DOMAIN SPECIFIC CONVOLUTIONAL NEURAL NETS FOR DETECTION OF ARCHITECTURAL DISTORTION IN MAMMOGRAMS

Rami Ben-Ari, Ayelet Akselrod-Ballin, Leonid Karlinsky, Sharbell Hashoul

IBM Research - Haifa, Israel

ABSTRACT

Detection of Architectural distortion (AD) is important for ruling out possible pre-malignant lesions in breast, but due to its subtlety, it is often missed on the screening mammograms. In this work we suggest a novel AD detection method based on region proposal convolution neural nets (R-CNN). When the data is scarce, as typically the case in medical domain, R-CNN yields poor results. In this study, we suggest a new R-CNN method addressing this shortcoming by using a pretrained network on a candidate region guided by clinical observations. We test our method on the publicly available DDSM data set, with comparison to the latest faster R-CNN and previous works. Our detection accuracy allows binary image classification (normal vs. containing AD) with over 80% sensitivity and specificity, and yields 0.46 false-positives per image at 83% true-positive rate, for localization accuracy. These measures significantly improve the best results in the literature.

Index Terms— Breast Mammography, Architectural Distortion, Deep Learning, Region Proposal, Convolution Neural Net, Computer-Aided Diagnosis

1. INTRODUCTION

Breast cancer is the second most common cause of death in women. While computer-aided diagnosis reports a high level of sensitivity in revealing lesions and calcifications, they typically fall short of detecting architectural distortion (AD). Due to its subtle appearance in the breast normal tissues, architectural distortion accounts for 12% - 45% of breast cancers, overlooked or misinterpreted in screening mammography [1]. Automatic detection of AD in mammograms raise a particular challenge due to: (1) Low visual signature (2) Ambiguous boundaries, hampering supervised learning (3) Unavailability of large data sets of AD. A limiting factor in many medical applications is lack of labeled and annotated data. Particularly large data sets of annotated AD are rare or publicly unavailable.

Highly subtle spiculations radiating out from a center point is the major hallmark of AD and many research initiatives seek to capture this characteristics by designing *handcrafted* features. This unique pattern of AD has led previous studies to embrace different approaches e.g. filter response [2, 3], search for linear structures [4, 5] or employing texture features [6]. The main shortcoming of the current methods, restricting their clinical use, is the high false-positive rate [5]. This paper addresses this important problem closing the gap towards a practical automatic tool for mammogram classification and AD localization.

Supervised CNN feature representations have been recently shown to be extremely effective for visual tasks out of their original training domain, in natural scenery [7, 8] as well as medical imaging [9, 10]. Latest computer vision benchmarks present the region proposal convolution neural neural nets (R-CNN) [11], often based on transfer learning, as the one of the top performing methods in object detection. The R-CNN combines the following ingredients: (1) Set region proposals as candidates for objects (2) Create high capacity CNN representation for each region of interest (RoI). (3) Classify the proposed RoIs as belonging to a certain type of object or background (4) Bounding box regression for precise localization. Yet, typical R-CNN [11, 18] requires hundreds of images per class in order to properly train the region proposal and then the whole network. However, large labeled data sets are often hard to obtain in many medical applications. In this paper, we present a new domain specific R-CNN, where the region proposals are not learned but extracted from the specific parenchymal regions according to radiologist best practices. This allows engaging human perception into the machine region proposal process and yields a superior method with high sensitivity and significantly lower false positive rate. In this study we use a pretrained network for AD representation. The resulting features are then classified through a cascade SVM classifier, first for the challenging task of discriminating between normal and AD containing mammograms, suggesting a demanding use case in screening mammography. An aggregation scheme then allows the localization of the AD findings. Our performance assessment is based on both calssification and localization accuracy while comparing the results to an implementation of a faster-RCNN approach for AD detection [1].

Our work entails five major contributions: (1) The celebrated R-CNN approach [11, 16] is applied here for the first

[1] A similar approach for tumor detection was shown at [18]

978-1-5090-1172-8/17/$31.00 ©2017 IEEE

time in the context of AD detection in mammograms. (2) We hereby suggest a novel domain specific R-CNN (DS-RCNN) highly performing while trained on just tens of positive AD containing mammograms. (3) Classification results imply that features obtained from pretrained CNN on natural images (from ImageNet[2]) are strong candidates for AD recognition task. (4) The suggested domain specific R-CNN outperforms the faster R-CNN and previous methods.

2. METHOD

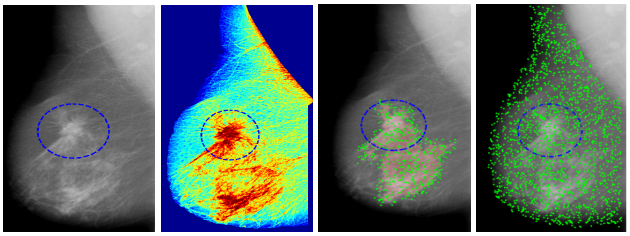
The proposed method pipeline is depicted in Figure 1. First, the breast outline and pectoral muscle are detected to limit the search domain into the breast relevant region. The segmentation of the parenchymal (a.k.a fibro-glandular) tissues is based on the followintg *unsharp mask* filter resulting a high response on the spiculated pattern, characterizing AD:

$$I_{hpf} = \alpha I - G * I \quad (1)$$

where I is the input image, α is a constant factor, G is a Gaussian and $*$ denotes standard convolution. Fig. 2 (b) shows the filter response on a sample case. Regions with high response (over a constant threshold) are then consolidated to create a mask for the parenchymal region as described in [12]. Clinical observations show that AD is likely to be found in the parenchymal region and often it's boundaries. Our region proposal therefore is based on sparse sampling from these domains, extracting bounding boxes around each sample point. The sparse sampling reduces redundancy and is justified by the translation invariance property of CNN representation (which is satisfied to a certain degree due to pooling) as well as our shift image augmentations in the training set. Figure 2 (c-d) demonstrate the sampling regime in the parenchymal domain with comparison to sampling from the whole breast. Note, the high concentration of sample points in the true region (labeled by blue circle) when samples are restricted to the parnchymal domain. To obtain the same density of samples in the true region, fourfold higher sampling is required in the breast. Training on many irrelevant samples not only raises the computational cost but also yields a weaker classifier. Although ADs missed by the segmentation appear as false-negative (in our test bed it reached up to 2%), the overall performance of the detector is significantly improved.

Each sample point defines a set of square regions of interest r_i^k (RoI) at different scales where, k indexes the image and i the particular bounding box. We follow the *transfer learning* approach [7, 8, 9] using a pretrained CNN for feature representation. To this end, we use the CNN-M pretrained convolutional network by the visual geometry group (VGG) of the University of Oxford, trained on ImageNet data set [13]. This

[2]http://www.image-net.org/



(a) (b) (c) (d)

Fig. 2. Region proposal and sampling. (a) The image. AD region is outlined in blue. (b) Filter response (red is higher) (c) Sampling in the suspicious region (d) Equivalent case with $\times 4$ sampling rate through the whole breast to obtain in average a similar sampling density in the AD region.

network consists of five convolution layers, three fully connected layers and a SoftMax layer with 1000 output nodes. In our model we use the output of the last fully connected layer (full7 layer-20) as the feature vector. Considering a small size of positive data set (containing an AD finding) we employ a modification of CNN-M network, with low dimensional full7 layer, resulting in 128D output. In order to fit the data into the network, the image is first transformed to 8 bit and, replicated to 3 RGB channels. Each RoI image was then resized to 224×224 according to the network input size. We further enhance our training set by image augmentation conducted on positive samples with 5 random shifts, 3 rotations and 2 flips (total of 10 augmentations).

To handle false positives in the detection stage we train a cascade linear-SVM where negative samples are random RoIs extracted from the parenchymal region in **normal** images, representing various patterns of healthy parynchemal tissues. This yields 520 positive (including augmentations) against 21 thousand negative RoIs. This highly imbalanced data is then handled by a cascade classifier scheme similar to the Viola-Jones [14] method in face detection. However, we found the following modifications to be to yield a significant improvement: At each cascade level the classifier is trained on positive samples balanced by a random subset of the false positives from the previous stage (hard negatives). Striving for high sensitivity, positive samples incorrectly rejected among the cascade training stages, are returned to the process at the following cascade level. Drawing RoI samples from the clinically meaningful parenchymal domain, creates a domain specific classifier with higher capacity in discriminating healthy and AD affected tissues.

In the testing set-up, bounding boxes (represented by their corresponding feature vector) are fed into the cascade classifier obtaining a score (i.e. probability), $\mathcal{C} : r_i^k \rightarrow s$. Next, positively classified RoIs are clustered using the Mean-Shift method [15]. The Mean-Shift algorithm is a non-parametric clustering technique which does not require prior knowledge

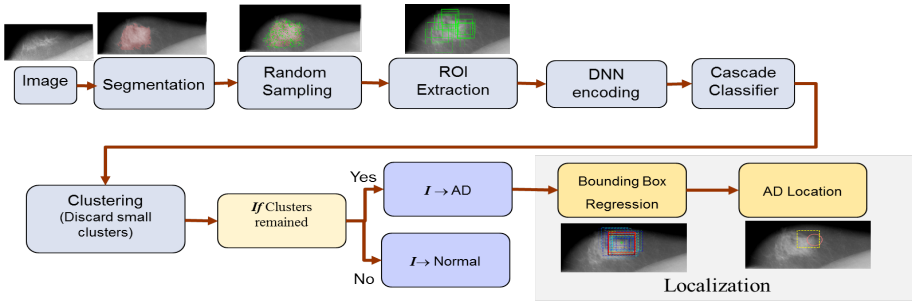


Fig. 1. The Processing Flow Chart. The breast parenchymal tissues are first detected by segmentation, then sampled in the interior and boundaries to extract candidate RoIs. Next, CNN-encoded candidates are classified. The resulting positive samples are then clustered. After rejection of small clusters and image classification the AD is localized by bounding box regression.

for the number of underlying clusters. Next, small clusters are filtered out as potentially being caused by sporadic false-detections (see parameter setting in sec. 4). Dense clusters however present strong indication for a true finding and therefore an image is classified as positive, containing a valid cluster, and negative otherwise. The last step of our pipeline addresses the localization. To this end, a regressor is trained similar to R-CNN [16], to yield a prediction as bounding box around the AD finding. Each predicted bounding box is then associated by a confidence level according to the mean score of the cluster members, as obtained from the classifier.

3. EVALUATION EXPERIMENTS

We address two tasks: (1) Binary image classification to AD (positive) vs. normal (negative) and (2) AD localization. Our test bed is taken from the publicly available DDSM data set (digital database for screening mammograms) [17]. A subset of 52 images from 42 patients were selected by an expert radiologist (last author) to have a relatively high certainty for presence of (spiculated/primary) AD. The negative set included 84 normal images from 37 patients. Of the breasts with AD abnormalities 2% were fatty 71% were glandular and 27% were extremely dense (with high prevalence of parenchymal tissues in the breast). The performance on highly dense breasts attracts particular interest, due to normal tissue occluding effect, hampering the detection. Normal mammograms selected from DDSM were 12% fatty, 56% glandular and 32% extremely dense breasts. Our performance assessment was carried out with leave-one-patient-out and 5-fold patient-wise cross-validation i.e. at each train and test iteration, all the images from the patient under test were strictly excluded from the training set.

4. RESULTS

We analyze the performance of two approaches, the top performing Faster R-CNN [18], trained for AD detection and the proposed domain specific R-CNN (DS-RCNN). The set up for DS-RCNN included 480 randomly sampled points in the parenchymal tissue region (320 interior and 160 on the boundary) with RoIs extracted at 3 different scales 1,1.2 and 1.4 from the nominal size of 128×128 (approx. 25×25 $[mm]$). Number of cascade levels in our classifier was set to 5 and clusters below 15 members were discarded. Our Faster R-CNN implementation was based on VGG-16 architecture [13] and Caffe with a deep fully convolutional network that generates RoI candidates [16]. This model was tested on the same data set with 5 image level augmentations (3 rotations and 2 flips) and 5-fold patient-wise cross validation .

AD findings in our data set present a large scale variability of up to $\times 5$ scale factor, caused by the true size variation, ambiguous boundaries and loose annotation. We therefore find standard intersection over union to be an improper measure in this use case. In a practical detection scenario of a CAD system, it is often sufficient to imply the presence of a suspicious finding rather than exact delineation or a tight bounding box, which is apriori ill-defined (see Fig. 2-a, and fourth column in Fig. 3). To validate the localization, we therefore consider a symmetric overlap ratio. Let t denote the annotation set and p_i a predicted RoI. The symmetric overlap ratio is then defined as:

$$\mathcal{R}_i = \mathcal{R}(p_i) = \max\left(\frac{|t \cap p_i|}{|t|}, \frac{|t \cap p_i|}{|p_i|}\right), s.t.\ m \leq |p_i| \leq M \quad (2)$$

Here, both subset and superset of t are considered as true within the scale factor limits of the predicted RoI $[m, M]$. The predicted mask is then obtained as the union of all valid RoIs (that overlap the true mask over a certain ratio) i.e, $p_k = \bigcup_i \{p_i | \mathcal{R}_i \geq \alpha\}$, where i indexes the RoI and k the image. We therefore allow an extremely large finding to be covered by several bounding boxes. The parameter α is a threshold on the overlap ratio set as 0.4 (40% overlap). Accordingly, two false-positive measures are defined for localization: FP_D as the average false positive detection in the true-positive images

(containing AD) and FP_T, commonly used in the literature presenting the average false positive per image with respect to the total size of the data set (including normal images):

$$TPR = \frac{\#\{\mathcal{R}(p_k) \geq \alpha\}}{\#\{\text{AD findings in the images classified as true}\}} \quad (3)$$

$$FP_D = \frac{\#\{\mathcal{R}_i < \alpha\}}{\#\{\text{images classified as true}\}} \quad (4)$$

$$FP_T = \frac{\#\{\mathcal{R}_i < \alpha\}}{\#\{\text{images in the cohort}\}} \quad (5)$$

Table 1 shows a comparative analysis for different approaches including DS-RCNN, R-CNN without object proposal (applied to the whole breast - BR×4), Faster R-CNN with data driven object proposal and three recent approaches in the literature with their corresponding test bed. For fair comparison results for two sets of validations are shown in Table 1, leave-one-out (LOO) and 5-fold cross validation (CV) both patient-wise. The results are aligned around the image classification sensitivity. For the image classification task, the suggested DS-RCNN yields the best result with over 80% sensitivity and specificity rates. In terms of localization accuracy, DS-RCNN achieves 83% detection rate at $FP_T = 0.46$ false positives per-image (FFPI). As expected, the performance for 5-fold CV is lower than LOO, due to the smaller training sets associated with CV method.

Applying the same strategy without segmentation (BR×4 in Table 2) doubled FPPI, reduced image classification specificity by 30% while resulting a slight improvement of 6% in localization sensitivity. The faster R-CNN shows low specificity in the image classification task and inferior accuracy in object localization. We assume the reason for this trend to be insufficient training samples for the region proposal network and the end to end training.

With comparison to previous work, best results were reported by Yoshikawa et al. [3] with 1 FPPI @ 0.83 TPR. According to Table 1 our results demonstrates a significant improvement, reducing FPPI by 54% at same sensitivity. Note that error rate of below 1 FPPI is obtained for the first time in AD detection, at over 80% sensitivity [3]. Analysis for the dense breast mammograms shows that DS-RCNN yields 80% detection rate, where human sensitivity is particularly low, suggesting a use case of second reader in screening mammography. Several successful localization results are shown in Fig. 3. The false-positive/negative failure cases were mostly found in dense parenchymal tissues or low contrast images.

With respect to runtime, the faster R-CNN method took $\sim$ 30 hours to train on a i7 Intel CPU with 64G RAM and TitanX GPU. Testing is fast and takes only 0.2 sec per image. The DS-RCNN was implemented in Matlab using *matconvnet*. Training the system with our cascade classifiers on 21K RoIs takes $\sim$ 1.5 min on the same platform. Testing time was 8.3 sec/image.

[3]The accuracy indexes provide a subjective evaluation due to different data sets and evaluation measures

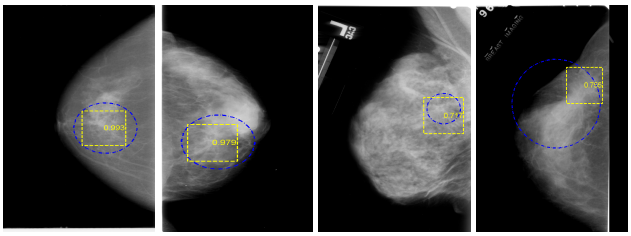


Fig. 3. Localization results. Annotation in blue circle and detection in yellow box. Images show successful detection without false positives. Values indicate confidence measures. Note the successful detection and accurate localization despite the high breast density. Annotation is marked by circle for improved visibility.

5. SUMMARY AND DISCUSSION

Detection of architectural distortion (AD) is one of the most challenging tasks in mammography for both radiologists and computer algorithms. In this study, we addressed the problem of binary image classification (AD vs. normal) and the AD localization. We have analyzed the high performing region proposal CNN (R-CNN) under k-fold cross validation test, for detection and localization of AD in mammograms. A new domain specific R-CNN was suggested, guided by the radiologist best practices for region proposals in training and testing. The benefit of guided sampling is compared to random sampling of whole breast and data driven learning of object proposals. The suggested domain specific R-CNN outperforms faster R-CNN being able to detect AD findings with relatively high accuracy using as little as 52 positive images from 42 patients.

Our approach does not rely on accurately delineated findings, accepting loosely rectangular annotation. This has a clear advantage for supervised training on large data sets. The suggested CNN based method eliminates the need for handcrafted features, and allows transferring the method to new modalities and organs with minimal overhead.

The DDSM public data set consists of scanned screen-film mammograms, mostly replaced today with the full-field digital mammography (FFDM). This data set is commonly used in literature since FFDM data sets with sufficient AD cases are publicly unavailable. Due to the different nature of the digital and scanned-film image acquisition, a separate model should be trained for each type of image.

	Classification		Localization			Data Set	Data Source	Validation
	Sens.(%)	Spec. (%)	TPR	FP_D	FP_T			
DS-R-CNN	80.8	**81.0**	0.83	**0.88**	**0.46**			Leave-One-patient-Out
BR×4	80.8	57.1	**0.88**	1.41	0.91	52 AD & 84 N.	DDSM	Leave-One-patient-Out
DS-R-CNN	80.8	77.4	0.79	0.95	0.52			*5-fold patient cross valid*
Faster R-CNN	80.8	68.7	0.41	1.24	0.69			*5-fold patient cross valid*
Rangayyan [5]			0.80		3.7	102 AD & 52 N.	Proprietary	Leave-One-patient-Out
Matsubara [4]			0.81		2.6	168 AD & 580 N.	Proprietary	
Yoshikawa [3]			0.83		1.0	40 AD & 160 N.	DDSM	

Table 1. Classification and localization accuracy for DS-RCNN at two sensitivity work points. R-CNN without object proposal (BR×4) with ×4 sampling. Results for faster R-CNN with learned object proposal and several methods from literature are shown. FP_D and FP_T present two measures for average false-positives defined in (5). Best results are in bold.

6. REFERENCES

[1] H. Burrell, A. Evans, A. Wilson, and S. Pinder, "False-negative breast screening assesment: What lessons we can learn?," *Clin. Radiol.*, vol. 56, pp. 385–388, 2001.

[2] S. K Biswas and D. P. Mukherjee, "Recognizing architectural distortion in mammogram: A multiscale texture modeling approach with GMM," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 2023–2030, 2011.

[3] R. Yoshikawa, A. Teramoto, T. Matsubara, and H. Fujita, "Automated detection of architectural distortion using improved adaptive gabor filter," in *Proc. of IWDM*, 2014.

[4] T. Matsubara, A. Ito, A. Tsunomori, T. Hara, C. Muramatsu, T. Endo, and H. Fujita, "An automated method for detecting architectural distortions on mammograms using direction analysis of linear structures," *EMBC*, pp. 2661–2664, 2015.

[5] R. M. Rangayyan, S. Banik, J. Chakraborty, S. Mukhopadhyay, and J. E. L. Desautels, "Measures of divergence of oriented patterns for the detection of architectural distortion in prior mammograms," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 8, no. 4, pp. 527–545, 2013.

[6] Amit Kamra, V. K. Jain, S. Singh, and S. Mittal, "Characterization of Architectural Distortion in Mammograms Based on Texture Analysis Using Support Vector Machine Classifier with Clinical Evaluation," *J. Digit. Imaging*, pp. 104–114, 2016.

[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," *CVPR*, pp. 1717–1724, 2014.

[8] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," *CVPPRW*, pp. 512–519, 2014.

[9] Yaniv Bar, Idit Diamant, Lior Wolf, and Hait Greenspan, "Chest pathology detection using deep learning with non-medical training," in *ISBI*, 2015.

[10] M. S. Hefny, T. Okada, M. Hori, Y. Sato, and R. E. Ellis, "Unregistered multiview mammogram analysis with pre-trained deep learning models," *MICCAI*, vol. 9350, pp. 238–245, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, 2014.

[12] Rami Ben-Ari, Aviad Zlotnick, and Sharbell Hashoul, "A weakly labeled approach for breast tissue segmentation and breast density estimation in digital mammography," in *ISBI*, 2016.

[13] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.

[14] Paul Viola and Michael J. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[15] E. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 603619, 2002.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[17] J. E. D. Oliveira, M. Guld, A. Araujo, B. Ott, and T. M. Deserno, "Towards a standard reference database for computer-aided mammography," in *Proceedings of SPIE Medical Imaging*, 2008.

[18] Ayelet Akselrod Balin, Leonid Karlinsy, Sharon Alpert, Sharbel Hasoul, Rami Ben-Ari, and Ella Barkan, "A region based convolutional network for tumor detection and classification in breast mammography," in *DLMIA-MICCAI*, 2016.