Exploring Deep-Based Approaches for Semantic Segmentation of Mammographic Images

Hugo Neves de Oliveira[1](✉), Claudio Saliba de Avelar[2],
Alexei Manso Corrêa Machado[3,4], Arnaldo de Albuquerque Araujo[1],
and Jefersson Alex dos Santos[1]

[1] Computer Science Department, Universidade Federal de Minas Gerais,
Belo Horizonte, Brazil
{oliveirahugo,arnaldo,jefersson}@dcc.ufmg.br
[2] Clinical Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
claudiosaliba.rad@gmail.com
[3] School of Medicine, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[4] Computer Science Department, PUC Minas, Belo Horizonte, Brazil
alexei@pucminas.br

Abstract. Pectoral muscle and background elimination are common steps for automated software in mammographic image preprocessing. We investigate FCNs, U-nets and SegNets in the task of mammogram segmentation, addressing three subtasks: pectoral muscle, background and breast region segmentation. The MIAS and INbreast datasets were used for evaluating Deep Neural Networks on the segmentation of these regions. Several objective evaluation metrics were used in order to compare our results with the ones available in the literature. State-of-the-art results were observed in most comparisons, significantly surpassing the baselines in most metrics. Best Jaccard values (in %) for Deep Learning algorithms were 89.7 ± 2.5, 98.4 ± 0.1 and 97.0 ± 0.4 for pectoral muscle, background and breast region segmentation, respectively, in the MIAS dataset. For INbreast, the best Jaccard value achieved for pectoral muscle segmentation was 90.8 ± 2.5.

Keywords: Pectoral muscle segmentation · Breast segmentation · Mammography · Deep Learning

1 Introduction

Pectoral muscle segmentation has been used as a preprocessing step for breast cancer analysis in Computer-Aided Detection/Diagnosis (CAD) systems. Due to density similarities with potentially cancerous breast tissue, the rate of False Positive results in detection tasks tends to increase [2,3,8,14]. The Medio-Lateral Oblique (MLO) view of mammograms is the most affected by the presence of pectoral muscles. Depending on anatomy and patient positioning during image

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 690–698, 2019.
https://doi.org/10.1007/978-3-030-13469-3_80

acquisition, the pectoral muscle could occupy as much as half of the breast region, or as little as a few percent of it. Pectoral muscles can appear concave, convex or have irregular shapes in mammograms, possibly with homogeneous boundaries between it and the breast tissue. Thus, pectoral muscle segmentation is a computationally demanding task, requiring the algorithm to be able to discriminate between different shapes, sizes and breast density variations. Other important preprocessing steps in mammogram analysis are background and breast region segmentation, which present challenges due to artifacts in the background of Screen Film Mammograms (SFMs), the low-contrast of the skin-air boundary region, and to large amounts of noise, mainly present in digitized SFMs [14].

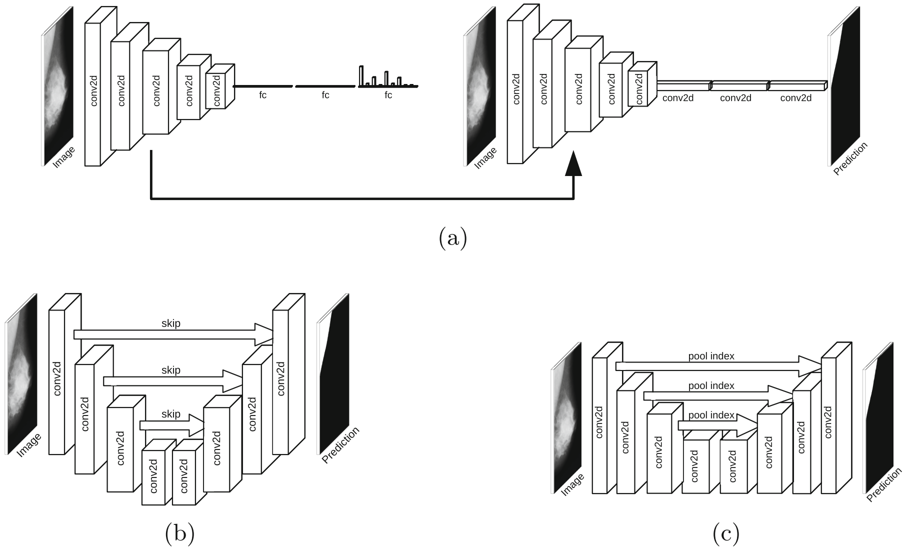
Despite their success in many Computer Vision tasks, Machine Learning methods have been noticeably absent in the mammogram segmentation body of research. As shown by Ganesan *et al.* [4], preprocessing tasks rely mostly on low-level techniques or simple statistical modelling, which may suffer from generalization/stability problems. Due to advances in high-performance parallelism, Deep Learning-based methods have evolved to comprise the state-of-the-art of most Computer Vision tasks and, recently, in Biomedical Image tasks [9].

The most recent survey [4] divides the methods used for pectoral muscle segmentation into five categories: intensity-based methods, line detection, statistical techniques, wavelet methods and active contour [3]. Further explanations of these methods is out of the scope of this paper and, therefore, readers should refer to Ganesan *et al.* [4] for a more detailed analysis of the state-of-the-art of mammogram segmentation. Several factors hamper the comparison among methods in the mammogram segmentation field. The main one is that there is not one standard set of metrics for comparison, so most of the literature uses only subjective evaluation metrics for the segmentation results, such as non-standardized specialist assessments of segmentation quality. This problem is aggravated by the lack of standardized datasets and ground truths, leading to the use of private data, severely hampering the reproducibility of most results. The most recent attempt to standardize the area was presented by Rampun *et al.* [14], which used only objective segmentation metrics and publicly available datasets, as well as ground truths obtainable upon email request. This work used active contour for modeling both the background and pectoral muscle boundary layers. It also comprised the state-of-the-art for the researched tasks, therefore it will be used as the main baseline throughout this paper.

Ganesan *et al.* [4] argues that there is not one specific method which works perfectly well for the problem of pectoral muscle segmentation. Therefore, following recent advances in Semantic Segmentation [1,10,15], the main contribution of this work is to evaluate Deep Learning-based approaches for mammogram segmentation. Secondary contributions include: (i) State-of-the-art results in pectoral muscle, background and breast region segmentation; (ii) Assessment of the superiority in stability of Deep Learning methods compared to classical approaches; (iii) Segmentation predictions and pretrained models are publicly available online for future academic use and reproducibility.

2 Deep Semantic Segmentation Approaches

Most Deep Neural Networks (DNNs) for image analysis has been based on convolution operations. Vanilla implementations of Convolutional Neural Networks (CNNs) [7] are essentially stackings of three types of layers: convolutional layers, pooling layers and fully connected layers. Convolutional and pooling layers are often stacked in the beginning of these networks and serve as learnable feature extractors, while fully connected layers play the role of the classifier at the end of the network, as can be seen in Fig. 1. In the following paragraphs we introduce the DNNs used in our experimental setup.



(a)

(b)

(c)

Fig. 1. Architecture examples for (a) FCNs [10]. (b) U-nets [15]. (c) SegNets [1].

Fully Convolutional Networks (FCNs). The most basic architectures [10], they can be understood as a patchwise approach, wherein each pixel in an image is a sample (Fig. 1a). Whole image fully convolutional training is identical to patchwise training where each batch consists of all the pixels in a set of images. Replacing fully connected layers by convolutional layers and adding a spatial loss produces an efficient machine for end-to-end dense learning [10].

U-Nets. Ever since FCNs, several attempts to mitigate the vanishing gradient problem have been proposed, most relying on alternative paths for information flow [6]. Skip connections are the most common way to create alternative paths, serving as highways for backpropagation to reach earlier layers in the network without passing through all the layers in front of them. U-nets [15] take advantage of skip connections to map higher-level contextual information to low-level pixel information. These networks are Encoder-Decoder architectures wherein the downsampling half (Encoder) is symmetrical to the upsampling half (Decoder), as shown in Fig. 1b. There is also a larger amount of feature channels in the upsampling layers compared to FCNs, which allows for more information to be propagated to higher resolution layers [15].

SegNets. SegNets [1], like U-nets, are Encoder-Decoder architectures for segmentation with symmetric layers. The Encoder half of the network is composed of VGG-like [16] convolutional layers. The construction of the Decoder network is accomplished by simply mirroring the Encoder layers and replacing the pooling layers for upsampling components (Fig. 1c). One main advantage of SegNet is the use of the pooling indices in the upsampling processes. SegNet uses the max pooling indices to upsample (without learning) the feature maps and convolves with a trainable decoder filter bank [1].

3 Methodology

Mammographic Datasets. Following the experimental procedure described by Rampun *et al.* [14], this paper's experiments were performed only on publicly available datasets. The main publicly available datasets are the Mammographic Image Analysis Society (MIAS[1]) dataset [17], the Digital Database for Screening Mammography (DDSM) [5], the Breast Cancer Digital Repository (BCDR) [11] and the INbreast[2] dataset [12]. Despite several attempts, we could not contact the BCDR team for access to their dataset, therefore it was not possible to run tests on these data. Also, to our knowledge, there are no publicly available ground truths to DDSM images, which were also removed from the analysis. The ground truths for the MIAS dataset were provided by Oliver *et al.* [13].

While DDSM, MIAS and BCDR are all SFM datasets, INbreast [12] is the only one acquired with the Full-Field Digital Mammography (FFDM) technique, rendering it the best dataset regarding image quality. This dataset contains accurate pixelwise annotations for the lesions and pectoral muscle regions. A total of 200 MLO images (from the 208 in the dataset) from INbreast were used in our tests. The 8 remaining MLO mammograms were not used due to problems with decoding the ground truths. All 322 images in MIAS were used in our experiments. Our experimental procedure was performed on the three tasks using MIAS and one task using INbreast (pectoral muscle segmentation). There is no need for breast region nor background segmentation on INbreast images, as the background can be easily segmented with a thresholding operation.

Experimental Procedure. We resized the mammograms and ground truths to 256 × 256 pixels and slightly changed the U-net architecture (setting the padding to 1) to receive these sizes of images. As the predictions of the DNNs match the 256×256 pixel size, after forwarding the images through the networks,

[1] https://www.repository.cam.ac.uk/handle/1810/250394.

[2] http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database.

we upsampled the images to their original sizes again. Results were obtained using a 5-fold cross-validation methodology over the datasets. In order to avoid artificially high results due to the similarities in breast structures of a subject, fold division was done per subject, assuring that all images of a patient are placed in the same fold. For each test fold, one of the other 4 training folds was not used in training and served as a second validation step in order to select the epoch with the best results. Details regarding the implementation and hyperparameters of the DNNs can be found in the supplementary material. Preprocessing was comprised only by the rescaling, normalization by mean and standard deviation and by horizontally flipping some mammograms in order for all images to have the same orientation. A simple post-processing of keeping only the largest contiguous white region and filling the gaps on the DNNs' binary predictions was also applied and was observed to consistently improve the results.

Evaluation Metrics. Based on previous works, we used several different segmentation metrics for validating the results. Rampun *et al.* [14] used Jaccard ($\ddot{J}$), Dice ($\ddot{D}$), Accuracy ($\ddot{A}$), Sensitivity ($\ddot{S}$), Specificity ($\bar{S}$) and Correctness ($\ddot{C}$). Other works [2,3,8] rely mostly on FP and FN metrics. In order to compare with all these works, we provide the values of all previously mentioned metrics for FCNs, U-nets and SegNets in our results (Sect. 4).

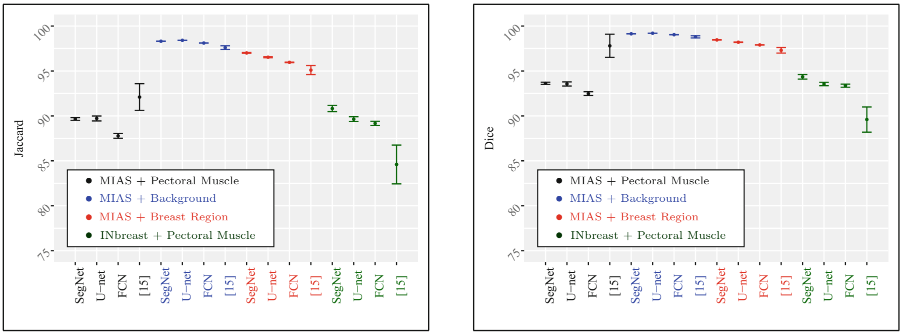
4 Results and Discussion

Due to its thorough methodology and state-of-the-art results, the main baseline used for comparison is the work of Rampun *et al.* [14]. The values for all metrics are presented in percentages (%) for easier comparisons, as some metrics (mainly FP and FN) yielded tiny proportional values, hampering the readability of the results. For assessing the statistical significance of the results compared with the baseline, we performed z-score hypothesis tests, as the number of samples is relatively large (322 images for MIAS and 200 for INbreast). Full numerical values for all metrics can be found in tables in the complementary material. In order to improve reproducibility, the best pretrained models used in our experimental procedure – as well as other complementary materials such as a script for running the pretrained models on other sets of images and additional results – are publicly available on our team's website[3].

Comparison with the Main Baseline. Figure 2 shows Confidence Intervals (CIs) with $p \leq 0.05$ for the $\ddot{J}$ and $\ddot{D}$ metrics in both INbreast and MIAS using both DNNs and Rampun *et al.*'s [14] method. One can see that in all cases the deep strategies obtained state-of-the-art results with $\ddot{D}$ and $\ddot{J}$ values close to 90% for pectoral muscle segmentation and accuracies for all tasks above 98%. Background segmentation proved to be the easiest task, with $\ddot{J}$ values over 98% and $\ddot{D}$ over 99%, configuring almost perfect background eliminations. One could argue that breast region segmentation is the most important task of all three, as most CADs are interested only on the breast region area, ignoring both the

[3] http://www.patreo.dcc.ufmg.br/deep-mammography-segmentation/.

background and pectoral data in the images. The best $\ddot{J}$ and $\ddot{D}$ values for this task were of 97.01% and 98.46%, respectively, again yielding highly precise segmentation predictions.



(a) (b)

Fig. 2. CIs for the MIAS dataset [17] in the pectoral muscle segmentation, breast region segmentation and background segmentation tasks and INbreast [12] in pectoral muscle segmentation. Vertical axis represents (a) $\ddot{J}$ and (b) $\ddot{D}$ metrics for $p \leq 0.05$. The lower end of the plot was trimmed at 75% to improve visualization of the CIs.

SegNet [1] and U-net [15] achieved similar results in most tasks and metrics, suggesting that both DNNs are good choices for these kinds of biomedical image segmentation tasks. The FCN architecture with a VGG16 basis achieved slightly lower results, but still was found to be significantly better than shallow approaches in most metrics and tasks.

Standard deviation values for DNNs were typically around one order of magnitude lower than the baseline, rendering deep segmentation schemes more reliable alternatives for fully automatic segmentation of mammograms. The only case where Rampun *et al.* [14] surpassed the deep architectures in the $\ddot{J}$ and $\ddot{D}$ metrics was in pectoral muscle segmentation, even though the standard deviations are considerably higher for the baseline. Results for the INbreast dataset [12] showed similar trends to the ones for the MIAS dataset [17]. DNNs outperformed the baseline for almost all metrics, but Specificity, reaching significantly better results than Rampun *et al.* [14] for all other metrics.

Comparisons with Other Baselines. It is hard to perform thorough comparisons between most papers in the area of mammographic image segmentation mainly due to the different evaluation metrics and the different datasets and subsets of images selected for most papers [14]. Many works use either a combination of FP and FN or custom metrics. One can see in Table 1 that deep methods achieved state-of-the-art-results also in FP and FN results, surpassing all methods in the FN metric with significantly better results. U-nets reached significantly better FP values compared to the best baseline (Ferrari *et al.* [3]), with the advantage of not compromising the FN metric, while SegNets and FCNs achieved comparable results in FP and vastly better results in FN. Also, while other methods in the state-of-the-art achieved $FP < 5\%$ and $FN < 5\%$ in between 50% and 60% of cases, FCNs, U-nets and SegNets reached this level of accuracy in 97.8%, 98.1% and 98.4% of images, respectively. Besides vastly larger percentages of images with $FP < 5\%$ and $FN < 5\%$, no DNN predictions had qualities worse than $min(FP, FN) < 5\%$ and $max(FP, FN) > 10\%$. Comparisons with other baselines can be found in the complementary material.

Table 1. FP and FN comparison with other baselines for pectoral muscle segmentation on MIAS [17]. FP and FN values are shown as percentages followed by standard deviation, when available. Values followed by % represent the percentage of images stratified according to the corresponding quality metric (row).

Metrics	FCN	U-net	SegNet	[2]	[3]	[8]
FP	0.68 ± 0.25	**0.53 ± 0.16**	0.62 ± 0.13	0.64	0.58	1.45
FN	0.39 ± 0.12	0.38 ± 0.12	**0.30 ± 0.9**	5.58	5.77	5.52
$FP < 5\%$ & $FN < 5\%$	97.8%	98.1%	98.4%	51.2%	53.6%	57.1%
$min(FP, FN) < 5\%$ & $5\% < max(FP, FN) < 10\%$	1.2%	1.2%	0.6%	22.6%	0%	33.3%
$min(FP, FN) < 5\%$ & $max(FP, FN) > 10\%$	0.9%	0.6%	0.9%	26.2%	0%	8.3%
$5\% < FP < 10\%$ & $5\% < FN < 10\%$	0%	0%	0%	0%	26.2%	0%
$5\% < min(FP, FN) < 0.10\%$ & $max(FP, FN) > 10\%$	0%	0%	0%	0%	0%	1.2%
$FP > 10\%$ & $FN > 10\%$	0%	0%	0%	0%	20.2%	0%

5 Conclusion

As far as the authors are aware, this paper reported the first use of Deep Learning for the segmentation of breast regions. We performed exhaustive tests using three DNN architectures for semantic segmentation and compared the results with state-of-the-art methods in the literature using several metrics and two publicly available datasets. In an effort to improve reproducibility and standardize the area, we only used objective evaluation metrics and provided a website containing several supplementary materials, including segmentation predictions, pretrained models and code for researchers to test the pretrained models in their own datasets. Even though the amount of data used in our experiments was suboptimal for deep methods, our experimental evaluation found that DNNs significantly surpassed the baselines in most cases and presented much better stability, that is, lower standard deviations. Data Augmentation techniques should improve DNN performance even further, as these methods tend to better converge with large sets of data. Most methods previously shown in the literature rely mostly on simple image processing filtering and segmentation and often stack several preprocessing and post-processing modules to the methodology. Despite their expensive training procedure, DNNs have a more plug-and-play nature, achieving state-of-the-art results with minimal pre and post-processing. Also the cost of forwarding images in pretrained DNNs is very computationally inexpensive, even without GPUs.

Future work includes a post-processing for shape regularization. Experimental evaluation also revealed that errors in distinct DNNs occurred in different places, therefore a late fusion scheme should improve the results.

Acknowledgments. Authors would like to thank NVIDIA for their support with GPUs; CAPES, CNPq and FAPEMIG (APQ-00449-17) for their financial support and Drs. Andrik Rampun and Arnau Oliver for the MIAS ground truths.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. TPAMI **39**(12), 2481–2495 (2017)
2. Camilus, K.S., Govindan, V., Sathidevi, P.: Pectoral muscle identification in mammograms. J. Appl. Clin. Med. Phys. **12**(3), 215–230 (2011)
3. Ferrari, R., Frere, A., Rangayyan, R., Desautels, J., Borges, R.: Identification of the breast boundary in mammograms using active contour models. Med. Biol. Eng. Comput. **42**(2), 201–208 (2004)
4. Ganesan, K., Acharya, U.R., Chua, K.C., Min, L.C., Abraham, K.T.: Pectoral muscle segmentation: a review. CMPB **110**(1), 48–57 (2013)
5. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The digital database for screening mammography. In: Digital Mammography, pp. 431–434 (2000)
6. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS, pp. 1097–1105. Curran Associates, Inc., New York (2012)
8. Li, Y., Chen, H., Yang, Y., Yang, N.: Pectoral muscle segmentation in mammograms based on homogenous texture and intensity deviation. Pattern Recogn. **46**(3), 681–691 (2013)
9. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440, June 2015
11. Lopez, M.G., et al.: BCDR: a breast cancer digital repository. In: ICEM (2012)
12. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: INbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012)

13. Oliver, A., Lladó, X., Torrent, A., Martí, J.: One-shot segmentation of breast, pectoral muscle, and background in digitised mammograms. In: ICIP (2014)
14. Rampun, A., Morrow, P.J., Scotney, B.W., Winder, J.: Fully automated breast boundary and pectoral muscle segmentation in mammograms. Artif. Intell. Med. **79**, 28–41 (2017)
15. 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
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Suckling, J., et al.: The mammographic image analysis society digital mammogram database. In: Exerpta Medica. International Congress Series, vol. 1069, pp. 375–378 (1994)