

Ultrasound Breast Lesion Detection using Extracted Attention Maps from a Weakly Supervised Convolutional Neural Network

Dalia Rodríguez-Salas¹, Mathias Seuret¹, Sulaiman Vesal¹, Andreas Maier¹

¹Pattern Recognition Lab, FAU Erlangen-Nürnberg
dalia.rodriguez@fau.de

Abstract. In order to detect lesions on medical images, deep learning models commonly require information about the size of the lesion, either through a bounding box or through the pixel-/voxel-wise annotation of the lesion, which is in turn extremely expensive to produce in most cases. In this paper, we aim at demonstrating that by having a single central point per lesion as ground truth for 3D ultrasounds, accurate deep learning models for lesion detection can be trained, leading to precise visualizations using Grad-CAM. From a set of breast ultrasound volumes, healthy and diseased patches were used to train a deep convolutional neural network. On the one hand, each diseased patch contained in its central area a lesion’s center annotated by experts. On the other hand, healthy patches were extracted from random regions of ultrasounds taken from healthy patients. An AUC of 0.92 and an accuracy of 0.87 was achieved on test patches.

1 Introduction

Early detection of high-risk breast lesions is crucial to prevent them from progressing to invasive diseases, and in turn breast cancer incidence and mortality can be likely reduced [1]. Although mammography imaging is the breast screening modality par excellence, breast ultrasound is often used as a supplementary screening modality and it provides a lower cost alternative for low- and middle-income countries. The increasing development of deep learning techniques for computer vision tasks in recent years has shown a clear raise on performance of automated breast lesion detection and classification systems [2].

In this work, we investigate the usage of a 3D Convolutional Neural Network (CNN) for detecting and localizing lesions in 3D ultrasound scans of breasts. The labels of our training data consists of lesion classes according to the American College of Radiology Breast Imaging Reporting and Data System (ACR BI-RAD) system, as well as an estimation of the lesion central point for volumes containing a lesion. We consider our system as weakly-supervised because while the shape and dimensions of lesions are absent from our training data, we aim at visualizing these nevertheless. Weak supervision has the advantage of requiring

less expensive ground truth, however this makes Machine Learning (ML) tasks more challenging.

For the detection of breast ultrasound lesions, more traditional machine learning techniques outnumber those using deep learning. Most of the works on lesion detection on breast imaging using deep learning techniques focus their efforts on 2D modalities such as mammogram and Magnetic Resonance Imaging (MRI).

Among the most recent works using traditional machine learning techniques, the one presented in [3] can be highlighted. They used a genetic algorithm to optimize a Random Forest (RF) classifier to generate candidate regions. A region is then classified as lesion depending on its probability built using pixel-level probabilities returned by the RF. They achieved a sensitivity of 84.4% on a dataset of 56 samples breast ultrasounds.

In [2], the performance of different deep learning network architectures for localization was compared when trained on breast ultrasounds to detect lesions. YOLO [4], R-CNN [5], and SSD [6] were the base architectures used in their experiments, where SSD trained with patches of size 300×300 reached the best performance. Despite of working with ultrasound imaging and deep learning models, their work is not directly comparable to ours because 1) they work with 2D images sampled from the 3D volumes by the specialist, and 2) they counted with expert’s annotations about the shape and size of the tumors.

Different works using deep learning techniques on breast ultrasound volumes target to classify whether the patient has a benign or malignant lesion. An overview of such works and a general overview of deep learning on ultrasound imaging can be found in [7].

The main contributions of this work are:

1. A system which detects and localizes lesions on breast ultrasound imaging.
2. A weakly supervised pipeline for lesion/abnormality detection and localization that can be extended to other medical imaging modalities.

The rest of the document is organized as follows: the materials and the description of the methodology for building our system is presented in section 2. The results of our experiments are shown in section 3. Lastly, an overall discussion of our results is drawn in section 4, including a subsection for our conclusion and possible paths for future work.

2 Materials and methods

The dataset used in our experiments consists of 904 volumes obtained from 461 patients, as shown in Table 1. Note that patients with at least one lesion are counted in the lesion class, regardless of whether they might have no lesions in one of their breasts.

The core of our system is a 3D CNN, more specifically a Residual Neural Network (ResNet) [8]. While there are different standard versions of the ResNet, having from 18 to 152 layers, the dimensionality of our data allows us to use only the smallest ones for memory reasons. Indeed, convolving the data with filters

Dataset	Total	Train		Test		Per Class	
		Healthy	Lesion	Healthy	Lesion	Healthy	Lesion
Patients	461	141	189	61	70	202	255
Samples	904	393	253	163	95	556	348

Table 1. Number of patients and samples in our dataset.

requires significantly more memory in 3D than in 2D. Because the default ResNet architecture has 1000 units in its output layer, we replace them with a single unit, which has a sigmoid activation on top to facilitate binary classification.

To train the CNN, using whole volumes is impracticable for three main reasons. First, as mentioned before, memory requirements would be extremely high. Second, loading such an amount of data from the hard drive would be too slow to train a CNN in a reasonable amount of time. And, third, this can lead to overfitting issues due to the small amount of training data that we have. Indeed, if the CNN is trained on large volumes, then it is more difficult for the network to learn what it is supposed to spot, and a greater number of training samples is then required. For these reasons, we train our CNN on 3D crops.

We get healthy crops by randomly sampling volumes labeled as healthy only, since there is no information about lesion size in the other volumes: crops containing only healthy tissues would have not guaranteed even if they had been extracted far from lesion centers. For positive samples, we extract crops centered on lesions. These crops are stored in new files, such that they can be obtained without loading the whole volumes in memory.

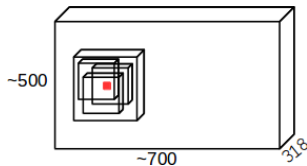


Fig. 1. Some variation is added in the training data by extracting random crops of size $16 \times 160 \times 160$ from a neighborhood around the lesion's patch sized $28 \times 280 \times 280$.

Two data augmentation methods are applied during the training. First, the crops that are stored in files, as described above, are created larger than what is given to the CNN, with a size of $28 \times 280 \times 280$. Random crops of the right size, i.e. $16 \times 160 \times 160$, are selected at run time, as illustrated in Fig. 1, which adds some variability to lesion locations during the training. Thus, the network learns to recognize lesions regardless of where in the input they are. The other augmentation which we use is flipping the data on horizontal and depth axes with a probability of 50%. This is justified by the fact that the shape of lesions is not constrained, and the CNN should have the same output regardless of their orientation.

The CNN is trained for 250 epochs, and thanks to the crop size, we can give it batches of 14 samples. We use the binary cross-entropy as loss function, and update the weights of the network using Adam [9]. In practice, we found that an initial learning rate of 0.001 performs well.

It is then possible to use a trained CNN to produce heatmaps indicating how likely each part of a given volume are to belong to a given class. In this work,

this is done with Gradient-weighted Class Activation Mapping (Grad-CAM) [10]. Given an input processed by the CNN, the gradient of an output class, in our case the lesion class, is back-propagated to the output of the last convolutional layer. Averaging these gradients for each location in the convolution indicates how much this location contributes to the class output. Thus, we obtain a heatmap which has the same dimension as the output of the last convolutional layer of the CNN. This heatmap can then be resized, with interpolation, to match the size of the input. The heatmap and the input are then added with a weight of 0.3 for the heatmap and 0.7 for the input. Lastly a colormap is applied to the image to produce hot-spots on the regions which are more likely to be lesions.

3 Results

In Table 2, the AUC and accuracy for a threshold of 0.5 are reported for both the test-set evaluated on the network trained with the complete training set, and the validation sets under a 5-fold cross validation process for which the average and standard deviation are reported. For the evaluation, the complete patch of $28 \times 280 \times 280$, as well as smaller patches circumscribed by it of size $16 \times 160 \times 160$, either with the centered in the lesion patch or in random locations were given to the network.

Table 2. Accuracy and AUC.

Central point location	Center (28x280x280)		Center (16x160x160)		Random (16x160x160)	
Patch size						
Metric	AUC Accuracy		AUC Accuracy		AUC Accuracy	
5-fold CV	0.8536 ± 0.0876	0.7433 ± 0.0836	0.9307 ± 0.0267	0.8743 ± 0.0260	0.9253 ± 0.0371	0.8744 ± 0.0291
Test set	0.9139	0.8411	0.9357	0.8798	0.9253	0.8721

In Fig. 2, the accuracy for the train- and test-set are plotted along the 250 epochs the network was trained for. On training patches, the two augmentation methods described in Section 2 were applied in each epoch. On testing patches, only random crops were extracted. The variation added to the testing patches was merely for experimental reasons, nevertheless, results without this variation are reported for the final model on Table 2.

Finally, in Fig. 3, the resulting heat-maps of using our network on patches of $9 \times H \times W$ as well as their overlapping (weighted addition) with their respective input is shown for four samples in the testing set. Where H and W are the height and width of the entire ultrasound volumes.

4 Discussion

Results in Table 2 show that the network performs better when using the smallest patches. This does not mean that the network is not capable of detecting lesions

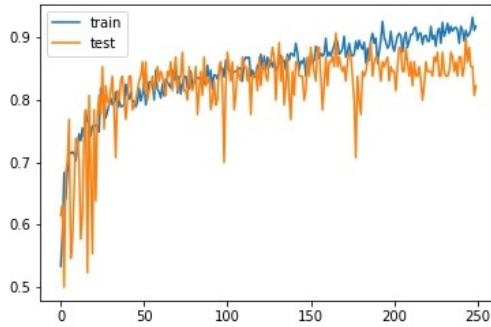


Fig. 2. Accuracy plot over 250 epochs for the training (blue) and test (orange) patches of size $28 \times 280 \times 280$. On training patches, random flip on the depth- and horizontal-axis as well as random crops of size $16 \times 160 \times 160$ were applied in each epoch. On test patches, only the random crop operation was used.

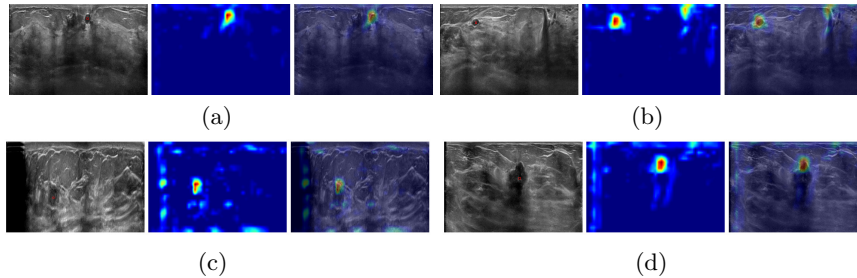


Fig. 3. Input, heat-map, and their overlapping for 4 ultrasounds in the test set. The displayed slice corresponds to the first coordinate of the estimated center of the lesion. In the input images, the center of the lesion which was annotated by experts is shown with a small red point.

on bigger patches. As shown in Fig. 3, even in much bigger patches, the network focuses its attention on regions which are more likely to belong to a lesion. This behavior is not unexpected, despite of being trained on much smaller patches, the network’s convolutional filters learn location variations of the input data whenever those variations were shown on the training samples. A clear example of this, can be found in [11], where a CNN is trained on small patches of text and the generated heat-map on a entire page, it is capable to distinguish text areas from background or illustrations. We attribute the decrease on accuracy to the average pooling layer on the ResNet architecture: it is expected that the larger is the crop the more is the area without possible lesion. This implies that relatively low values will come out from the average pooling layer, even if a lesion is present in the patch.

In Fig. 3, the heat-maps indicate that our network does not only pay attention to the lesion but also on different locations -commonly other black areas-; nevertheless, the major attention corresponds to the lesion area.

An adequate metric to evaluate the localization pipeline has to be developed as metrics such as Intersection over Union (IoU), or similar, make little sense when having a single point of the lesion as ground truth. For instance, in Fig.3-d, the hot-spot of the heatmap clearly shows the localization of the lesion; however, it does not overlap with the center of the lesion annotated by experts.

4.1 Conclusion and future work

In this paper, it was shown that an accurate CNN for patch-level classification of lesions in ultrasound volumes can be trained with the information of a single central point of the lesion. A key point of the pipeline resides on the augmentations used during the training phase: feeding the network with different random crops circumscribed in the extracted patches, gave the network enough information on the variability of the ultrasound volumes.

The generation of attention maps of our network using Grad-CAM provides information about the localization of the lesions. As future work, we plan to apply post-processing on the heat-maps to extract quantitative information of the size and localization of the lesions.

References

1. Morrow M, Schnitt SJ, Norton L. Current management of lesions associated with an increased risk of breast cancer. *Nature reviews Clinical oncology*. 2015;12(4):227.
2. Cao Z, Duan L, Yang G, et al. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC Med Imaging*. 2019 Jul;19(1):51. Available from: <https://doi.org/10.1186/s12880-019-0349-x>.
3. Torres F, Escalante-Ramirez B, Olveres J, et al. Lesion Detection in Breast Ultrasound Images Using a Machine Learning Approach and Genetic Optimization. In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer; 2019. p. 289–301.
4. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 779–788.
5. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2014. p. 580–587.
6. Chen LC, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell*. 2017;40(4):834–848.
7. Liu S, Wang Y, Yang X, et al. Deep learning in medical ultrasound analysis: a review. *Engineering*. 2019;5(2):261–275.
8. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 770–778.
9. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*. 2014;.
10. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Proceedings of the IEEE international conference on computer vision*; 2017. p. 618–626.
11. Seuret M, Limbach S, Weichselbaumer N, et al. Dataset of Pages from Early Printed Books with Multiple Font Groups. In: *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*; 2019. p. 1–6.