High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks

Krzysztof J. Geras[a], **Stacey Wolfson**[c], **S. Gene Kim**[c,d], **Linda Moy**[c,d], **and Kyunghyun Cho**[a,b,1]

[a]Center for Data Science, New York University; [b]Courant Institute of Mathematical Sciences, New York University; [c]Center for Biomedical Imaging, Radiology, NYU School of Medicine; [d]Perlmutter Cancer Center, NYU Langone Medical Center

Recent advances in deep learning for object recognition in natural images has prompted a surge of interest in applying a similar set of techniques to medical images. Most of the initial attempts largely focused on replacing the input to such a deep convolutional neural network from a natural image to a medical image. This, however, does not take into consideration the fundamental differences between these two types of data. More specifically, detection or recognition of an anomaly in medical images depends significantly on fine details, unlike object recognition in natural images where coarser, more global structures matter more. This difference makes it inadequate to use the existing deep convolutional neural networks architectures, which were developed for natural images, because they rely on heavily downsampling an image to a much lower resolution to reduce the memory requirements. This hides details necessary to make accurate predictions for medical images. Furthermore, a single exam in medical imaging often comes with a set of different views which must be seamlessly fused in order to reach a correct conclusion. In our work, we propose to use a multi-view deep convolutional neural network that handles a set of more than one high-resolution medical image. We evaluate this network on large-scale mammography-based breast cancer screening (BI-RADS prediction) using 103 thousand images. We focus on investigating the impact of training set sizes and image sizes on the prediction accuracy. Our results highlight that performance clearly increases with the size of training set, and that the best performance can only be achieved using the images in the original resolution. This suggests the future direction of medical imaging research using deep neural networks is to utilize as much data as possible with the least amount of potentially harmful preprocessing.

neural networks | convolutional networks | deep learning | mammography

Breast cancer is the second leading cause of death among women in the United States. It is estimated that in 2015, 232 thousand (k) women were diagnosed with breast cancer and approximately 40k died from the disease (1). Screening mammography is the main imaging test used to detect occult breast cancer. Multiple randomized clinical trials have shown a 30% reduction in mortality in asymptomatic women who were undergoing screening mammography (1, 2). Although mammography is the only imaging test that reduced breast cancer mortality (3–6), the appropriate screening interval for mammograms has been the subject of public debate with different professional societies offering varying guidelines for mammographic screening (1, 3–9). In particular, there has been public discussion regarding the potential harms of screening. These harms include false positive recalls and false positive biopsies as well as anxiety caused by recall for diagnostic testing after a screening exam. Overall, the recall rate following a screening mammogram is between 10-15%. This equates to about 3.3 to 4.5 million callback exams for additional testing (10).

The vast majority of the women asked to return following an inconclusive mammogram undergo another mammogram and/or ultrasound for clarification. Most of these false positive findings are found to represent normal breast tissue with the additional imaging. Only 1% to 2% of women who have an abnormal screening mammogram are recommended to undergo a biopsy. Only 20-40% of these biopsies yield a diagnosis of cancer (11). In 2014, over 39 million screening and diagnostic mammography exams were performed in the US. Therefore, in addition to the anxiety from undergoing a false positive mammogram, there are significant costs associated with unnecessary follow ups and biopsies. Clearly, there is an unmet need to shift the balance of routine breast cancer screening towards more benefit and less harm.

Mammography-based Breast Cancer Screening

Screening mammography has been recognized as one of the leading contributors to reducing breast cancer mortality by 30% since the 1980s. Despite such successes, breast cancer remains the second leading cause of cancer death in women and the leading cause of death in women aged between 45 and 55 (2, 12). This fact illustrates key limitations in the sensitivity of current approaches to breast cancer screening largely based on mammography. The specificity of the screening mammography is also limited as the continued large fraction of false positives often leads to unnecessary follow up procedures including biopsies that cause considerable pain, anxiety and cost.

Significance Statement

Breast cancer is the second leading cause of cancer death in women in the United States. Although screening mammography has decreased breast cancer mortality, its specificity remains low. The development of deep convolutional networks (DCN) to aid in the evaluation of screening mammography would save significant health care costs. It is a common practice to use an existing DCN architecture, designed for natural images, to classify medical images. This approach often requires heavy downscaling of an image, which may hide details important for diagnosis. In this work, we propose a novel DCN architecture to efficiently handle an unprecedentedly large data set of high-resolution breast mammograms. We experimentally show that the image resolution and the data size have significant impact on the accuracy of breast cancer screening with deep learning.

Author contributions: L.M., S.G.K., K.C. and K.J.G. designed research; K.J.G. and K.C. performed research; K.J.G., K.C. and S.W. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: kyunghyun.cho@nyu.edu.

It is therefore necessary to improve diagnostic accuracy of mammography in terms of both sensitivity for early detection of breast cancer and specificity to keep the recall rate low.

Breast Cancer Screening as a Deep Learning Task. Deep learning has recently seen enormous success in challenging problems such as object recognition from natural images, automatic speech recognition and machine translation (13). This success has prompted a surge of interest in applying deep convolutional networks (DCN) to medical imaging. Many recent studies have shown the potential of applying such networks to medical imaging, including breast screening mammography; however, without investigating the fundamental differences between medical and natural images and their impact on the design choices and performance of proposed neural networks. For instance, much of recent works have either significantly downscaled a whole image or solely focused on classifying a small region of interest. This might be detrimental to performance of such models given the well-known dependency of breast cancer screening on fine details, such as the existence of a cluster of microcalcifications, as well as global structures, such as the symmetry between two breasts. Furthermore, the potential of DCN's has only been assessed in limited settings of small data sets often consisting of less than 1k images, while the success of such networks in natural object recognition is largely attributed to the availability of more than 1M annotated images. This further hinders our understanding of the true potential of DCN's in medical imaging, particularly in breast cancer screening.

In this work, we conduct an investigation into analyzing and understanding fundamental properties of deep convolutional networks in the context of breast cancer screening. We start by building a large-scale data set of approximately 23k screening mammographic exams (103k images) collected at multiple sites of our institution. We develop a novel DCN that is able to handle multiple views of screening mammography and to utilize large high-resolution images without downscaling. We refer to this DCN as a multi-view deep convolutional network (MV-DCN). Our network learns to predict the assessment of a radiologist, classifying an incoming example as BI-RADS 0 ("incomplete"), BI-RADS 1 ("normal") or BI-RADS 2 ("benign finding"). We study the impact of the data set size and image resolution on the screening performance of the proposed MV-DCN, which would serve as a *de facto* guideline for optimizing future deep neural networks for medical imaging. We further investigate the potential of the proposed MV-DCN by visualizing predictions made by MV-DCN.

High-Resolution Multi-View Deep Convolutional Neural Networks

Deep Convolutional Neural Network. A deep convolutional neural network (14, 15) is a classifier that takes an image $\mathbf{x}$ as input, often with multiple channels corresponding to different colors (e.g., RGB), and outputs the conditional probability distribution over the categories $p(y|\mathbf{x})$. This is done by a series of nonlinear functions that gradually transform the input pixel-level image. A major property of the deep convolutional network, which distinguishes it from a multilayer perceptron, is that it heavily relies on convolutional and pooling layers, which make the network invariant to local translation of visual features in the input.

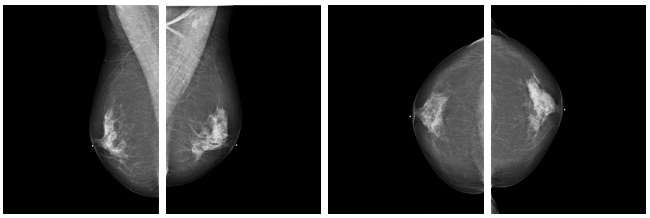

Fig. 1. The four views used in our experiments.

Multi-View Deep Convolutional Neural Network. Unlike an object recognition task with natural images, an exam in medical imaging often comes with a set of views. For instance, it is standard in screening mammography to obtain two views, cranial caudal (CC) and mediolateral oblique (MLO), for each breast of a patient, resulting in a set of four images. We will refer to them as L-CC, R-CC, L-MLO and R-MLO (these are illustrated in Figure 1).

There is a rich literature in building a deep neural network for multi-view examples. Most of them fall into one of two major families. First, there are works on unsupervised feature extraction from multiple views using a variant of deep autoencoders (16–18). They usually train a multi-view deep neural network with unlabeled examples, and use the output of such a network as a feature extractor, followed by a standard classifier. On the other hand, (19) proposed to build a multi-view deep convolutional network directly for classification.

We propose a variant of multi-view deep convolutional network which was motivated by (19). This multi-view deep convolutional network computes the output in two stages. In the first stage, a number of convolutional and pooling layers is separately applied to each of the views. We denote such view-specific representation by $\mathbf{h}_{\mathrm{v}}$, where v refers to the index of the view. These view-specific representations are concatenated to form a vector, $[\mathbf{h}_{\mathrm{L-CC}}, \mathbf{h}_{\mathrm{R-CC}}, \mathbf{h}_{\mathrm{L-MLO}}, \mathbf{h}_{\mathrm{R-MLO}}]$, which is an input to the second stage - a softmax layer producing output distribution $p(\mathbf{y}|x)$.

The whole network is trained jointly by stochastic gradient descent with backpropagation (20). Furthermore, we employ a number of regularization techniques to avoid the behavior of overfitting due to the relatively small size of training dataset, such as data augmentation by random cropping (21) and dropout (22). These will be describe later in detail.

High-Resolution Convolutional Neural Network. It is common in object recognition and detection in natural images to heavily downscale an original high-resolution image. For instance, the input to the deep convolutional network by the best performer of ImageNet Challenge 2015 (classification task) was an image downscaled to 224×224 (23). This is often done in order to improve the computational efficiency, both in terms of computation and memory, and also because no significant improvement has been observed with higher-resolution images. It reflects an inherent property of natural images, in which the objects of interest are usually presented in relatively larger portions than other objects and what matter most are their macro-structures, such as shapes, colors and other global descriptors. However, downscaling of an input image is not desirable in the case of classifying medical images, and in particular for early-stage screening based on breast mammography. Often a cue for diagnosis is a subtle finding which may

be identified only at the original resolution.

In order to address the computational issues of handling full-resolution images, we propose to use aggressive convolution and pooling layers. First, we use convolution layers with strides larger than 1 in the first two convolutional layers. Also, the first pooling layer has a larger stride than the other pooling layers. Thereby, we greatly reduce the size (width and height) of feature maps early in the network. Although this aggressive convolution and pooling loses some spatial information, the parameters of the network are adjusted to minimize this information loss during training. This is unlike downscaling of the input, which loses information unconditionally. Second, we take the average of output vectors from the last feature map instead of concatenating them (24), which has been a more common practice (21, 25). This drastically reduces the dimensionality of the view-specific vector without much, if any, performance degradation (26). Using both of these approaches, we are able to build a multi-view deep convolutional neural network that takes four 2600×2000 pixels images (one per view) as input without any downscaling.

Related Work

In this section, we briefly review recent deep learning based approaches to breast mammography. We summarize these recent works in Table 1.

Multi-Stage vs. End-to-End Approaches. Traditionally disjoint breast cancer screening or lesion detection is done in three stages: detection, analysis and final assessment/management. In the first stage, a breast mammography image is segmented into different types of regions, such as foreground (breast) and background. Within the segmented region of breast, the second stage focuses on extracting a set of regions of interest (ROI) that will be examined in more detail. In the third stage, each of those ROI's is determined to be a malignant lesion or not. The outcome of the third stage is used to make the final decision on a given case consisting of multiple views.

Most of the recent research on applying deep learning to breast mammography have focused on replacing one or more stages in this existing multi-stage pipeline; for instance, mass detection (27–29). In their work, a deep neural network is trained to determine whether a small patch is a mass. Others have focused on training a deep neural network for classifying a small region of interest into one of a few categories, assuming an existing mass detection system (9, 30–32).

On the other hand, a small number of research groups have considered replacing the whole multi-stage approach with a single, or a series of, trainable machine learning algorithms. Kooi et al. (33) proposed to use a random forest classifier for mass detection followed by a deep convolutional network that classifies each detected mass. A similar approach was proposed by Becker et al. (34). Akselrod-Ballin et al. (35) further proposed to use deep convolutional networks for both mass detection and classification, potentially enabling end-to-end training. Two groups (36, 37) went even further by proposing a single deep convolutional network that classifies a whole image, or a set of multiple views. The work by Carneiro et al. (37) is closest to our approach in this paper. In both of the works, a single deep convolutional neural network takes as input a set of multiple views of an exam and predicts its BI-RADS label.

Data Size. Although it is recognized that one of the driving forces behind the success of deep learning is the availability of large scale data, this has not been well exploited when applying deep learning to breast mammography screening. As evident in Table 1, most of the recent works use less than 1k images for both training and testing. In order to avoid the issue of small training data, most of the earlier works resorted to training a deep neural network with many small patches, or ROI's, avoiding end-to-end training. One exception is (37) in which Carneiro et al. use the whole image, however, with the deep convolutional network pretrained for object recognition in natural images. Unlike these earlier approaches, we collect and use a large-scale data set of an unprecedented size, consisting of 103k images. This allows us to carefully study the impact of the size of training data set.

Natural vs. Controlled Distribution. Breast screening is aimed at a general population rather than a selected group of patients. This implies that the distribution of the screening outcome is heavily skewed toward "normal" (BI-RADS 1). In our training set which closely follows a general population distribution, approximately half of the cases were assigned BI-RADS 1 ("normal"), while 35% were assigned BI-RADS 2 ("benign finding") and 15% BI-RADS 0 ("incomplete"). This is in contrast to two widely-used, publicly available datasets, INBreast (38), DDSM (39, 40) and other curated small-scale datasets from recent literature (see those in Table 1). These datasets are often constructed to include approximately the same proportions of normal and abnormal cases, resulting in, what we refer to as, a *controlled distribution* of outcomes which differs from a *natural distribution*. For instance, INBreast has approximately achieved a balance between benign and malignant cases. This type of artificial balancing, or equivalently upsampling of malignant cases, may bias a model to more often predict a given case as malignant and require a recall more often than necessary. Unlike these earlier works, in this paper, we use the full data without artificial balancing of outcomes to ensure that any trained deep convolutional network will closely reflect the natural distribution of outcomes.

Data

Collection. This is a Health Insurance Portability and Accountability (HIPAA)-compliant, retrospective study approved by our Institutional Review Board. Consecutive screening mammograms for 17,946 patients aged* between 19 and 99 (mean: 55.6, std: 11.8) collected within two years (2013-2014) at three imaging sites affiliated with New York University School of Medicine were used in this study. These imaging centers are located in the New York City metropolitan area (a large academic center and two large ambulatory care practices), where, altogether, over 70k mammograms are performed annually. The ethnic makeup of the patient cohort for this study reflects the population pool in NYC, which is 50% Caucasian, 30% African American, 5% Asian and 15% Hispanic.

Data Statistics. There are altogether 17,946 patients, 23,345 exams and 102,800 images in the data set. We divide the data into training, validation and test sets in the following manner. First we sort data for all patients according to the date of the

*When more than one exam for a patient was in the data set, we included the ages of that patient at the time of all exams to compute the values above.

Table 1. We summarize the most relevant previous works that apply deep learning to breast mammography. When more than one data set was used, we list the size of the largest data set. ⋆ denotes this paper. The table should be read with the following footnotes. □ The target task; BI-RADS: BI-RADS prediction, lesion: lesion classification (benign vs. malignant), mass: mass detection, and MC: microcalcification detection. • Whether the proposed system is trainable end-to-end. For instance, a system that requires an external system for extracting regions of interest (ROI) is not end-to-end, while a system that uses convolutional networks for both ROI extraction and lesion classification is. † In the parentheses there is the number of test images or "CV" if cross-validation is used. ♡ Whether multiple views per one exam are utilized. ○ Whether the data reflects natural distribution (N) or controlled distribution (C). ♠ Whether the input to a deep neural network is a whole image (IMG) or a small subset (ROI). ♣ Did not use images as input to the learning algorithm.

task□	ref.	E2E•	#images†	image size	MV♡	input♠	dist.○
BI-RADS	⋆	√	95k (8k)	2.6k×2k	√	IMG	N
	(37)	√	680 (≈ 340)	264×264	√	IMG	C
	(36)	√	410 (CV)	224×224	√	IMG	C
	(35)	√	850 (≈ 170)	800×800		IMG	C
lesion	(30)		607 (CV)	512×512		ROI	C
	(33)		44k (18k)	250×250		ROI	N
	(31)		1820 (182)	224×224		ROI	C
	(32)		736 (≈300)	150×150		ROI	C
	(9)		1606 (≈378)	13×13	√	ROI	N
mass	(27)		116 (CV)	32×32		ROI	C
	(28)		410 (CV)	264×264		ROI	C
	(29)	√	2500 (250)	256×256		ROI	C
MC	(41)		1000 (204)	N/A♣		ROI	C
	(42)		1410 (N/A)	N/A♣		ROI	C

last visit of the patient. We use the first 80% of the patients in this order as the training data, the next 10% as the validation data and the last 10% as the test data. For each patient in the test set, we evaluate our model's performance in predicting only the label for the last visit of each patient. This reflects best the way the model would be used once deployed.

Table 2. Distribution of data associated with different BI-RADS in training, validation and test data. Each cell in the table has the following format: number of exams / number of images.

	BI-RADS 0	BI-RADS 1	BI-RADS 2
Training	2866 / 12517	9295 / 40852	6588 / 29355
Validation	304 / 1314	1428 / 6232	1068 / 4676
Test	203 / 885	910 / 3964	683 / 3005

Data preprocessing and augmentation. We normalize the images in the following way. For each image we compute the mean, μ, and the standard deviation, σ, of its pixels. We then subtract μ from each pixel and divide each pixel by σ. Additionally, we flip horizontally the images of R-CC and R-MLO views so that the breast is always on the same side of the image. Since the images vary in size, we crop all of them to the size of 2600×2000 pixels. The position of the crop is determined in the following manner. First, the crop area is placed leftmost on the horizontal axis and centrally on the vertical axis. Then noise is added to this position. Let us denote the number of pixels between the top border of the crop area and the top border of the image by b_{top} and analogously define b_{bottom} and b_{right}. We draw a number, t_{vertical} from a uniform distribution $\mathcal{U}(-\min(b_{\text{top}}, 100), \min(b_{\text{bottom}}, 100))$ and $t_{\text{horizontal}}$ from $\mathcal{U}(0, \min(b_{\text{right}}, 100))$. Finally we translate

layer	kernel size	stride	#maps	repetition
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

Fig. 2. Description of one deep convolutional network column for a single view. It transforms the input view (a gray-scale image) into a 256-dimensional vector.

the position of the crop area by $t_{\text{horizontal}}$ pixels horizontally and t_{vertical} pixels vertically. During training this noise is sampled independently every time an image is used. During validation there is no noise added to the position of the crop area. At test time, we feed ten sets of four randomly cropped views to the network. The final prediction is made by averaging predictions for all crops. The aim of this averaging is twofold; first, to use information from outside the center of the image while keeping the size of the input fixed and second, to make prediction of the network more stable. A small fraction of data contains more than one image per view. For such cases one image per view is sampled randomly uniformly each time an exam is used during training and testing. During validation the image with the earliest time stamp is always used.

Settings

Evaluation Metrics. When there are two classes the most frequently applied performance metric is the AUC (area under the ROC curve). However, since there are three classes in our learning task, we cannot apply this metric directly. Instead we compute three AUCs, each time treating one of the three classes as a positive class and the remaining two as negative. We use the macro average of the three AUCs, abbreviated as macAUC, as the main performance metric in this work.

Unlike other widely used nonlinear classifiers, such as a support vector machine or a random forest, a deep convolutional neural network outputs a proper conditional distribution $p(y|\mathbf{x})$. It allows us to compute the network's confidence in its prediction by computing the entropy of this distribution, i.e.,

$$H(y|\mathbf{x}) = -\sum_{y' \in \mathcal{C}} p(y'|\mathbf{x}) \log p(y'|\mathbf{x}), \qquad [1]$$

where y' iterates over all possible classes $\mathcal{C}$. The larger the entropy, the less confident the network is about its prediction. Based on H, we can quantify the change in accuracy (measured by AUC) with respect to the network's confidence. We refer to AUC computed with the most confident 30% of test examples as HC-AUC and we define HC-macAUC to be the average of the three HC-AUC's.

Model Setup. The overall architecture of our network is shown in Figure 3. Each of the four columns corresponding to different views has an architecture described in Figure 2. In addition

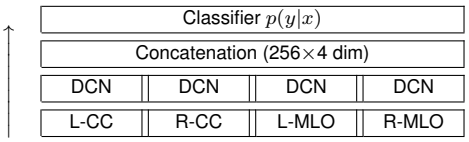

Fig. 3. An overview of the proposed multi-view deep convolutional network. DCN refers to the convolutional network network column from Figure 2. A single DCN is shared across the four views: L-CC, R-CC, L-MLO and R-MLO. The arrow indicates the direction of information flow.

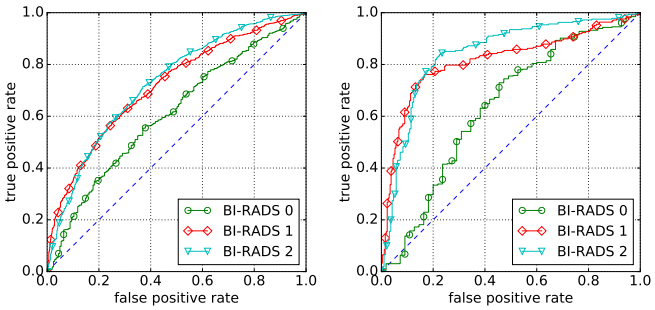

Fig. 4. ROCs computed with all test data (left) and ROCs computed with test data which the network was confident about (right).

to augmenting the data set by cropping the images at random positions, we regularize the network in three ways. First, we tie the weights in the corresponding columns, i.e., the parameters of the columns processing L-CC and R-CC views are shared as are those of the columns processing L-MLO and R-MLO views. Second, we add Gaussian noise to the input (with the mean of zero and the standard deviation of 0.01). Third, we apply dropout (with a rate of 0.2) after each pooling layer. We turn off the input noise and dropout during validation and testing.

The parameters of the network are initialized using the recipe of Glorot & Bengio (43) and learned using the Adam algorithm (44) with the initial learning rate of 10^{-5}. Due to the memory limitations of our hardware, the mini-batch size is set to four. We train the network for up to 100 epochs. After each training epoch we compute the macAUC on the validation set. We report the test error of the model which achieved the lowest macAUC on the validation set.

Quantitative Results Analysis

Effect of Scale. First, we validate our earlier claim on the need of large-scale data for harnessing the most out of deep convolutional neural networks. We train separate networks on the training sets of different sizes; 100%, 50%, 20% and 10% of the original training set.

In Table 3, we observe that the classification performance improves as the number of training examples increases. This shows the importance of using a large training set for training a deep neural network. This is consistent with observations made in many other fields such as computer vision, natural language processing and speech recognition (13).

Effect of Resolution. We then investigate the effect of resolution of input images. Using the full training set, we train networks with varying input resolutions; scaling both dimensions of the input by $\times 1/8$, $\times 1/4$ and $\times 1/2$. We use bicubic

Table 3. The effect of changing the fraction of the training data used. Increasing the amount of data yields better results. Further improvements with even more training data is to be expected.

fraction	10%	20%	50%	100%
0 vs. others	0.501	0.571	0.587	0.609
1 vs. others	0.559	0.614	0.69	0.717
2 vs. others	0.555	0.631	0.702	0.728
macAUC	0.538	0.605	0.659	0.685
HC-macAUC	0.584	0.618	0.702	0.765

Table 4. The effect of decreasing the resolution of the image.

scale	$\times 1/8$	$\times 1/4$	$\times 1/2$	$\times 1$
0 vs. others	0.577	0.565	0.575	0.609
1 vs. others	0.599	0.614	0.679	0.717
2 vs. others	0.613	0.617	0.703	0.728
macAUC	0.596	0.599	0.652	0.685
HC-macAUC	0.656	0.643	0.701	0.765

interpolation to downscale the input. When the input resolution is significantly smaller than the original image some convolutional layers in the later stages cannot be applied because the size of the feature maps becomes smaller than the size of a convolutional kernel. In that case, we simply skip the remaining layers until the global average pooling. As shown in Table 4, we already see a drop in performance when each dimension of the input was downscaled by half. Further degradation of performance was observed with more aggressive downscaling.

Confidence. We measure confidence of predictions in terms of the entropy of the output distribution (cf. Equation 1). As shown in Table 5, we observe that confident predictions of the proposed model are more accurate. This phenomenon was apparent in in all the experiments (see Table 3, Table 4 and Figure 4).

Visualization

A flip side of high effectiveness of a deep convolutional neural network is the difficulty in interpreting its internal processing. Although we have full access to its internal working, it is non-trivial to understand it. Only recently there have been some efforts on visualizing deep convolutional neural networks for computer vision (45, 46). These recent approaches however are not computationally efficient and are not easy to apply to medical images due to a number of reasons, including the need for training with a large data set (45) and the availability of good image statistics (46). Instead, we propose a simpler visualization technique in this paper that does not require any further training.

We look at the sensitivity of the network's output to the

Table 5. The accuracy (macAUC) using the top-P% validation examples according to entropy (examples balanced according to the prior distribution). The lower the P, the more confident predictions. When $P = 30\%$, we refer to the macAUC as a high-confidence macAUC (HC-macAUC). When reporting HC-AUC on the test set, we use the cut-off entropy values computed on the validation set.

P	10%	20%	**30%**	50%	100%
HC-macAUC	0.839	0.8	0.766	0.735	0.687

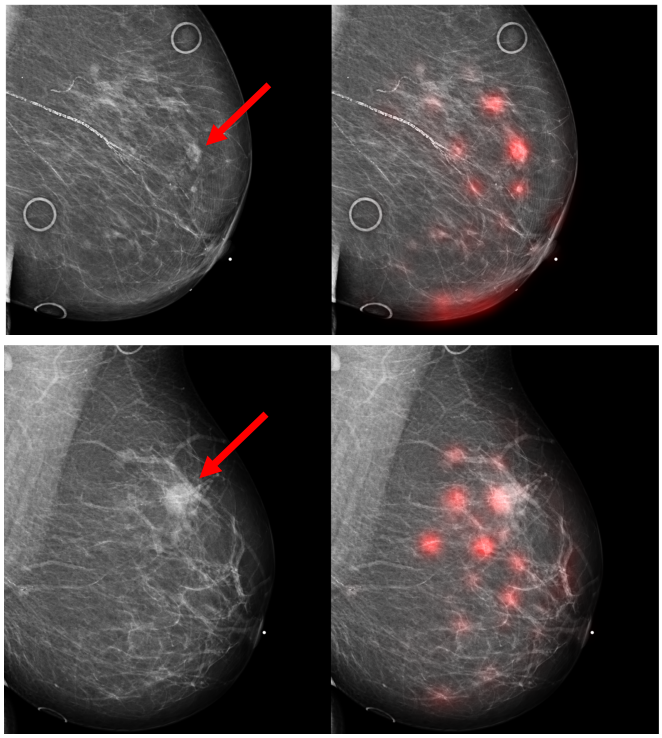

Fig. 5. On the left there are images of breast of two different patients confirmed to have breast cancer by biopsy. The patients were 88 (top) and 54 (bottom) years old. The red arrows indicate the suspicious finding which proved to be malignant. Both cancers were invasive ductal carcinoma. On the right there are the same images with regions of the images (highlighted in red) which influence confidence of predictions of our neural network. Please note that our visualization highlights parts of the image that are relevant for all classes (BI-RADS 0, BI-RADS 1, and BI-RADS 2) and that those highlighted areas include the biopsy locations shown on the left images

perturbation of each input pixel. The network outputs the conditional distribution over all the categories, and we can measure the entropy (or confidence) of the predictive distribution $\mathcal{H}(y|\mathbf{x})$. We can use standard backpropagation to compute $\left|\frac{\partial \mathcal{H}}{\partial \mathbf{x}_{ij}^{v}}\right|$ for the pixel (i,j) of the v-th view. Those input pixels that influence the confidence of the network will have high values, and those that do not contribute much will have low values (≈ 0). We show two examples of such visualization for patients which were confirmed by a follow-up examination to have breast cancer in Figure 5.

Conclusion

In this paper we have made a first step towards end-to-end large scale training of multi-view deep convolutional networks for breast cancer screening. We have shown experimentally that it is essential to keep the images at high-resolution. We expect that the same result would hold for other learning tasks with medical images where fine details determine the outcome. We also demonstrated that it is necessary to use a large number of exams. Although we used the largest breast cancer screening data set ever reported in literature, the performance of our model has not saturated and is expected to improve with more data.

Our network's performance was lowest on differentiating BI-RADS 0 from the other classes. We attribute it to the issue of noise in BI-RADS labeling. Doctors often disagree on how a particular exam should be classified (3) and in fact, less than 1% of the screening population has cancer (1, 8). To alleviate this problem we can instead use the information on whether a person actually went on to develop breast cancer in the future as a label.

It is also worth noting that, because of limited computational resources, we had to heavily rely on our experience in the choice of learning hyperparameters. We did not perform a systematic search for optimal hyperparameters, which often has a great impact on the performance of a neural network in limited data scenarios (47, 48). The methods we used in this work are powerful and our results can be improved simply by the means of applying more computational resources without significantly changing the methodology.

ACKNOWLEDGMENTS. We would like to thank Jure Žbontar, Yann LeCun, Artie Shen and Pablo Sprechmann for insightful comments on this work. K.C. thanks support by eBay, Facebook, Google and NVIDIA. L.M. and S.G.K. were partly supported by the NIH grant R01CA160620.

1. Duffy SW, et al. (2002) The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties. *Cancer* 95(3):458–69.
2. Jemal A, et al. (2009) Cancer statistics, 2009. *CA Cancer J Clin* 59(4):225–49.
3. Kopans DB (2002) Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 94(2):580–1; author reply 581–3.
4. Duffy SW, Tabar L, Smith RA (2002) The mammographic screening trials: commentary on the recent work by olsen and gotzsche. *CA Cancer J Clin* 52(2):68–71.
5. Hendrick RE, Smith RA, Rutledge, J. H. r, Smart CR (1997) Benefit of screening mammography in women aged 40-49: a new meta-analysis of randomized controlled trials. *J Natl Cancer Inst Monogr* (22):87–92.
6. Nelson HD, et al. (2009) *Screening for Breast Cancer: Systematic Evidence Review Update for the US Preventive Services Task Force*, U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. (Rockville (MD)).
7. Siegel RL, Miller KD, Jemal A (2015) Cancer statistics, 2015. *CA: a cancer journal for clinicians* 65(1):5–29.
8. Tabar L, et al. (2001) Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 91(9):1724–31.
9. Mordang JJ, et al. (2016) Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks in *International Workshop on Digital Mammography*. (Springer), pp. 35–42.
10. Tosteson AN, et al. (2014) Consequences of false-positive screening mammograms. *JAMA internal medicine* 174(6):954–961.
11. Kopans DB (2015) An open letter to panels that are deciding guidelines for breast cancer screening. *Breast Cancer Res Treat* 151(1):19–25.
12. Society AC (2009) Breast cancer facts & figures 2009-2010, (American Cancer Society, Inc.), Report.
13. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.
14. LeCun Y, et al. (1989) Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551.
15. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.
16. Ngiam J, et al. (2011) Multimodal deep learning in *International conference on machine learning*. pp. 689–696.
17. Srivastava N, Salakhutdinov R (2012) Multimodal learning with deep boltzmann machines in *Advances in neural information processing systems*. pp. 2222–2230.
18. Wang W, Arora R, Livescu K, Bilmes J (2015) On deep multi-view representation learning in *International Conference on Machine Learning*. pp. 1083–1092.
19. Su H, Maji S, Kalogerakis E, Learned-Miller E (2015) Multi-view convolutional neural networks for 3d shape recognition in *Proceedings of the IEEE International Conference on Computer Vision*. pp. 945–953.
20. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error-propagation in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1*. (MIT Press, Cambridge, MA) Vol. 1, pp. 318–362.
21. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks in *Advances in neural information processing systems*. pp. 1097–1105.
22. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.
23. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition in *IEEE Conference on Computer Vision and Pattern Recognition*.
24. Lin M, Chen Q, Yan S (2013) Network in network in *International Conference on Learning Representations*.
25. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition in *International Conference on Learning Representation*.
26. Szegedy C, et al. (2015) Going deeper with convolutions in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1–9.
27. Domingues I, Cardoso JS (2013) Mass detection on mammogram images: a first assessment of deep learning techniques.

28. Dhungel N, Carneiro G, Bradley AP (2015) Automated mass detection in mammograms using cascaded deep learning and random forests in *International Conference on Digital Image Computing: Techniques and Applications*. (IEEE), pp. 1–8.
29. Ertosun MG, Rubin DL (2015) Probabilistic visual search for masses within mammography images using deep learning in *IEEE International Conference on Bioinformatics and Biomedicine*. (IEEE), pp. 1310–1315.
30. Huynh BQ, Li H, Giger ML (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging* 3(3):034501–034501.
31. Lévy D, Jain A (2016) Breast mass classification from mammograms using deep convolutional neural networks. *arXiv:1612.00542*.
32. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Lopez MAG (2016) Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine* 127:248–257.
33. Kooi T, et al. (2017) Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis* 35:303–312.
34. Becker AS, et al. (2017) Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative Radiology*.
35. Akselrod-Ballin A, et al. (2016) A region based convolutional network for tumor detection and classification in breast mammography in *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. (Springer), pp. 197–205.
36. Zhu W, Lou Q, Vang YS, Xie X (2016) Deep multi-instance networks with sparse label assignment for whole mammogram classification. *arXiv:1612.05968*.
37. Carneiro G, Nascimento J, Bradley AP (2015) Unregistered multiview mammogram analysis with pre-trained deep learning models in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. (Springer), pp. 652–660.
38. Moreira IC, et al. (2012) Inbreast: toward a full-field digital mammographic database. *Academic radiology* 19(2):236–248.
39. Bowyer K, et al. (1996) The digital database for screening mammography in *Third international workshop on digital mammography*. Vol. 58, p. 27.
40. Heath M, et al. (1998) Current status of the digital database for screening mammography in *Digital mammography*. (Springer), pp. 457–460.
41. Wang J, et al. (2016) Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports* 6:27327.
42. Bekker AJ, Greenspan H, Goldberger J (2016) A multi-view deep learning architecture for classification of breast microcalcifications in *IEEE International Symposium on Biomedical Imaging*. pp. 726–730.
43. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. in *International Conference on Artificial Intelligence and Statistics*.
44. Kingma D, Ba J (2015) Adam: A method for stochastic optimization in *International Conference on Learning Representation*.
45. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks in *European Conference on Computer Vision*. (Springer), pp. 818–833.
46. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. *arXiv:1506.06579*.
47. Bergstra JS, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization in *Advances in Neural Information Processing Systems 24*. pp. 2546–2554.
48. Snoek J, et al. (2015) Scalable bayesian optimization using deep neural networks in *International Conference on Machine Learning*.