Siamese Network for Dual-View Mammography Mass Matching

Shaked Perek(✉), Alon Hazan, Ella Barkan, and Ayelet Akselrod-Ballin

IBM Research, Haifa, Israel
{shaked.perek,alon.hazan,ella,ayeletb}@il.ibm.com

Abstract. In a standard mammography screening procedure, two X-ray images are acquired per breast from two views. In this paper, we introduce a patch based, deep learning network for lesion matching in dual-view mammography using a Siamese network. Our method is evaluated on several datasets, among them the large freely available digital database for screening mammography (DDSM). We perform a comprehensive set of experiment, focusing on the mass correspondence problem. We analyze the effect of transfer learning between different types of dataset, compare the network based matching to classic template matching and evaluate the contribution of the matching network to the detection task. Experimental results show the promise in improving detection accuracy by our approach.

Keywords: Biomedical imaging · Deep learning · Mammography

1 Introduction

Mammography (MG), the primary imaging modality for breast cancer screening, typically utilizes a standard dual-view procedure. Two X-ray projection views are acquired for each breast, a craniocaudal (CC) and a mediolateral oblique (MLO) view. Examining the correspondence of a suspected finding in two separate compression views, enables the radiologist to better classify an abnormality. Studies have shown that using a two-view analysis helps radiologists reduce false positive masses caused by overlapping tissues that resemble a mass, and ultimately helps achieve a higher detection rate [17]. Although Computer Aided Diagnosis (CAD) algorithms were developed to assist radiologists, their usefulness has been debated. This is partially due to the many false positives they produce, especially for masses and architectural distortions. We propose a novel approach for identifying the correspondences between masses detected in different views, to further improve the detection and classification of MG algorithms.

Previous work on MG classification employed hand-crafted features, such as texture, size, histogram matching, distance from the nipple, and more. The extracted features were then classified together using various techniques to assess the similarity between image pairs. [11] demonstrated the positive effect of dual-view analysis, which detects suspicious mass in one view and its counterpart

D. Stoyanov et al. (Eds.): RAMBO 2018/BIA 2018/TIA 2018, LNCS 11040, pp. 55–63, 2018.
https://doi.org/10.1007/978-3-030-00946-5_6

in the other view. Based on geometric location, this analysis fuses both sets of features and classifies them with linear discriminant analysis. [1] used dual view analysis to improve single-view detection and classification performance by combining the dual-view score with the single-view score. Features were obtained manually using candidate location, shape, and image characteristics.

Deep learning approaches have already shown impressive results in MG detection and classification. [3] presents a micro-calcification (MC) classification that uses a dual-view approach based on two neural networks; this is followed by a single neuron layer that produces the decision based on the concatenated features from both full image views. [15] presents a multiscale convolutional neural networks (CNN) for malignancy classification of full images and sub-image patches integrated with a random forest gating network. Dhungelz et al. [5] proposed a multi-view deep residual network (Resnet) to automatically classify MG as normal/benign/malignant. The network consists of six input images, CC and MLO together with binary masks of masses and MC. The output of each Resnet is concatenated, followed by a fully connected layer that determines the class. Similarly, [6] proposed a two-stage network approach that operates on the four full images: CC and MLO of the left and right breasts. The second stage concatenates the four view-specific representations to a second softmax layer, producing the output distribution.

Most multi-view deep learning approaches to MG are applied on unregistered full images and concatenates the features obtained by the network on each view separately. In contrast, we propose a Siamese approach that focuses on matching localized patch pairs of masses from dual views. Siamese networks are neural networks that contain at least two sub-networks, with identical configuration, parameters, and weights. During training, updates to either path are shared between the two paths. To address the correspondence problem, previous works used the **Siamese network** [10] to simultaneously train inputs together. [4] uses this type of network for a face verification task, in which each new face image was compared with a previously known face image. [16] demonstrate the advantage of Siamese networks by detecting spinal cord mass in different resolutions. Sharing parameters leads to fewer parameters allowing training with smaller datasets. The subnetworks representation is related, and thus better suited for the comparison task.

Our work entails three key contributions: (1) A novel deep learning dual view algorithm for mass detection and localization in breast MG based on Siamese networks, which have not been used before to solve lesion correspondence in MG. (2) A careful set of experiments using several datasets to study the contribution of the network components, also showing that the network is better than the classic template matching approach. (3) Evaluation on the DDSM database.

2 Methods

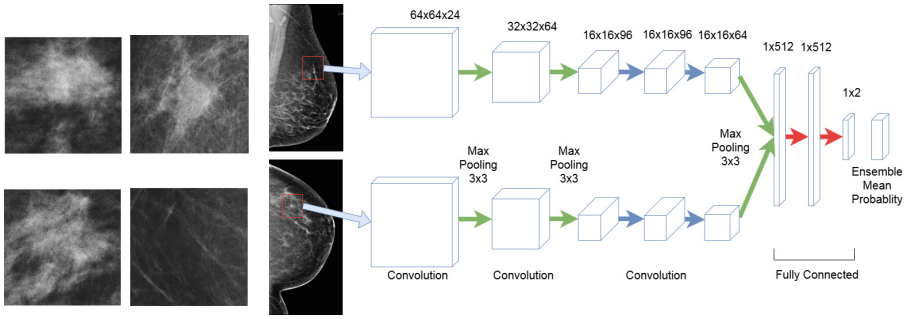
For this study, our input took unregistered CC/MLO MG images and matched between lesions appearance in both views. Below, we describe the network

matching architecture, the experimental methodology including fine-tuning and comparison to template matching and how the matching architecture is integrated into an automatic detection pipeline.

2.1 Matching Architecture

Our approach extends the work presented by Han et al. [8]. The authors developed MatchNet, a CNN approach for patch-based matching between two images. The network consists of two sub-networks. The first is a **feature network**, a Siamese neural network, in which a pair of patches, extracted from the CC and MLO views are inserted and processed through one of two networks. Both paths consist of interchanging layers of convolutions and pooling, which are connected via shared weights. The second is the **metric network**, which concatenates the two features, contains three fully connected layers and uses a softmax for feature comparison. Dropout layers were added after layers FC1 and FC2 with value of 0.5. The network is jointly trained with a cross entropy loss. Figure 1(c), presents the modified network, including the network's ensemble approach.

The mammography datasets employed for this study were created by defining a positive image pair label, as the detections annotated by a radiologist in each view, while a negative pair label is defined by matching false detection with annotated detections in the other view.



(a) Pair (b) Non-pair (c) Architecture

Fig. 1. The dual-view matching architecture. Columns (a, b) are illustration of ROI input patches from two views, CC and MLO. (a) Matching pair of images (b) Non matching images. (c) Patch pairs from CC and MLO views are inserted to the network. The feature network, consists of interchanging layers of convolutions and pooling, share parameters between paths. The metric network has fully connected layers with dropout, produce the final decision by networks ensemble.

2.2 Fine Tuning the Network

Fine-tuning and transfer learning have shown to improve performance results despite of specific application domains [14,18]. To adapt MatchNet to the task

of matching detections from different MG views, we first evaluated fine-tuning. We fine-tuned by training the layers of the metric network, i.e. the three fully connected layers and the last convolution layer from the feature network. We used three different datasets, as described in the Experiment and Results section, including: Photo tourism (natural image pairs)[12], Digital Database for Screening Mammography (DDSM) [9] and In-house dataset. We used the trained weights of one dataset domain to fine tune the other datasets.

2.3 Template Matching

Template matching, which extracts sub-image patches and computes a similarity measure that reflects the template and image patch correspondence, has been used extensively in computer vision [2]. We compare our deep learning network to template matching with normalized cross correlation. Intuitively, we assume that the similarity of image patches of a mass in one view with the same mass in the other view under deformations, will be higher than the similarity with a different mass or region of the breast [7].

2.4 Dual-View Automatic Lesion Detection

We integrated two components, a **matching architecture** and a single-view **detection algorithm** to exploit the contribution of the dual-view network to the full pipeline. The detection algorithm is based on a modified version of U-net [13], which was originally designed for the biomedical image processing field. In the original U-net, the output size is identical to the input size. However, for our task segmentation is not required at the pixel level, since the boundary of tumors and healthy tissue is ill-defined. Thus, we modified the U-net output, so that each pixel of the output, corresponds to a 16×16 pixels area of the input.

The system flow is such that, given a dual-view pair of images as input, the single-view detection algorithm is applied separately on the CC, MLO image I_{cc}, I_{MLO} and outputs a set of candidate patches, $P_{CC} = \{p^1_{CC}, ...p^N_{CC}\}, P_{MLO} = \{p^1_{MLO}, ...p^M_{MLO}\}$ respectively. The objective of the matching architecture is to identify the correspondences. If both patch candidates, CC and MLO views, from the detection flow, are identified as a true lesion, then the label for the pair will be true and accordingly considered a positive match. We assign labels to each pair based on the Dice Coefficient threshold δ, between two masks, defined by a detection contour and ground truth lesion contour respectively. For our experiments, we used $\delta = 0.1$ as the threshold. Any contour with a larger score is said to be a true lesion.

3 Experiments and Results

3.1 Data Description

We carried out the experiments on three datasets: (a) The Photo Tourism dataset [12], consists of three image datasets: Trevi fountain, Notre Dame and Yosemite.

Which is similar to the dataset used in the MatchNet paper [8]. It consists of 1024×1024 bitmap images, containing a 16×16 array of image patches. Each image patch has 64×64 pixels and has several matching images that differ in contrast, brightness and translation. (b) The Digital Database for Screening Mammography (DDSM) [9], contains 2620 cases of four-view MG screenings. It includes radiologist ground truth annotations for normal, benign and malignant image. 1935 images contain tumors. (c) The In-house dataset includes benign and malignant tumor ground truth annotations, from both CC/MLO MG views for either left, right or both breasts. It contains 791 tumor pairs. Figure 1(a, b) shows some tumor pairs from In-house dataset used as positive examples for the network versus negative examples. We randomly split the data into training (80%) and testing (20%) subsets of patients. The partitioning was patient-wise to prevent training and testing on images of the same patient.

3.2 Patch Preprocessing and Augmentations

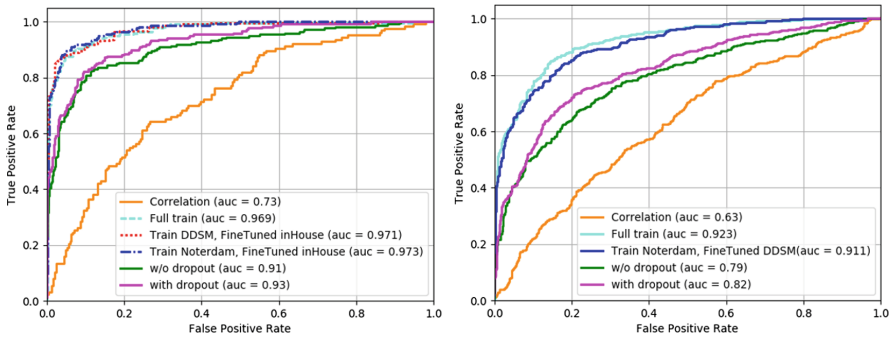
We extracted ROI patches from the full MG images of 4000×6000 pixels by cropping a bounding box around each detection contour. Each such bounding box was enlarged by 10% in each dimension to include useful information around the lesion border. The extracted patches were then resized to 64×64 to generate the input to the network. We normalized all the datasets by subtracting the mean of each image and dividing by the standard deviation of each patch, avoiding the proposed MatchNet normalization [12].

Augmentation was utilized throughout the training stage on all three datasets, such that each patch was flipped left and right and rotated by 90°, 180°, 270°. Each augmented patch was matched with all the others augmented patches. Medical datasets are generally unbalanced. Namely, the number of positive pairs are significantly smaller than the negative pairs. Thus, we train two networks, each network has a balanced input of positive pairs and randomly selected negative pairs. In the testing stage, we evaluate each test image through all networks, and achieve a final score using a mean probability.

We trained with a learning rate of 0.0001, Adam optimizer and batch size of 512. Experiments were performed on a Titan X Pascal GPU. Training time for DDSM models took 4 h. Testing time with model ensemble took 6 s.

3.3 Fine Tuning the Network

We studied the contribution of fine-tuning on the results in three experiments. Full training on Photo tourism and fine tuning with (i) In-house (ii) DDSM (iii) Full training on DDSM and fine tuning with In-house. (i+ii) were done using Notredam dataset. The results for these tests are presented in Fig. 2, where the upper and lower subfigures correspond to the In-house and DDSM dataset respectively. The comparison of the In-house and DDSM full training results (AUC 0.969, 0.92) with the fine tuning results (AUC 0.973, 0.91) did not show a clear advantage over the fine tuning process. This can be explained by two factors: the domain transfer effect, namely despite the Notredam large dataset



(a) In-house results (b) DDSM results

Fig. 2. Fine tuning ROC results. The figures demonstrate the different experiment performed to evaluate the ability of the matching architecture to classify MG pairs and non pairs. (a) In-house dataset shows no advantage for fine tuning. (b) DDSM dataset shows best result by full train (cyan). (Color figure online)

of image pairs, natural images are different than medical images. Second, the Noterdam dataset pairs are much more similar to each other than the different views pairs from the breast images, which go through deformation.

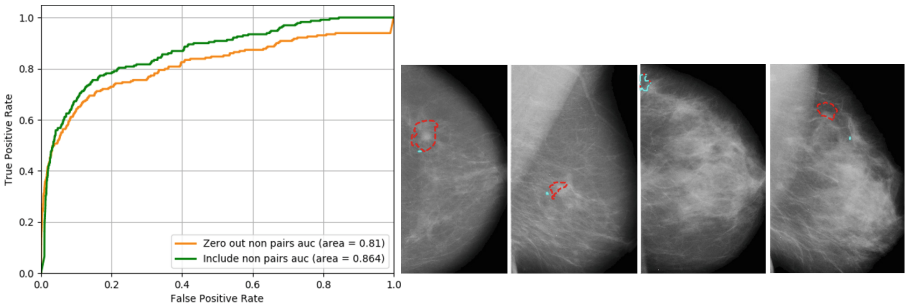
Fine tuning the DDSM with the In-house dataset in (iii), obtained (AUC 0.971) compared to full training of (AUC 0.969). DDSM is a large MG dataset, however it is acquired with a different imaging technique from the In-house data (full field digital mammography) and this might explain the similar results. The ROC plot also shows the improvement in AUC by adding dropout in Fig. 2.

3.4 Template Matching

The cross-correlation score was transformed from the range of $[-1, 1]$ to $[0, 1]$ to represent the score as probabilities. The correlation presented in Fig. 2 obtained significantly lower results of AUC 0.73, 0.63 on In-house, DDSM respectively.

3.5 Dual-View Automatic Lesion Detection

To evaluate the contribution of the matching architecture to the full detection pipeline, we applied the single-view detection algorithms on the CC, MLO image pairs followed by the matching architecture on the DDSM dataset. In some cases, detections will appear only for one view and not in the other. These cases cannot be evaluated using the matching architecture. Thus, two possibilities arise, exclude all detections without a pair or include them. Figure 3(a) shows the classification of the set of patches into positive and negative matches, generates an AUROC of 0.864, 0.81 depending on whether the small set of detections with no-pairs were included or excluded. We conclude that it is reasonable to include these detection as some tumors may be identified only in a single view.



(a) Patch matching ROC (b) CC/MLO (c) CC/MLO

Fig. 3. Results of automatic lesion detection pipeline. (a) Green curve includes detections with no-pair in second view, orange curve excludes those detection. Detection examples on DDSM dataset (b, c). Red contours denote automatically detected pairs that correspond to GT while, the cyan contours are false positive automatic detections that were reduced by the dual-view algorithm. (Color figure online)

Additionally, Fig. 3(a) shows that proposed approach can reduce the false positive detection rate while keeping a high sensitivity. For MG pairs matching, we can keep a sensitivity of 0.99 and specificity of 0.19. Namely, by keeping the standalone detections we are able to reduce the false positives by almost 20%. Fig. 3(b, c), illustrates the full pipeline prediction on MG images. Probabilities of the false detections pairs (in cyan) are omitted in the final detection output. This is similar to the approach used by human radiologists, first detecting suspicious findings and then analyzing them by comparing the dual-view appearance.

4 Discussion

Finding correspondence between patches from different views of the same breast is a challenging task. Each image from MLO/CC views undergoes nonlinear deformations which can make the lesions very different from each other. On the other hand, being able to detect the lesion in both views can help the radiologists reach more accurate findings. In this work, we propose a dual-view Siamese based network, in which the architecture learns a patch representation and similarity for lesion matching. We demonstrate the advantage of a learned distance metric implemented in the network and its value in addition to a single view detection. This work can also be extended to 3D mammography by applying 3D patches. Future work will extend this work to other types of findings such as calcifications and will utilize mass location information to better eliminate false positives.

References

1. Amit, G., Hashoul, S., Kisilev, P., Ophir, B., Walach, E., Zlotnick, A.: Automatic dual-view mass detection in full-field digital mammograms. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9350, pp. 44–52. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24571-3_6
2. Ballard, D.H., Brown, C.M.: Computer Vision, 1st edn. Prentice Hall Professional Technical Reference, New York (1982)
3. Bekker, A.J., Greenspan, H., Goldberger, J.: A multi-view deep learning architecture for classification of breast microcalcifications. In: IEEE 13th International Symposium on ISBI, pp. 726–730. IEEE (2016)
4. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on CVPR, vol. 1, pp. 539–546. IEEE (2005)
5. Dhungel, N., Carneiro, G., Bradley, A.P.: Fully automated classification of mammograms using deep residual neural networks. In: ISBI, pp. 310–314. IEEE (2017)
6. Geras, K.J., Wolfson, S., Shen, Y., Kim, S., Moy, L., Cho, K.: High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047 (2017)
7. 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)
8. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: MatchNet: unifying feature and metric learning for patch-based matching. In: CVPR. IEEE (2015)
9. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The digital database for screening mammography. In: Digital Mammography, pp. 431–434 (2000)
10. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2 (2015)
11. Paquerault, S., Petrick, N., Chan, H.P., Sahiner, B., Helvie, M.A.: Improvement of computerized mass detection on mammograms: fusion of two-view information. Med. Phys. **29**(2), 238–247 (2002)
12. http://phototour.cs.washington.edu/patches/default.htm (2007)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE TMI **35**(5), 1299–1312 (2016)
15. Teare, P., Fishman, M., Benzaquen, O., Toledano, E., Elnekave, E.: Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J. Digit. Imaging **30**(4), 499–505 (2017)
16. Wang, J., Fang, Z., Lang, N., Yuan, H., Su, M.Y., Baldi, P.: A multi-resolution approach for spinal metastasis detection using deep siamese neural networks. Comput. Biol. Med. **84**, 137–146 (2017)

17. Warren, R.M., Duffy, S., Bashir, S.: The value of the second view in screening mammography. Br. J. Radiol. **69**(818), 105–108 (1996)
18. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)