Breast density classification with deep convolutional neural networks

Nan Wu, Krzysztof J. Geras, Yiqiu Shen, Jingyi Su, S. Gene Kim, Eric Kim,
Stacey Wolfson, Linda Moy & Kyunghyun Cho

NYU

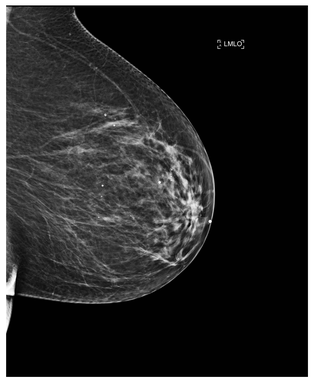
Abstract

Breast density classification is an essential part of breast cancer screening. Although a lot of prior work considered this problem as a task for learning algorithms, to our knowledge, all of them used small and not clinically realistic data both for training and evaluation of their models. In this work, we explored the limits of this task with a data set coming from over 200,000 breast cancer screening exams. We used this data to train and evaluate a strong convolutional neural network classifier. In a reader study, we found that our model can perform this task comparably to a human expert.

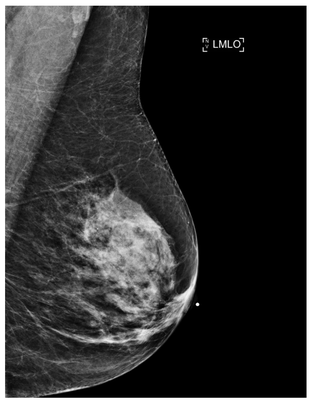
Data

We used a clinically realistic data set of over 200,000 screening mammography exams, each containing at least four images corresponding to the standard four views used in screening mammography Geras et al. [2017]. Each exam is assigned a BI-RADS label indicating a diagnosis of a radiologist. We supplemented this data with labels corresponding to breast density, which we extracted from the textual reports associated with the exams in our data set.

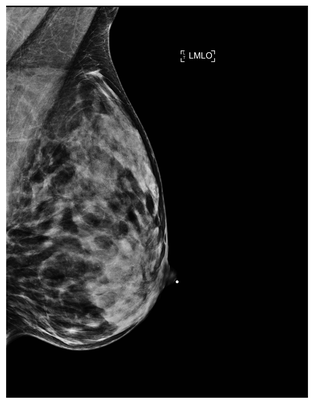
almost entirely fatty (0)



scattered areas of fibroglandular density (1)



heterogeneously dense (2)



extremely dense (3)

Examples of the four breast density classes.

Deep convolutional neural network

We used a multi-column deep convolutional neural network of an architecture loosely inspired by the earlier work of Simonyan and Zisserman [2015]. The input to the network is four **2600** $\times$ **2000** images corresponding to the standard views used in screening mammography. It is very similar to the architecture in Geras et al. [2017] with the exception of the number of the outputs in the softmax layer.

Classifier $p(y\|x)$			
Fully connected layer (1024 hidden units)			
Concatenation (256$\times$4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

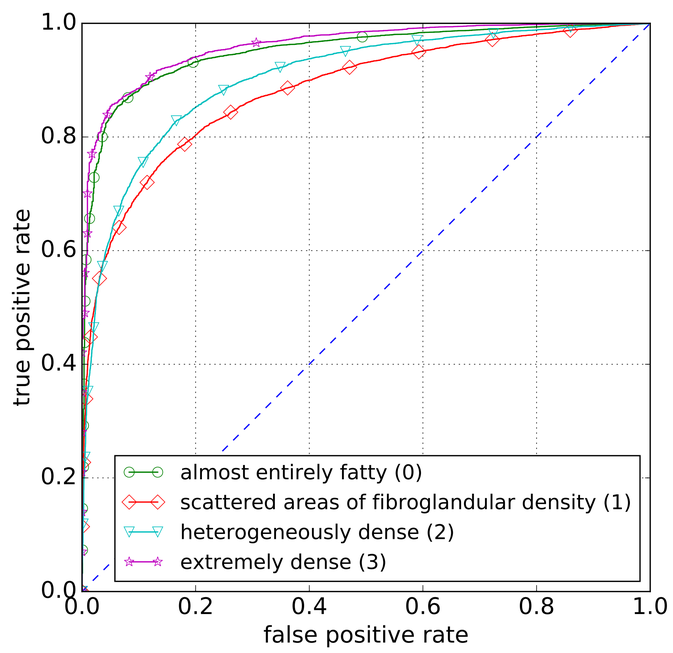
An overview of the structure of the convolutional neural network used in our experiments. DCN denotes a series of convolutional and pooling layers.

Experimental setup

We sorted the patients according to the date of their latest exam and divide them into training (first 80%), validation (next 10%) and test (last 10%) sets. For the test phase, we kept only the most recent exam for each patient. This way of partitioning the data allows us to estimate performance of our classifiers on future data accurately.

Our primary metric in this work is the standard classification accuracy. As the levels of breast density correspond to relative increases in the amount of fibroglandular tissue, two consecutive labels can be confused even by an experienced radiologist. This is why we also considered *top-k* accuracy. In this metric we consider a prediction to be correct if the ground truth is among the k most likely labels predicted. Additionally, we also considered accuracy only between the two superclasses: "dense" (classes 2 and 3) versus "not dense" (classes 0 and 1). Secondly, we evaluated our models with respect to the area under the ROC curve (AUC). We computed AUCs for all four binary problems of distinguishing between one of the density categories and the rest of the density categories, and then took the macro average, abbreviated as macAUC.

Results



ROC curves for all four classes. The classes 1 and 2 are the hardest for a neural network to distinguish from the rest. The AUC values are 0.955, 0.888, 0.907, 0.960 for classes 0, 1, 2, 3 respectively.

Impact of the size of the data set

To explore the effect of data set scale, we trained separate networks on training sets of different sizes; 100%, 10% and 1% of the original training set. Interestingly, even though training with more data increases performance in all metrics, the difference is not large.

Performance of our CNNs. The * symbol in the leftmost column indicates that a model was initialized using weights of a previously trained overall BI-RADS classifier.

data	macAUC	top-1	top-2	top-3	superclass
1%	0.888	0.729	0.967	0.998	0.849
10%	0.907	0.745	0.976	0.999	0.856
100%	0.916	0.767	0.982	0.999	0.865
*1%	0.892	0.733	0.974	0.998	0.848
*10%	0.909	0.753	0.980	0.998	0.856

Transferring knowledge from BI-RADS classifier

Considering the correlation between breast density and overall BI-RADS, we applied the idea of transfer learning to accelerate learning of our breast density prediction network. We used the weights of our model previously trained for breast cancer screening Geras et al. [2017] to initialize the parameters of the network trained for breast density prediction. The two networks have an identical architecture, except for the softmax layer.

The models trained with such initialization perform better than their counterparts trained from scratch in almost all metrics, however, only by a small margin. Intriguingly, models initialized with parameters of a previously trained overall BI-RADS classifier achieve the best performance in much fewer numbers of training epochs: 20 instead of 50 when using 1% of the original training data 15 instead of 25 when using 10% of the original training data.

Comparison to human performance

To understand what the limit of performance possible to achieve on this task is, we conducted a reader study with human experts with different levels of experience. The three participants in our reader study were: a medical student (S), a radiology resident (R) and an attending radiologist (A). The experts were all shown the same 100 exams randomly drawn from the test set. For each exam, the experts were asked to rank the breast density classes from the most likely to the least likely according to their judgement. Additionally, we computed analogous values with only two supercalsses.

Both human experts and learning models achieve a fair agreement with the labels in the data. Note that the agreement between the predictions of our model and the labels in the data are of similar magnitude to the agreement between the humans themselves.

We also compared our best CNN model to an average of the predictions of human experts. We achieved that by treating predictions of experts as one-hot vectors and averaging them. In this experiment the humans achieved macAUC of 0.892 (class 0: 0.960, class 1: 0.812, class 2: 0.807 and class 3: 0.990), while the CNN achieved macAUC of 0.934 (class 0: 0.971, class 1: 0.859, class 2: 0.905 and class 3: 1.000).

Agreement (Cohen's kappa) in choosing the most likely class between different readers (S, R, A), our neural network (N), our baseline (H) and labels in the data set (L).

	L	N	H	S	R	A
L		0.61	0.39	0.41	0.55	0.39
N			0.58	0.53	0.60	0.48
H				0.28	0.37	0.34
S					0.65	0.48
R						0.43

Agreement (Cohen's kappa) in distinguishing between dense breasts (classes 2 and 3) and not dense (classes 0 and 1) between different readers (S, R, A), our neural network (N), our baseline (H) and labels in the data set (L).

	L	N	H	S	R	A
L		0.65	0.50	0.50	0.73	0.46
N			0.72	0.62	0.83	0.57
H				0.48	0.69	0.48
S					0.69	0.64
R						0.60

References

Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, S. Gene Kim, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv:1703.07047v2*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.