in small patches over the entire image [10]. It is widely known that feature learning and deep learning methods usually perform better when the input data has been decorrelated and normalized, mainly because such properties help gradient-based optimization techniques to converge [10]. Notice that both, global and contrast normalizations are performed in an image-wise fashion, thus it is not required to store parameters in the training procedure.

2) Supervised feature learning: CNN is a neural network that shares connections between hidden units yielding to low computational time and translational invariances properties. CNNs have been successfully applied in shape recognition problems [11] as well as in medical diagnosis that involved texture as a discriminant feature [4]. Because mass characterization is highly correlated with shape and texture features [1], [2], a CNN model becomes a suitable strategy for mass lesion classification.

a) Architecture: A CNN layer comprises 3 main components: convolutional kernels, activation function and pooling function. Convolutional kernels are a set of several small squares that works as filters over the input image. The result of this convolution is the input for the activation function which includes non-linearities in an element-wise fashion. Finally, the pooling function aggregates contiguous values with an aggregation function such as $\max()$. To improve the capability of the model, these 3 components are stacked iteratively so that the output of one component is the input for the next one, building a deep neural network with two or more layers.

The proposed architecture, depicted in figure 2, has 11×11 local kernels in the first layer followed by a 5×5 pooling shape with stride of 4×4 pixels. The second layer has 4×4 local kernels with 4×4 pooling shape without overlapping. The third fully connected layer has 400 maxout units, while the final softmax layer has 2 output units used for classification.

b) Regularization: Usually, classifier models with large numbers of parameters tend to overfit training data, hindering their capability to generalize to unseen data. Neural networks are no exception and require different strategies to control this behavior. In this work dropout and max-norm regularization were used. Dropout [12] randomly set to 0 the input of a unit, while max-norm regularization forces the norm of each vector of incoming weights in a unit to a maximum value.

c) Optimization: The best proposed architecture has around 4.6×10^6 parameters. Large models training has to scale in both memory requirements and computational time. The strategy used in this work to train the CNN is stochastic gradient descent with momentum. Early stopping strategy monitoring the area under the ROC curve (AUC) on validation set was chosen as stop criteria. The implementation of the proposed method was carried out with the Pylearn2 framework which makes an efficient and intensive use of GPU computing cores [13]. A massive exploration was conducted using the CETA-CIEMAT[2] Research Center infrastructure. Bigger models that requires more intensive computation were carried out using a NVidia Tesla K40 GPGPU card.

3) Classification: Following the previous work [8], [2], Linear SVM was selected as classification strategy. To evaluate the CNN as a representation strategy, image regions are propagated through the network until the penultimate layer and their activations are used as representation. This stage can be seen as a fine-tuning process of the last layer, where a smaller model is adjusted.

III. Experimental setup

The original BCDR-F03 dataset was divided in training (60%) and test dataset (40%), following a stratified sampling per patient, that is, we make sure that all mammograms of each patient belong to only one of the two subsets. This setup guarantees that the model is not tested using seen patients during training stage.

In the preprocessing stage, the size of the cropped region was fixed to $r = 150$ according to the distribution of the lesions size and computational capability; and the kernel size for LCN was $k = 11$ pixels. Following previous results [2], 5×5 and 3×3 blocks sizes for HOG and 4 and 8 regions for HGD were explored. 8 and 16 bins were evaluated for both histograms.

The CNN parameter exploration was performed by taking out one sixth of the training set as validation set, training 25 models with random hyperparameter initialization and choose the best according to validation performance. It has been reported that this strategy is preferable over grid search when training deep models [14]. Before training the SVM model, a zero-mean unit-variance normalization process is carried out. Training set was used for finetunning the C parameter for the SVM classifier by performing 20 bootstrap runs without replacement. Final performance is reported in terms of area under the ROC curve (AUC) in the test set.

IV. Results

Table II shows summary of these experiments. The HCfeats set, that uses segmentation information, performs slightly better than HOG-based descriptors. This confirms the importance of shape information for mass characterization. Interestingly CNN models, that uses only the raw pixels,

[2] http://www.ceta-ciemat.es/, accessed on February 17, 2015

	Learned			**Baseline**		
Features	**CNN3**	DeCAF	CNN2	HOG	HGD	HCfeats
AUC	**0.860**	0.836	0.821	0.796	0.793	0.799

TABLE II: Summary of results in terms of AUC in test set.

outperform the state-of-the-art features [2]. It is also noteworthy that adding more layers to the model improves the representation capability yielding to the best results.

For comparative purposes, we included the evaluation of DeCAF [15], a pretrained model with the Imagenet dataset [3]. DeCAF is a model with a larger complexity, in comparison with all other evaluated representations. Thus, it is expected that performs better than using handcrafted features. However, a smaller CNN model trained with the images of the domain performs the best. This behavior yields to the two main conclusions of this work: On the one hand, CNN models outperforms state-of-the-art representations for automatic mammography image analysis. On the other hand, such automatic analysis is a problem with its own particularities, thus is not enough to learn the representation using a large CNN model, but also the learning process should be guided by a training set with a wide visual variability such that shows to the model textural and shape features presented in mass lesions.

V. Conclusions

This work proposes a framework to address classification of mass lesions in mammography film images. Instead of design particular features to explain the content of mammography images, this approach learns them directly from data in a supervised way. The proposed CNN architecture takes the raw pixels of the image as input, to learn a set of nonlinear transformations that represents better the image data. Our approach outperformed state-of-the-art methods, HOG and HGD descriptors [2], from 79.9% to 86.0% in terms of area under the ROC curve (AUC). Interestingly, this model also outperforms a set of handcrafted features that takes advantage of additional information given by the segmentation of the radiologist. Our future work includes the evaluation of bigger architectures as well as the inclusion of clinical information expecting to enhance the representation.

VI. Acknowledgements

This work was mainly supported by the Cloud Thinking project (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, "Mais Centro" program (Portugal). Also, it was partially supported by project "Diseño e implementación de un sistema de cómputo sobre recursos heterogéneos para la identificación de estructuras atmosféricas en predicción climatológica" number 1225-569-34920 through Colciencias contract number 0213-2013 and the computing facilities of Extremadura Research Center for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain. Arevalo also thanks Colciencias for its support through a doctoral grant in call 617/2013. The authors also thank for K40 Tesla GPU donated by NVIDIA and which was used for some feature learning experiments.

References

[1] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical Imaging*, vol. 37, no. 3, pp. 420 – 426, 2013.

[2] D. C. Moura and M. A. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 561–574, 2013.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, pp. 1798–1828, Aug 2013.

[4] J. Arevalo, A. Cruz-Roa, and F. A. González, "Hybrid image representation learning model with invariant features for basal cell carcinoma detection," 2013.

[5] G. Wu, D. Zhang, and L. Zhou, eds., *Machine Learning in Medical Imaging*. Springer International Publishing, 2014.

[6] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm, "Breast Tissue Segmentation and Mammographic Risk Scoring Using Deep Learning," in *Breast Imaging* (H. Fujita, T. Hara, and C. Muramatsu, eds.), vol. 8539 of *Lecture Notes in Computer Science*, pp. 88–94, Springer International Publishing, 2014.

[7] X.-S. Zhang, "A new approach for clustered MCs classification with sparse features learning and TWSVM.," *The Scientific World Journal*, vol. 2014, p. 970287, Jan. 2014.

[8] A. R. Jamieson, K. Drukker, and M. L. Giger, "Breast image feature learning with adaptive deconvolutional networks," 2012.

[9] N. P. Pérez, M. A. G. López, A. Silva, and I. Ramos, "Improving the mann–whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography," *Artificial Intelligence in Medicine*, no. 0, pp. –, 2014.

[10] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2146–2153, Sept 2009.

[11] Q. Ke and Y. Li, "Is rotation a nuisance in shape recognition?," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 4146–4153, June 2014.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[13] I. J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio, "Pylearn2: a machine learning research library," *arXiv preprint arXiv:1308.4214*, 2013.

[14] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[15] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[3] http://www.image-net.org/, accessed on February 17, 2015