Deep Learning for Automatic Detection of Abnormal Findings in Breast Mammography

Ayelet Akselrod-Ballin(✉), Leonid. Karlinsky, Alon Hazan,
Ran Bakalo, Ami Ben Horesh, Yoel Shoshan, and Ella Barkan

IBM Research, Haifa, Israel
ayeletb@il.ibm.com

Abstract. Automatic identification of abnormalities is a key problem in medical imaging. While the majority of previous work in mammography has focused on classification of abnormalities rather than detection and localization, here we introduce a novel deep learning method for detection of masses and calcifications. The power of this approach comes from generating an ensemble of individual Faster-RCNN models each trained for a specific set of abnormal clinical categories, together with extending a modified two stage Faster-RCNN scheme to a three stage cascade. The third stage being an additional classifier working directly on the image pixels with the handful of sub-windows generated by the first two stages. The performance of the algorithm is evaluated on the INBreast benchmark and on a large internal multi-center dataset. Quantitative results compete well with state of the art in terms of accuracy. Computationally the methods runs significantly faster than current state-of-the art techniques.

1 Introduction

Numerous studies have shown that early detection of breast cancer can both increase treatment options and reduce mortality [1]. There are two main types of abnormal objects detected in mammograms (MG) – **Masses** are seen as compact areas that appear brighter than the embedding tissue and **Calcifications** [2]. **Macro calcifications (macro)** are usually benign bigger bits of calcium, appearing as coarse calcium deposits in the breast, such as Coarse or "popcorn-like", Dystrophic or Rim-like (eggshell). **Micro calcifications (micro)** are tiny specks of mineral deposits that can be distributed in various ways, clusters, specific patterns or scattered. Certain features and presentations of micro calcifications, specifically, amorphous, pleomorphic shapes and clustered distribution can be associated with malignant breast cancer. However, detection and identification are extremely difficult, due to the subtle fine-grained visual categories and large variability of appearance of the different categories. Therefore, automatic detection, localization and quantification of abnormal findings, has the potential of separating the normal-negative from positive exams in the screening

A. Akselrod-Ballin and L. Karlinsky contributed equally to this work.

M.J. Cardoso et al. (Eds.): DLMIA/ML-CDS 2017, LNCS 10553, pp. 321–329, 2017.
DOI: 10.1007/978-3-319-67558-9_37

processes, allowing the radiologist to focus on the challenging cases and avoiding unnecessary breast biopsies.

Deep learning has shown remarkable performances in recognition tasks such as detection and classification [3–6]. Below we briefly review the dominant detection algorithms in the field, emphasizing the differences compared to our approach.

R-CNN and its variants use external region proposals instead of sliding windows to find objects in images. Commonly, classical methods are used to generate those region proposals. In R-CNN [5], Selective Search [7] is used to generate candidate object boxes, a CNN [3] is used to generate the feature vector for each box, an SVM is trained to classify the box feature vectors, and linear regression followed by non-maximal suppression is used to adjusts the bounding boxes and eliminate duplicate detections. Each stage of this complex pipeline must be precisely independently tuned and the resulting system is slow, taking more than 40 seconds per test on a GPU [8].

Other CNN based detectors, such as Deep MultiBox [9], Fast [8] and Faster [6] R-CNN, and YOLO [10], focus on speeding up the R-CNN framework by sharing computation and using Region Proposal Networks (RPN) – a CNN that effectively performs sliding window searches in a fully convolutional manner [11] over a grid of receptive fields, to generate object proposals instead of Selective Search [6–8]. YOLO [10] also puts spatial constraints on the grid cell proposals. While they offer speed and accuracy improvements over R-CNN, they still fall short of real-time performance, running in about 2–3 FPS similar to Faster-RCNN.

Advances in deep learning methods have recently been exploited successfully in the medical imaging domain [12, 13] and specifically in breast mammography (MG) for mass and calcification classification [14, 15]. Our study departs from the majority of previous work as it focuses on detection and localization of abnormalities rather than on binary classification of calcification or masses. An exception is the prominent work on mass detection of [16], which is based on a multiscale deep belief network classifier and is followed by a cascade of R-CNN and random forest classifiers. A recent study by the same authors [17] is currently considered the most effective study for detection of micro calcification, utilizing a cascade of boosting classifiers with shape and appearance features. Based on these approaches [18] recently presented a deep residual neural network aimed at classification of an MG as benign or malignant. The final segmentation maps of the masses and micro-calcifications are not evaluated, yet they demonstrate INBreast [19] separation into normal and benign with an ROC AUC of 0.8.

An important component of the Faster-RCNN framework differing it from similar detection frameworks like YOLO [10], is the classifier running after the RPN built on top of an ROI-pooling layer [6] that pools features from the top most layer of the base network shared between RPN and the classifier. This can be seen as a two stage cascade with RPN being the first stage and the classifier second. However, the representation generated by ROI-pooling is not specifically optimized for classification (being shared between the detector and the classifier). Therefore this work follows the Faster-RCNN architecture presented in [20] yet proposes to extend the scheme to a three stage cascade, the third stage being an additional classifier working directly on the image pixels with the handful of sub-windows generated by the first two stages.

Our contribution is three fold: First, to the best of our knowledge there are no studies utilizing a unified deep learning approach to combine both detection, localization and classification of multiple type of masses and calcifications. Second we utilize a three- stage cascade integrating a two-stage ensemble of RPN and Faster-RCNN models and a third CNN classifier and demonstrate its effectiveness for reducing the number of FP detections. Finally, our methodology competes well with state of the art in terms of inference time and accuracy even on a large dataset.

2 Methods

Problem Formulation: Given as input a set of training images, bounding boxes corresponding to abnormal findings $\{I_i, y_{1i}, x_{1i}, y_{2i}, x_{2i.}, c_i\}$ and a testing set of images, we seek to detect the abnormal findings in the test set and locate the bounding boxes with a corresponding confidence score.

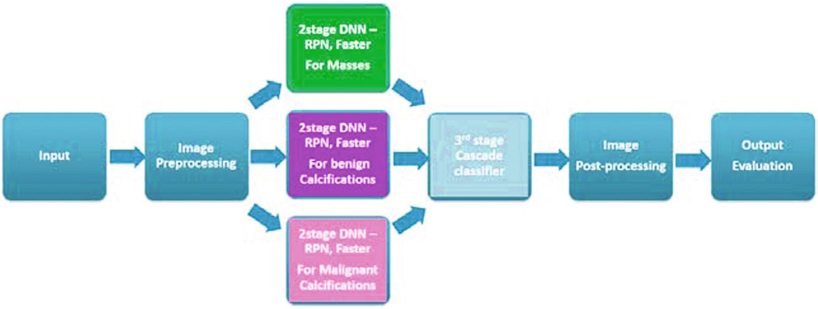
Input
Image Preprocessing
2stage DNN – RPN, Faster For Masses
2stage DNN – RPN, Faster For benign Calcifications
2stage DNN – RPN, Faster For Malignant Calcifications
3rd stage Cascade classifier
Image Post-processing
Output Evaluation

Fig. 1. The deep framework for detection and classification of abnormalities in Mammography.

The system preprocesses the input MG image by dividing the ~4 k × 3 k pixels onto an overlapping grid, utilized to train the three-stage cascade deep neural network (DNN) described below such that each subpart (grid cell) extracts candidate bounding boxes and predicts confidence scores for those boxes. Then, the image post-processing performs grid composition of all the DNN parts results. Figure 1 outlines the system architecture which is composed of three main components detailed below. Roughly speaking, the stages are 'initial detection', 'classification' and 'classification refinement'. In setting these stages we extend the Faster RCNN paradigm, by continuing the 're-classification' loop by adding one more classification iteration, and admitting the detected crops directly to the final classifier (rather than letting it look on them through the prism of the ROI pooling layer and a set of features optimized for detection).

(1) **A region proposal network (RPN):** a deep fully convolutional network that is trained to detect windows on the input image that are likely to contain objects of interest regardless of their class. The RPN simultaneously predicts objects bounds and objectness scores at each position on a wide range of scales and aspect ratios, following which top scoring 500 predictions are kept [11]. The way the network effectively operates on the image in a sliding window fashion, yet a lot of internal

computation are being inherently re-used due to local and hierarchical nature of the stacked network layers. The sliding window operates on a regular grid with the same step size in x and y equal to the final stride of the top most layer receptive field (32 pixels in our case). For each grid location 9 seed boxes (sampled with several different aspect ratios and sizes) are being classified. During training, these boxes, originally called anchors in [6], are being associated with ground truth object boxes according to IoU scores, where the most similar ones (according to IoU) are being labeled as object anchors as well as all those that pass a certain threshold (0.5) in IoU. For each anchor classified as being (part of, or containing) the object, a linear regression is employed to refine the anchor bounding box to the candidate object one.

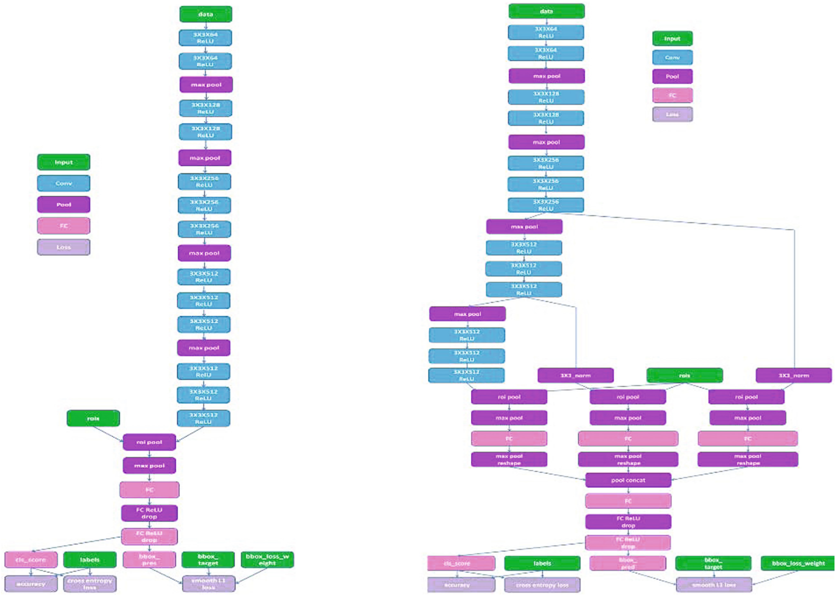

Fig. 2. Original Faster-RCNN architecture (left) Modified Faster-RCNN architecture (right)

(2) **A Fast R-CNN detection network:** that is trained to classify candidate object windows returned by the RPN, each window is classified into one of the classes of interest or rejected as a false alarm. Both RPN and Fast-RCNN share a large common set of bottom layers allowing computation re-use and hence a significant speed-up at test time. For both RPN and the classification network we employ a modified version of VGGNet by [4]. Originally trained on the ImageNet dataset [1], and fine-tuned by us on the task at hand. The system was trained on a single TitanX GPU with 12 GB on chip memory, and i7 Intel CPU with 64 GB RAM. Training times required ~36 h, while testing takes 0.2 seconds per image. During training 2000 top-scoring boxes are sampled from the RPN, during testing top scoring 500 boxes are sampled using standard non-maximal suppression

(NMS) based on box overlap. We used the SGD solver with learning rate of 0.001, batch size 2, momentum 0.99, and 60 epochs. Figure 2 shows the modified Faster-RCNN architecture we followed (see details in [20]).

(3) **Third stage Cascade Classifier:** The true positive (TP) and false positive (FP) boxes candidates of the DNN are computed on the training set and selected to be the positive and negative samples for the next training phase respectively. These boxes are then cropped and used to train a VGG-16 classifier optimized to separate between the TP and the hard negatives of the last step (the classifier) of Faster-RCNN. During training the boxes are randomized and randomly re-cropped and augmented with random geometric and photometric transformations in order to enrich the training and improve generalization. The key idea behind adding this third stage to Faster-RCNN two stage cascade is to allow a separately optimized network to have an alternative look on the ROIs whose original Faster-RCNN internal representation is constrained to be shared between the first two stages (the detector and the classifier). The full training process details include: each batch contained four images chosen at random; out of each image up to three positive boxes (true detection of the previous stage) and one negative box (false detection of the previous stage) were chosen at random; each box was randomly rotated up to 30°, randomly re-cropped by ±25% of original size, randomly flipped, mean subtracted, and finally resized to 224 × 224 size as required by the standard VGG-16 classifier. During testing the original detected boxes from the previous stage were processed for speed (no augmentation, single crop).

3 Experiments and Results

We evaluated the algorithm for detection of masses and calcifications. The data set consists of (1) the publicly available INBreast dataset [19] (2) an Internal dataset consisting of approximately 3500 images, collected from several different medical centers with ground truth annotation by experts including 750, 360, 2400 images with masses, malignant calcifications, and benign calcifications respectively. The mass category includes Breast Imaging-Reporting and Data System (BI-RADS) subtypes of 2, 3, 4, 5 namely benign and malignant masses. The dataset was split into training 80% and testing 20% so that all test patients image were excluded from the training set.

Figure 3, reports the free response operating characteristic curve (FROC) calculating the number of true positive rate (TPR) as a function of false positives per image (FPI). The figure shows the results obtained by each class-model separately on one type of images, and also results of joining of all models on all images including normal images. We compare joining the ensemble by ‘naïve’ concatenation of results obtained by all models or by utilizing the 3rd cascade step on the ensemble of models. The results show clear reduction of FP’s and performance improvement by the cascade approach.

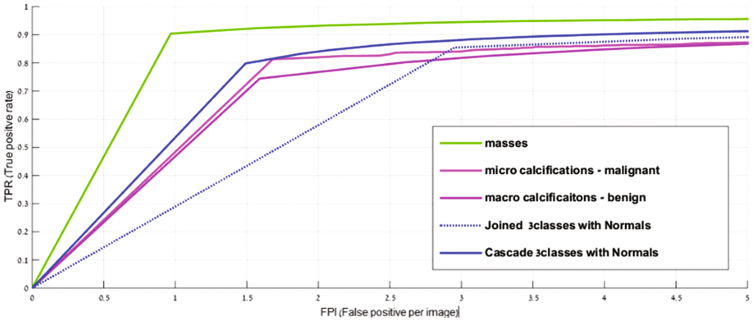

Fig. 3. Detection performance per class and entire set. FROC curves showing the result on various operating points with true positive rate (TPR) versus false positive per image (FPI) for mass images (green), malignant micro calcification images (magenta) and benign macro calcifications images (purple), for 'naïve' joining of all model results on all images (dotted blue), and for joining all models results on all images based on 3rd cascade (bold blue). (Color figure online)

Table 1 summarizes our results compared to the best state-of-the-art results (see details in [16, 17]). [17] Reported the leading results for detection of calcifications on the INBreast (INB) dataset, having a TPR for individual calcification of 40% at one FPI and a TPR of 80% at 10 FPI. The authors noted that these results are significantly better than the current state of the art, which has a TPR of less than 0.01@1 FPI and a TPR of 0.1@10 FPI. [16] Provided a detailed table for mass detection with the best TPR 0.96 ± 0.03@1.2 FPI and TPR = 0.87 ± 0.14@0.8 FPI for INBreast. The papers also obtained the best results in respect to running time of 20 s.

The INB results were obtained by testing separately on INB calcification and mass images with the models trained by our internal data. This is not optimal as there are significant differences in the images characteristics between the two sets. Accordingly, removal of small calcifications (radius < 10 pixels) yields a significant improvement in

Table 1. Presents our results compared to best state-of-art algorithms

Calcifications results	TPR@FPI	#images	Runtime
Ours on micro & macro calcifications	All: TPR 0.4@1 FPI No small: TPR 0.85@1.5 FPI	310/410 INB	5 s
Ours on macro benign calcifications	TPR 0.8@1.5 FPI TPR 0.52@1 FPI TPR 0.9@10 FPI	2400 internal	5 s
Ours on micro malignant calcifications	TPR 0.81@1.7 FPI TPR 0.48@1 FPI	360 internal	5 s
Micro & macro calcifications [17]	TPR 0.4@1 FPI TPR 0.8@10 FPI	410 INB	20 s
Mass results	TPR@FPI	#images	Runtime
Ours on masses	TPR 0.93@0.56 FPI	100/410 INB	5 s
Ours on masses	TPR 0.9@1 FPI	750 internal	5 s
Benign/Malignant [16]	TPR 0.96 ± 0.03@1.2 FPI TPR 0.87 ± 0.14@0.8 FPI	410 INB	20 s

performance. Comparison to published results in the field is difficult as most of the previous work focused on binary classification and also mainly report on DDSM. Yet, Table 1 and Fig. 3 demonstrate the high accuracy obtained by our approach, the advantage of adding the cascade to the Faster-RCNN model and the generalization ability of our model to different classes of abnormalities. The performance on calcifications, specifically the small ones remains to be further investigated and improved.

The visual results in Fig. 4, shows a zoom in on the malignant calcification detection results. A representative example of true-positive, false-negative and false positive detections is given in Fig. 5 for various abnormal subtypes including micro calcifications macro calcification and masses.

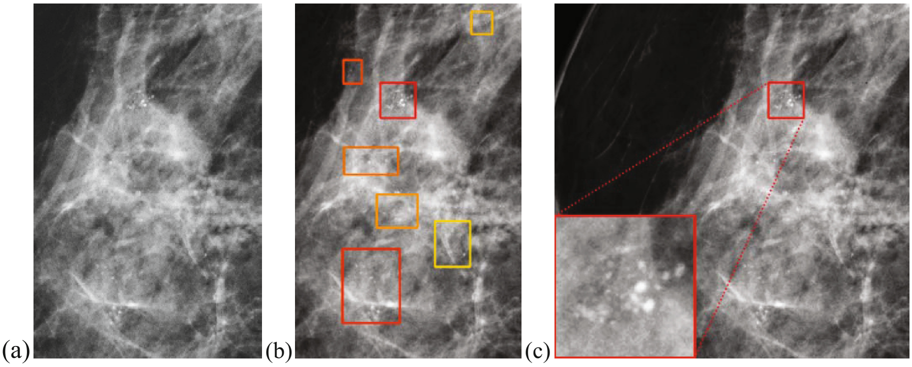

Fig. 4. **Qualitative Detection Results** displayed on a cropped image with clustered malignant micro calcifications (a) Original (b) With accurately detected boxes, where the lines in autumn colors correspond to score (red and yellow represents high and low score respectively) (c) zoom in on highest scoring detected box. (Color figure online)

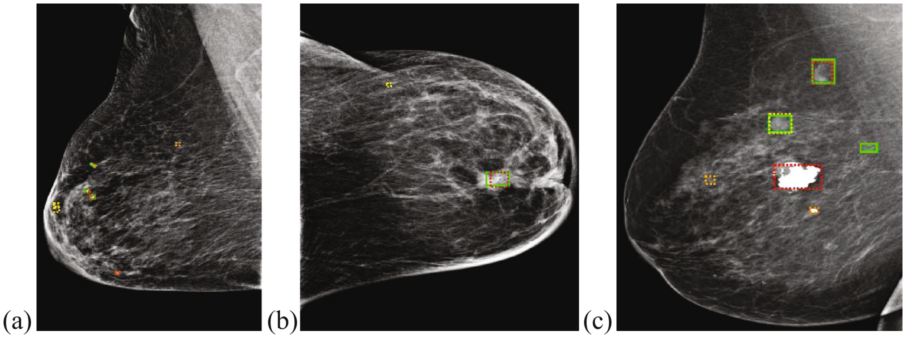

Fig. 5. Examples of detection results on full breast images. Ground truth annotation is highlighted in green for the class depicted while the candidates of automatic detection can be viewed in dashed line in autumn colors corresponding to score (red and yellow represents high and low score respectively). (a) Macro calcification (b) Micro calcification and (c) Masses. (Color figure online)

4 Summary

The paper presents an efficient DNN framework for detection of abnormalities in MG images focusing on the demanding task of detection localization and classification of masses and calcifications. The unified DNN built, is composed of a three-stage cascade, including RPN, a modified Faster-RCNN and a classifier CNN aimed at reducing the FP's. Our results are competitive with the state of the art in terms of efficiency and accuracy showing promising results for mass and calcification detection on large datasets. Future work will extend this approach to a multi view, bilateral approach to integrate information from the four mammography screening views of the right and left breast and generalize this approach to other abnormal classes.

References

1. American Cancer Society: 2015 Cancer Facts and Figures. American Cancer Society, Atlanta (2015)
2. Saranya, R., Bharathi, M., Showbana, R.: Automatic detection and classification of microcalcification on mammographic images. IOSR J. Electron. Commun. Eng. (IOSR-JECE) **9**(3), 65–71 (2014)
3. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS, vol. 25, pp. 1097–1105 (2012)
4. Simonyan K., Zisserman A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 (2014)
5. Girshick R., Donahue J., Darrell T., Malik J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 (2014)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS. arXiv:1506.01497 (2015)
7. Uijlings, J.R., Van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
8. Girshick R.: Fast R-CNN. In: ICCV, pp. 1440–1448 (2015)
9. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: CVPR, pp. 2155–2162 (2014)
10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR (2016)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, Boston, MA, pp. 3431–3440 (2015)
12. Gulshan, V., Peng, L., Coram, M., et al.: Development and validation of a DL algorithm for detection of diabetic retinopathy in retinal fundus photos. JAMA **316**, 2402–2410 (2016)
13. Esteva, A., Kuprel, B., Novoa, R.A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **2017**(542), 115–118 (2017)
14. Giger, M.L., Karssemeijer, N., Schnabel, J.A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu. Rev. Biomed. Eng. **15**, 327–357 (2013)
15. Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J., Denton, E., Zwiggelaar, R.: A review of automatic mass detection and segmentation in MG images. MIA **14**, 87–110 (2010)
16. Dhungel, N., Carneiro, G., Bradley, A.P.: Automated mass detection in mammograms using cascaded deep learning and random forests. In: DICTA, pp. 1–8 (2015)

17. Lu, Z., Carneiro, G., Dhungel, N., Bradley, A.P.: Automated detection of individual microcalcifications from MG using a multistage cascade approach. arXiv:1610.02251v (2016)
18. Dhungelz, N., Carneiroy, G., Bradley, A.P.: Fully automated classification of mammograms using deep residual neural networks. In: ISBI 2017 (2017)
19. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: INBreast: toward a full-field digital MG database. Acad. Radiology. **19**(2), 236–248 (2012)
20. Akselrod-Ballin, A., Karlinsky, L., Alpert, S., Hasoul, S., Ben-Ari, R., Brakan, E.: A region based convolutional network for tumor detection and classification in breast mammography. In: DLMIA (2016)