Breast Density Classification with Convolutional Neural Networks

Pablo Fonseca[1(✉)], Benjamin Castañeda[2], Ricardo Valenzuela[3], and Jacques Wainer[1]

[1] RECOD Lab, Institute of Computing, University of Campinas, Campinas, Brazil
pablo.arroyo@ic.unicamp.br

[2] Laboratorio de Imágenes Médicas, Pontificia Universidad Católica del Perú, Lima, Peru

[3] Imaging Lab, Eldorado Research Institute, Campinas, Brazil

Abstract. Breast Density Classification is a problem in Medical Imaging domain that aims to assign an American College of Radiology's BIRADS category (I-IV) to a mammogram as an indication of tissue density. This is performed by radiologists in an qualitative way, and thus subject to variations from one physician to the other. In machine learning terms it is a 4-ordered-classes classification task with highly unbalance training data, as classes are not equally distributed among populations, even with variations among ethnicities. Deep Learning techniques in general became the state-of-the-art for many imaging classification tasks, however, dependent on the availability of large datasets. This is not often the case for Medical Imaging, and thus we explore Transfer Learning and Dataset Augmentationn. Results show a very high squared weighted kappa score of 0.81 (0.95 C.I. [0.77,0.85]) which is high in comparison to the 8 medical doctors that participated in the dataset labeling 0.82 (0.95 CI [0.77, 0.87]).

1 Introduction

Breast cancer is a major health treat as it accounts for the 13.7% of cancer deaths in women according to the World Cancer Report [1]. Moreover, it is the second most common type of cancer worldwide and recent statistics show that one in every ten women will develop it at some point of their lives. However, it is important to notice that when detected at an early stage, the prognosis is good, opening the door to Computer Aided Diagnosis Systems that target the prevention of this disease. Medical research towards the prevention of breast cancer has shown that breast parenchymal density is a strong indicator of cancer risk [2]. Specifically, the risk of developing breast cancer is increased only in 5% related to mutations in the genetic biomarkers BRCA 1 and 2; this risk, on the other hand, is increased to 30% for breast densities higher than 50% [3,4]. Because

B. Castañeda—Supported by grant DGI 2012-0141 from the Pontificia Universidad Católica del Perú.

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 101–108, 2017.
DOI: 10.1007/978-3-319-52277-7_13

of this, the breast density can be seen as very valuable information in order to perform preventive tasks and assess cancer risk. However, this behavior varies from one ethnicity to the other, even with different breast density distributions across populations. A comparative study of our dataset used in this research with other populations can be found in Casado et al. [5].

1.1 Breast Density

According to Otsuka et al. [6], mammographic density refers to the proportion of radiodense fibroglandular tissue relative to the area or volume of the breast. In order to assess breast composition, there are both qualitative and quantitative methods. One of the best known qualitative methods is the Breast Imaging Reporting and Data System (BI-RADS), the target of this research [7]. Among quantitative methods, there is one developed by Boyd, the quantitative ACR and other computer-assisted methods. Some previous work of automatic breast composition classification include Oliver et al. [8] where several methods are tested on the MIAS database using the BIRADS standard and Oliver et al. [9] where a method that included segmentation, extraction of morphological and texture features and bayesian classifier combination. Also there are commercial tools such as Volpara (TM) [10] and Quantra (TM) [11].

1.2 American College of Radiology: BIRADS Categories

The American College of Radiologists (ACR) developed four qualitative categories for breast density which are presented below along with the meaning of each composition category and sample mammograms in the study. The main goal of this research is to classify mammograms into the BIRADS categories using convolutional neural networks based techniques. There are two approaches that were explored: Random Filter Convolutional networks as feature extractors coupled with a Linear SVM and the well known Krizhevsky Deep Convolutional Network.

These four categories are qualitatively judged by radiologists according to their density. When it comes to assessing breast density in mammograms some challenges might arise even for experienced radiologists such as reported in the comparative study of inter- and intra-observer agreement among radiologists [12].

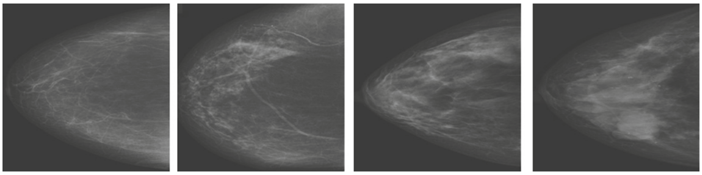
1.3 Radiologist Agreement and Performance Evaluation

Rater agreement need to be measured in a way that it is consistent with the actual performance of radiologists as well as automated algorithms. Accuracy would not be very informative as kappa indexes for this kind of task. Cohen's kappa can target agreement on both categorical and ordinal variables. In Rendondo et al. [12] the performance of radiologists is evaluated for a stratified sample of 100 mammogram in the BIRADS categories which include two separated target measuares: assesment and breast density. For breast density, the study found high interobserver and intraobserver agreement ($\kappa = 0.73$, with a 95% confidence interval 0.72–0.74 and $\kappa = 0.82$, with a 95% confidence interval 0.80–0.84, respectively) where the squared weighted kappa is used. In that study, there were 21 radiologists with average experience in reading mammograms of 12 years (range 4–22). Even if according to the Landis and Koch's strenght of agreement [13] the intraobserver kappa value shows an almost perfect agreement, the value 0.82 can show the subjectivity of the BIRADs measure, even for trained professionals.

2 The Breast Density Dataset

In this paper, we work with the dataset presented by Casado et al. [5], here we briefly survey its idiosyncrasies, but for more details please refer to that paper. The mammograms were obtained from two medical centers in Lima, Peru. Some of those images are shown in Fig. 1. All subjects were women who underwent routine breast cancer screening. The age of the subjects ranged from 31 to 86 years (mean age of 56.7 years, standard deviation of 9.5 years). A total of 1060 subjects were included in the sample population. The mammograms were collected in a craniocaudal view using two different systems. The first one was a Selenia Dimensions (Hologic, Bedford, MA) which produced digital mammograms with a pixel pitch of 100 μm, The second system was a Mammomat 3000 (Siemens Medical, Iselin, NJ) in combination with a CR 35 digitizer (Agfa Healthcare, Mortsel, Belgium) that allowed producing digital images with a depth of 16 bits and a pixel pithch of 50 μm. Approximately 16% and 84% of the mammograms used in this study were acquired with the Selenia Dimensions and Mammomat 3000 systems, respectively.

The mammograms were blindly classified by eight radiologists with varying degrees of experience assessing mammograms between 5 and 25 years. The mode of the breast density classification by the eight radiologists was considered as ground truth for this study. These medical doctors are named from A to H in no particular order. To serve the purposes of this research the Region of Interest (ROI) was manually selected to include only the breast region. The cropped



(a) BIRADS I (b) BIRADS II (c) BIRADS III (d) BIRADS IV

Fig. 1. Sample Mammograms in the study (BIRADS I-IV). Here 4 mammograms are shown along with their BIRADS density classification: the mammogram in figure (a) is the less dense and the mammogram in figure (d) is the most dense. Images were resized to an 1:1 ratio for feature extraction.

Table 1. Distribution of densities across dataset

BIRADS	I	II	III	IV
Total dataset (1060 images)	32.64% 346 images	45.19% 479 images	19.62% 208 images	2.55% 27 images
Training subset (744 images)	32.66% 243 images	45.16% 336 images	19.62% 146 images	2.55% 19 images
Test subset (316 images)	32.59% 103 images	45.25% 143 images	19.62% 62 images	2.53% 8 images

images were also resized to a fixed size of 200 × 200 pixels, it must be noticed that radiologists had the full resolution DICOM mammogram. The semantics of the problem shows that finding regions of high radio dense tissue is meaningful only with respect of the total area of the breast. Initial experiments show that a better behavior of CNNs was found when the resize was performed. The general machine learning setup was to do a stratified division of the dataset in training and test sets as shown in Table 1.

3 Methods

We explore two methods based on convolutional neural networks that might deal with small datasets as ours. First, the HT-L3 architecture in Fig. 2 in Sect. 3.1 and previously studied for Breast Density Classification in [14] and initially proposed by [15]. In this first case, the filters are generated randomly, so a major focus is quantifying how stable are under different initializations. The second method in Sect. 3.2 is the well known Krizhevsky network [16] also identified as AlexNet. The network was trained on the ImageNet dataset an fine tuned for Breast Density Classification.

3.1 HT-L3 Visual Features with Random Filters

The HT-L3 family of features described by Cox and Pinto [17] and Pinto et al. [15] can be seen as a parameterizable image description function that was inspired on the primates visual system. It performs three consecutive layers of filtering, activation, pooling and normalization leading to a high dimensional representation of the image which can be used to feed a classifier such as the Support Vector Machines (SVM) with linear kernel.

Random noise filters were used in the filtering stage, although they were normalized to have zero mean and unit standard deviation no further training was performed on them. It is the aim of some of the experiments carried on to characterize the stability of such setup. Grid search was used to find the most suitable architectures. This means that several candidate architectures of the HT-L3 family are screened in order to chose the top performing ones. More

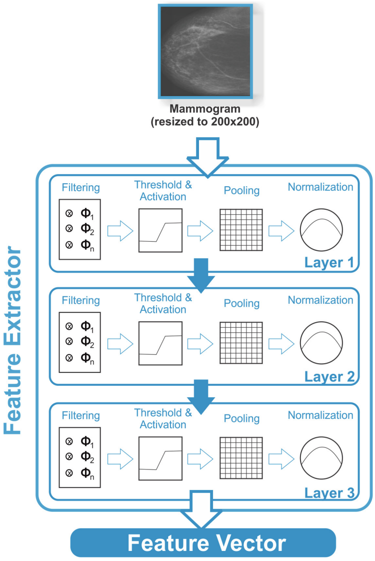


Fig. 2. HT-L3 Convolutional Neural Network: Here the linear and non-linear operations are shown in the order in which they are performed. Each layer (1–3) has the same operations, and the output of one layer is the input for the next one. However, each layer will produce more deep multiband images.

details on the implementation of this architecture for Breast Density Classification were reported in [14], a previous work of our group. From these results we chose the top three performing architectures which are presented in Table 2.

Table 2. Top performing architectures: here the top 3 performing architectures are shown alongside they parameters. S. L1, 2, 3 defines the filter size of the layer, # L1, 2, 3 defines the number of filters per layer, exponent ans α are fixed to 2.

	S. L1	S. L2	S. L3	# L1	# L2	# L3	Exp.	α
Arch1	5×5	5×5	5×5	128	128	64	2	2
Arch2	5×5	7×7	7×7	128	32	128	2	2
Arch3	5×5	5×5	7×7	32	32	64	2	2

3.2 Deep Learning Trained with the Image.net and Fine-Tuned with the Breast Density Dataset

In their 2012 paper, Krizhevsky et al. [16] presented a convolutional neural network architecture that got the best results on the Image.net dataset, a natural images dataset. Using the Caffe framework for Deep Learning [18], a convolutional network as described by [16] trained with the Image.net 1k classes subset was fine-tuned with the breast density dataset.

Changes to the net included the last layer being changed for a 4-way softmax in order to perform BIRADS classifications. The same sets of training and testing in the evaluating the HT-L3 features were used. Data aumengtation was used generating random crops of the mammogram.

4 Results

The HT-L3 features with random filters worked well and they are quite stable on multiple runs (we initialized 30 times with different seeds each architecture), which implied generating random filters with different seeds can indicate that a great deal of the discriminative power of these features lay on the architecture, leaving a classifier such as the SVM with linear kernel the task of finding separations across different classes, given the feature representations.

On the other side, in the Krizhevsky network the learning is performed with back-propagation and stochastic gradient descent, which means that the filters are actually learned as well as the weights of the fully connected layers that produce the final classification. As the size of the dataset is small for Deep Learning standards (1k images). On Table 3 we report the behavior of four automatic classification approaches, the first one is the Krizhevsky network trained with Image.net database and then fine-tuned with our breast density database and the three top performing HT-L3 architectures.

Table 3. Agreement with the golden standard on the test set

	κ-squared
Radiologist A	0.72
Radiologist B	0.88
Radiologist C	0.89
Radiologist D	0.74
Radiologist E	0.89
Radiologist F	0.88
Radiologist G	0.80
Radiologist H	0.88
Radiologists (95% CI)	[0.77, 0.87]
Krizhevsky NET (95% CI)	[0.77, 0.85]
HTL3 Arch1 (95% CI - 30 runs)	[0.75, 0.77]
HTL3 Arch2 (95% CI - 30 runs)	[0.72, 0.74]
HTL3 Arch3 (95% CI - 30 runs)	[0.71, 0.73]

5 Conclusions

Deep learning techniques for image classification aim an end-to-end learning procedure, specially in contrast to hand-engineered features as a previous approach presented in [19]. Convolutional neural networks, on the other hand, use a series of well-defined operations however parameterizable that both obtain a representation of the image and then perform the classification. Two techniques were explored, a three layer convolutional network with random filters used to produce high dimensional features later used to train a Linear SVM classifier and the Krizhevsky Network trained with the Image.net database and then fine tuned with our Breast Density Dataset. The main difference between these two approaches is that in the first one, no learning was performed by the convolutional neural network and thus the filter were randomly generated and the classification was actually performed by a linear SVM.

We found satisfactory results as measure by the kappa score, as a measure of agreement between the golden standard and the proposed methods. We would like to stress the point that even if results show a very high squared-weighted kappa score of 0.81 (0.95 C.I. [0.77,0.85]) of our best performing approach which is high in comparison to the 8 medical doctors that participated in the dataset labeling 0.82 (0.95 CI [0.77, 0.87]), radiologist-like behavior might not be fully well measured with this agreement metric.

We also been successfully in proving the stability of classification when random filters are used, this might indicate that a great deal of the discriminative power of these features lay on the model assumptions, at least for the problem targeted here. However, this HT-L3 based approaches did not performed better than the Krizhevsky Network.

Future work should include exploration of a better metric for assessing radiologist-like performance and further Deep Learning Architectures as well as Open Mammogram Databases.

References

1. Boyle, P., Levin, B., et al.: World Cancer Report 2008. IARC Press, International Agency for Research on Cancer (2008)
2. Wolfe, J.N.: Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer **37**(5), 2486–2492 (1976)
3. Boyd, N., Martin, L., Yaffe, M., Minkin, S.: Mammographic density and breast cancer risk: current understanding and future prospects. Breast Cancer Res. **13**(6), 1 (2011)
4. Ursin, G., Qureshi, S.A.: Mammographic density-a useful biomarker for breast cancer risk in epidemiologic studies. Nor. Epidemiol. **19**(1), 59–68 (2009)
5. Casado, F.L., Manrique, S., Guerrero, J., Pinto, J., Ferrer, J., Castañeda, B.: Characterization of Breast Density in Women from Lima, Peru (2015)
6. Otsuka, M., Harkness, E.F., Chen, X., Moschidis, E., Bydder, M., Gadde, S., Lim, Y.Y., Maxwell, A.J., Evans, G.D., Howell, A., Stavrinos, P., Wilson, M., Astley, S.M.: Local Mammographic Density as a Predictor of Breast Cancer (2015)

7. D'Orsi, C.J.: Breast Imaging Reporting and Data System (BI-RADS). American College of Radiology (1998)
8. Oliver, A., Freixenet, J., Martí, R., Zwiggelaar, R.: A comparison of breast tissue classification techniques. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 872–879. Springer, Heidelberg (2006). doi:10.1007/11866763_107
9. Oliver, A., Freixenet, J., Marti, R., Pont, J., Perez, E., Denton, E., Zwiggelaar, R.: A novel breast tissue density classification methodology. IEEE Trans. Inf. Technol. Biomed. **12**(1), 55–65 (2008)
10. Volpara Solutions (2015). http://volparasolutions.com
11. Hologic (2015). www.hologic.com/
12. Redondo, A., Comas, M., Maciá, F., Ferrer, F., Murta-Nascimento, C., Maristany, M.T., Molins, E., Sala, M., Castells, X.: Inter-and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br. J. Radiol. **85**(1019), 1465–1470 (2012). PMID: 22993385
13. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
14. Fonseca, P., Mendoza, J., Wainer, J., Ferrer, J., Pinto, J., Guerrero, J., Castañeda, B.: Automatic Breast Density Classification using a Convolutional Neural Network Architecture Search Procedure (2015)
15. Pinto, N., Doukhan, D., DiCarlo, J.J., Cox, D.D.: A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput. Biol. **5**(11), e1000579 (2009)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
17. Cox, D., Pinto, N.: Beyond simple features: a large-scale feature search approach to unconstrained face recognition. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 8–15 (2011)
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093 (2014)
19. Angulo, A., Ferrer, J., Pinto, J., Lavarello, R., Guerrero, J., Castañeda, B.: Experimental Assessment of an Automatic Breast Density Classification Algorithm Based on Principal Component Analysis Applied to Histogram Data (2015)