Large-scale classification of breast MRI exams using deep convolutional networks

Shizhan Gong[2,1], **Matthew Muckley**[1], **Nan Wu**[2], **Taro Makino**[1], **S. Gene Kim**[1], **Laura Heacock**[1], **Linda Moy**[1], **Florian Knoll**[1], **Krzysztof J. Geras**[1,2]
[1]Department of Radiology, NYU School of Medicine
[2]Center for Data Science, New York University

Abstract

In this paper we trained an end-to-end classifier using supervised deep convolutional neural network on a data set composed of a very large set of 8632 3D MR exams. Our best model can achieve an AUC of 0.8486 in identifying malignant cases on the test set reflecting the full spectrum of the patients who undergo the breast MRI examination. We also studied the effect of the data set size and the effect of using different T1-weighted images in the series on the performance of our model. Our goal is that this work will serve as a guideline for optimizing future deep neural networks for breast MRI interpretation.

1 Introduction

Breast cancer is the second leading cancer-related cause of death among women in the US. Magnetic resonance imaging (MRI) is one of the imaging modalities used for the diagnosis and monitoring of breast cancer, which has higher sensitivity in comparison to other modalities [10, 12]. However, reading and interpreting MR images is both labor and time intensive. Recently, the emergence of practical tools based on deep neural networks has generated enormous interest in the medical imaging community. These techniques can be adapted to improve the diagnostic and prognostic performance of breast MRI. In this paper, we trained and evaluated a set of neural networks on a breast MRI data set of an unprecedented scale. Our best model achieves an AUC of 0.8486 in identifying malignant cases in the test set reflecting the full spectrum of the patients who undergo the breast MRI examination.

Although automated radiomics systems for aiding radiologists in breast MRI interpretation have been under development for a long time, most previous approaches to this task involve two separate steps. In the first step, the location of the lesion is determined and the border of the lesion is delineated [4, 14, 5, 8], after which the extracted patches are fed to a final classifier [11, 2, 13, 1, 15]. However, this two-step approach has limitations. Since most segmentation methods can only detect mass-type lesions, it is difficult to analyze non-mass enhancement. Thus, lesion-level analysis may ignore some aspects of the data, which can be discovered with image-level analysis [6, 16]. In our paper, we trained an end-to-end classifier using a supervised deep convolutional neural network on 3D MR images. Therefore, we avoid time-consuming lesion annotation, which may also introduce subjective biases.

The goal of our research is to build models more accurate than those previously available in order to assist radiologists in interpreting breast MRI exams. By improving the quality of their assessments, we aim to decrease the number of false positive biopsies and short-term follow up exams, a known drawback of breast MRI [12].

2 Data

Our study was approved by the IRB and is compliant with the HIPAA. The data set is clinically realistic, and contains 8632 MRI exams from 6295 patients scanned between 2008 and 2018 at NYU

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Langone.[1] The data includes 1114 exams with biopsy-confirmed malignancies, 2148 exams with benign findings, and 918 exams with both benign and malignant findings. We randomly divided the patients into training (3789 patients, 5192 exams), validation (1251 patients, 1704 exams) and test (1255 patients, 1736 exams) subsets. The target labels were derived from pathology reports. For each breast, we assigned two binary labels separately indicating the presence of malignant and benign findings. Summed over the left and right breasts, each exam had a total of four labels.

3 Methods

We define the task of the breast MRI classification in the following manner. For each breast, we aim to predict two binary labels separately indicating the presence of malignant and benign findings. As input, we take one or more 3D MR images.

3.1 Model architecture

We use a novel neural network architecture called the Doubly Deep Convolutional Neural Network (see Figure 1). The input to the network is a 3D image, which is composed of several 2D slices. The sizes of 2D slices vary across different exams, and the network processes each 3D input image in two stages. In the first stage, an input image is divided into several 2D slices, and each slice is passed through a ResNet [7] of our design.[2] Each 2D slice is transformed into a 3D feature map, and a convolutional layer with single output channel is applied separately to all 3D feature maps to collapse them into 2D feature maps. With the network flow explained so far, we transformed 2D slices into a group of 2D feature maps whose sizes are significantly smaller than the original images. In the second stage, we assemble these 2D feature maps to form a 3D feature map. A 3D ResNet[3] of our design is then applied to the 3D feature map to generate four predictions corresponding to the four tasks.

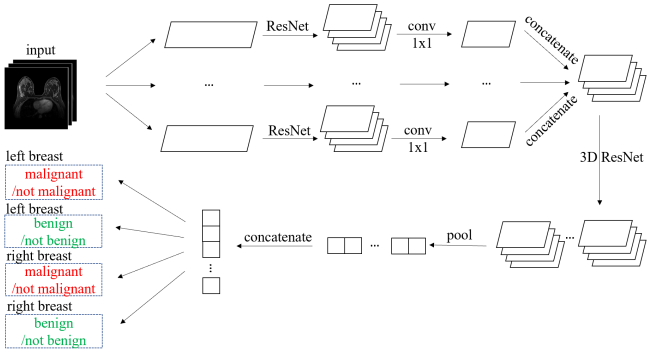

Figure 1: The Doubly Deep Convolutional Neural Network.

3.2 Training and evaluation

The loss function we use to train our model has the following form:

$$\mathcal{L}\left(\{(\mathbf{x}^n, y^n_{L,B}, y^n_{L,M}, y^n_{R,B}, y^n_{R,M})\}_{n\in\{1,\dots,N\}}\right) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{s\in\{L,R\}}\sum_{c\in\{M,B\}}\log(\hat{y}_{s,c}(\mathbf{x}^n)),$$

where M and B stand for malignant and benign, L and R represent the left and right breast, $y^n_{s,c}$ represents whether the n^{th} patient has a lesion of type c in her breast s, and $\hat{y}_{s,c}(\mathbf{x}^n)$ represents the predicted probability of the corresponding correct label. As it is framed above, we solve this learning task with multi-task learning. That way, we can utilize information from benign cases to help learning the task of discriminating between malignant and not malignant exams.

[1] All women underwent bilateral breast DCE-MRI using a 3.0T magnet (TimTrio, Siemens) or 1.5T magnet (Symphony, Siemens) in the prone position. The diagnostic protocol at our institution included both sagittal images and axial images. We collected fat-suppressed T1-weighted pre and three post-contrast acquisitions beginning 140 s - 760 s after the injection of a contrast agent (Magnevist or Gadavist). T1-weighted imaging parameters included: repetition time / echo time (TR/TE) = (3.57-8.90)/(1.04-4.76) msec, flip angle 8°-20°, slice thickness 1 - 1.8 mm, matrix (320 - 480) × (182 – 388).

[2] Its architecture is similar to ResNet-18 without the global average pooling layer and the softmax layer, except that the kernel size in the first convolutional layer is 5×5 and we remove all batch normalization layers (as we train with a minibatch size of one example).

[3] This network consists of one 3D residual block with $3 \times 3 \times 3$ convolutional kernel and 128 output channels followed by a global maximum pooling layer and four separate softmax layers.

The parameters of the network were learned using the Adam algorithm with a initial learning rate of 10^{-5}. Our training data was augmented with random flips and cropping of the original images. All augmentations were off during the validation and test phases. We trained the network for 50 epochs, allowing it to overfit, and picked the best model based on performance on the validation set.

Our primary evaluation metric was the AUC. For each of the malignant and benign prediction tasks, we evaluated predictions for both breasts independently, and pooled them to obtain a breast-level AUC. As detecting malignant lesions is more important than detecting benign lesions in clinical practice, we evaluated our model primarily with respect to breast-level AUC for predicting malignant lesions. We picked the best training epoch on the validation set using this metric.

4 Experiments

4.1 Impact of the training set size

First, we explored the relationship between data set size and test error when using the T1-weighted second post-contrast subtraction images. We trained separate networks on randomly sampled subsets of the full training set: 100%, 50%, 20%, 10%, 5%, 2% and 1% of the original training set. In Figure 2, we observed that the classification performance improves as the number of training examples increases, and that even with 100% of the data, the performance of the network has not yet saturated.

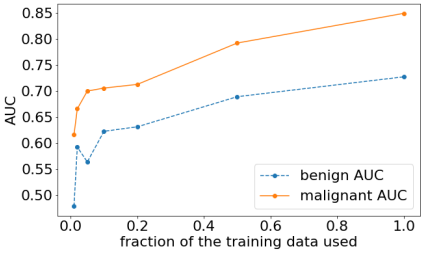

Figure 2: The effect of changing the fraction of the training data used on the performance on the test set.

4.2 Performance of different images in the series

There are four different T1-weighted images in the series for each exam, which were acquired at different times before (denoted by $\mathbf{x}_{t_0}$) or after (denoted by $\mathbf{x}_{t_1}$, $\mathbf{x}_{t_2}$ and $\mathbf{x}_{t_3}$) the injection of a contrast agent. In this set of experiments, we explored the predictive value of different images in the series and their subtractions (denoted by $\mathbf{x}_{t_1}$ - $\mathbf{x}_{t_0}$, $\mathbf{x}_{t_2}$ - $\mathbf{x}_{t_0}$ and $\mathbf{x}_{t_3}$ - $\mathbf{x}_{t_0}$). In order to do so, we trained separate networks on the training sets composed of different images in the sequence. As shown in Table 1, the model trained on the second post-contrast subtraction image has the best performance, which likely correlates to the greatest lesion conspicuity for lesions with plateau and wash-out temporal kinetics, which are more likely to be malignant [9].

4.3 Effects of model ensemble

We also tried ensembling the classifiers from the previous section to further improve the performance. When averaging the predicted probability using the first, second and third post-contrast subtraction images, the malignant AUC on the test set increased to 0.8497. Then we tried following the approach of Caruana et al. [3], to combine all seven models trained with different images in the series, which resulted in the malignant AUC on the test set decreasing to 0.8459. These results show that model ensembling brings us limited benefits. Therefore, we conclude that from the perspective of our neural networks, the information in different images is redundant to a large degree.

Table 1: The performance of different images in the series used for training.

image	AUC	
	benign	malignant
$\mathbf{x}_{t_0}$	0.6304	0.6714
$\mathbf{x}_{t_1}$	0.6055	0.6999
$\mathbf{x}_{t_2}$	0.6766	0.7620
$\mathbf{x}_{t_3}$	0.7038	0.7809
$\mathbf{x}_{t_1}$ - $\mathbf{x}_{t_0}$	0.7011	0.8135
$\mathbf{x}_{t_2}$ - $\mathbf{x}_{t_0}$	0.7267	0.8486
$\mathbf{x}_{t_3}$ - $\mathbf{x}_{t_0}$	0.7227	0.8418

5 Discussion

In this paper, we set a strong baseline for future research on deep neural networks for classification of breast MRI exams. We show that the performance of our neural network, although already strong, can be further improved by collecting more data. We also demonstrate a curious result showing that the information in different T1-weighted images in the series appears almost completely redundant.

There are a few possible improvements to our work. For example, we did not perform a systematic search for optimal hyperparameters, and our preprocessing of the original images was minimal. Changing both aspects could potentially significantly improve our results.

Acknowledgements

The authors would like to thank Catriona C. Geras for correcting earlier versions of this manuscript, Marc Parente for help with importing the image data and Mario Videna and Abdul Khaja for supporting our computing environment. We also gratefully acknowledge the support of Nvidia Corporation with the donation of some of the GPUs used in this research. This work was supported in part by grants from the National Institutes of Health (R21CA225175 and P41EB017183).

References

[1] Natalia O Antropova, Hiroyuki Abe, and Maryellen L Giger. Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. *Journal of Medical Imaging*, 5(1):014503, 2018.

[2] Hongmin Cai, Yanxia Peng, Caiwen Ou, Minsheng Chen, and Li Li. Diagnosis of breast masses from dynamic contrast-enhanced and diffusion-weighted MR: a machine learning approach. *PloS One*, 9(1):e87387, 2014.

[3] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM, 2004.

[4] Weijie Chen, Maryellen L Giger, and Ulrich Bick. A fuzzy c-means (fcm)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Academic Radiology*, 13(1):63–72, 2006.

[5] Mehmet Ufuk Dalmış, Suzan Vreemann, Thijs Kooi, Ritse M Mann, Nico Karssemeijer, and Albert Gubern-Mérida. Fully automated detection of breast cancer in screening MRI using convolutional neural networks. *Journal of Medical Imaging*, 5(1):014502, 2018.

[6] Christoph Haarburger, Michael Baumgartner, Daniel Truhn, Mirjam Broeckmann, Hannah Schneider, Simone Schrading, Christiane Kuhl, and Dorit Merhof. Multi scale curriculum CNN for context-aware breast MRI malignancy classification. *arXiv preprint arXiv:1906.06058*, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] P Herent, B Schmauch, P Jehanno, O Dehaene, C Saillard, C Balleyguier, J Arfi-Rouche, and S Jégou. Detection and characterization of MRI breast lesions using deep learning. *Diagnostic and Interventional Imaging*, 2019.

[9] Christiane Katharina Kuhl, Peter Mielcareck, Sven Klaschik, Claudia Leutner, Eva Wardelmann, Jurgen Gieseke, and Hans H Schild. Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211(1):101–110, 1999.

[10] Monica Morrow, Janet Waters, and Elizabeth Morris. MRI for breast cancer screening, diagnosis, and treatment. *The Lancet*, 378(9805):1804–1811, 2011.

[11] Ke Nie, Jeon-Hor Chen, J Yu Hon, Yong Chu, Orhan Nalcioglu, and Min-Ying Su. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. *Academic Radiology*, 15(12):1513–1525, 2008.

[12] Susan G Orel and Mitchell D Schnall. MR imaging of the breast for the detection, diagnosis, and staging of breast cancer. *Radiology*, 220(1):13–30, 2001.

[13] Zhiyong Pang, Dongmei Zhu, Dihu Chen, Li Li, and Yuanzhi Shao. A computer-aided diagnosis system for dynamic contrast-enhanced MR images based on level set segmentation and ReliefF feature selection. *Computational and Mathematical Methods in Medicine*, 2015, 2015.

[14] Ashirbani Saha, Michael R Harowicz, and Maciej A Mazurowski. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. *Medical Physics*, 45(7):3076–3085, 2018.

[15] Daniel Truhn, Simone Schrading, Christoph Haarburger, Hannah Schneider, Dorit Merhof, and Christiane Kuhl. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. *Radiology*, 290(2):290–297, 2018.

[16] Juan Zhou, Lu-Yang Luo, Qi Dou, Hao Chen, Cheng Chen, Gong-Jie Li, Ze-Fei Jiang, and Pheng-Ann Heng. Weakly supervised 3d deep learning for breast cancer classification and localization of the lesions in MR images. *Journal of Magnetic Resonance Imaging*, 2019.