

Expert identification of visual primitives used by CNNs during mammogram classification

Jimmy Wu^a, Diondra Peck^b, Scott Hsieh^c, Vandana Dialani, MD^d,
Constance D. Lehman, MD^e, Bolei Zhou^a, Vasilis Syrgkanis^f, Lester Mackey^f,
and Genevieve Patterson^f

^aMIT, Cambridge, USA

^bHarvard University, Cambridge, USA

^cDepartment of Radiological Sciences, UCLA, Los Angeles, USA

^dBeth Israel Deaconess Medical Center, Cambridge, USA

^eMassachusetts General Hospital, Cambridge, USA

^fMicrosoft Research New England, Cambridge, USA

ABSTRACT

This work interprets the internal representations of deep neural networks trained for classifying diseased tissue in 2D mammograms. We propose an expert-in-the-loop interpretation method to label the behavior of internal units in convolutional neural networks (CNNs). Expert radiologists identify that the visual patterns detected by the units are correlated with meaningful medical phenomena such as mass tissue and calcificated vessels. We demonstrate that several trained CNN models are able to produce explanatory descriptions to support the final classification decisions. We view this as an important first step toward interpreting the internal representations of medical classification CNNs and explaining their predictions.

Keywords: medical image understanding, deep learning for diagnosis, interpretable machine learning, expert-in-the-loop methods

1. PURPOSE

State-of-the-art convolutional neural networks (CNNs) can now match and even supersede human performance on many visual recognition tasks;^{1,2} however, these significant advances in discriminative ability have been achieved in part by increasing the complexity of the neural network model which compounds computational obscurity.³⁻⁶ CNN models are often criticized as black boxes because of their massive model parameters. Thus, lack of interpretability prevents CNNs from being used widely in clinical settings and for scientific exploration of medical phenomena.^{7,8}

Deep learning based cancer detection in 2D mammograms has recently achieved near human levels of sensitivity and specificity, as evidenced by the large-scale Digital Mammography DREAM Challenge.⁹ Recent computer aided diagnostic systems have also applied advanced machine learning to a combination of imaging data, patient demographics, and

Corresponding authors Jimmy Wu jimmywu@alum.mit.edu and Genevieve Patterson gen@microsoft.com; This paper is being submitted solely to SPIE for publication and presentation.

medical history with impressive results.⁸ However, applications such as breast cancer diagnosis and treatment heavily depend on a sense of trust between patient and practitioner, which can be impeded by black-box machine learning diagnosis systems. Thus, automated image diagnosis provides a compelling opportunity to reevaluate the relationship between clinicians and neural networks. Can we create networks that explain their decision making? Instead of producing only a coarse binary classification, e.g. does a scan reveal disease or not, we seek to produce relevant and informative descriptions of the predictions a CNN makes in a format familiar to radiologists. In this paper, we examine the behavior of the internal representations of CNNs trained for breast cancer diagnosis. We invite several human experts to compare the visual patterns used by these CNNs to the lexicon used by practicing radiologists. We use the Digital Database for Screening Mammography (DDSM)¹⁰ as our training and testing benchmark.

Contributions

This work is the first step toward creating neural network systems that interact seamlessly with clinicians. Our principal contributions, listed below, combine to offer insight and identify commonality between deep neural network pipelines and the workflow of practicing radiologists. Our contributions are as follows:

- We visualize the internal representations of the CNNs trained on cancerous, benign, benign without callback, and normal mammograms;
- We develop an interface to obtain human expert labels for the visual patterns used by the CNNs in cancer prediction;
- We compare the internal representations to the BI-RADS lexicon,¹¹ showing that many interpretable internal CNN units detect meaningful factors used by radiologists for breast cancer diagnosis.

2. METHODS

To gain a richer understanding of which visual primitives CNNs use to predict cancer, we fine-tuned several strongly performing networks on training images from the Digital Database for Screening Mammography (DDSM).¹⁰ We then evaluated the visual primitives detected by individual units that emerged for each fine-tuned model using Network Dissection, a technique to visualize the favorite patterns detected by each unit.¹² Three authors who are practicing radiologists or experts in this area manually reviewed the unit visualization and labeled the phenomena each unit identified. Finally, we compared the named phenomena used by internal units of each CNN to items in the BI-RADS lexicon.¹¹ Here we denote the convolutional filters at each layer as the unit, as opposed to the 'neuron', to disambiguate them from the biological entity. The unit visualization is the set of the top activated images segmented by the unit's feature map.

2.1 Dataset

We conduct our experiments with images from the Digital Database for Screening Mammography, a dataset compiled to facilitate research in computer-aided breast cancer screening. DDSM consists of 2,500 studies, each including two images of each breast, patient age, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of

abnormalities, and information about the imaging modality and resolution. Labels include image-wide designations (e.g., cancerous, benign, benign without callback, and normal) and pixel-wise segmentations of lesions.¹⁰

For the experiments in the following sections, we divided the DDSM dataset scans into 80% train, 10% validation, and 10% test partitions. All images belonging to a unique patient are in the same split, to prevent training and testing on different views of the same breast.

2.2 Network Architectures

We adapted several well-known image classification networks for breast cancer diagnosis as shown in Table 1. We modified the final fully connected layer of each architecture to have two classes corresponding to a positive or negative diagnosis. Network weights were initialized using the corresponding pretrained ImageNet¹³ models and fine-tuned on DDSM. We trained all models in the PyTorch¹⁴ framework using stochastic gradient descent (SGD) with learning rate 0.0001, momentum 0.9, and weight decay 0.0001.

Architecture	AUC
AlexNet ¹⁵	0.8632
VGG-16 ¹⁶	0.8929
Inception-v3 ⁴	0.8805
ResNet-152 ⁵	0.8757

Table 1: The network architectures used and their performance as the AUC on the validation set.

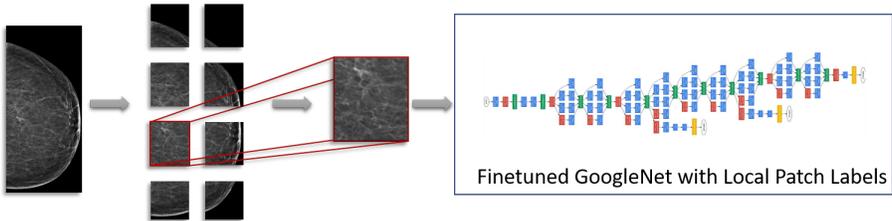


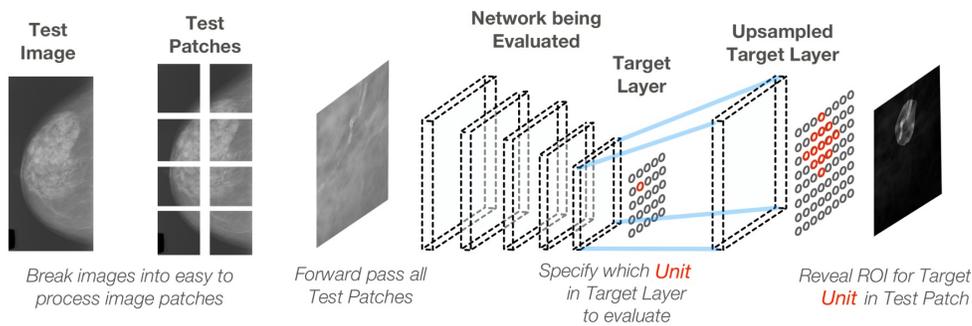
Figure 1: GoogleNet Inception-v3 fine-tuned with local image patches and their labels. Multiple patches (overlapping with a sliding window) are extracted from each image and then passed through a CNN with the local patch label determined by the lesion masks from DDSM. After fine-tuning each network we tested performance on the task of classifying whether the patch contains a malignant lesion. Performance on out-of-sample prediction was as follows: Area Under Curve (AUC) 0.8805, Area Under Precision-Recall Curve: 0.325. We could correctly detect 68% of positive patch examples (radiologist’s positive detection rate ranges between 0.745 and 0.923^{17,18}), while only incorrectly predicting as positive 2% of the negative examples.

Figure 1 illustrates how we prepared each mammogram for training and detection. Because of the memory requirements of processing an high-resolution image with any neural network, we split the mammograms into patches then process image patches instead. We applied a sliding window at 25% the size of a given mammogram with a 50% patch stride. This

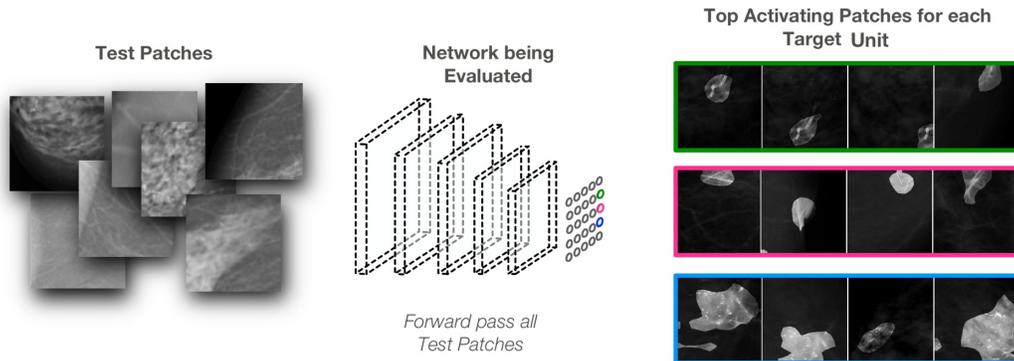
gave us a set of image patches for each mammogram that may or may not have contained a cancerous lesion. The ground truth label for each mammogram patch was calculated as positive for cancer if at least 30% of the lesion was contained in the image patch or at least 30% of the image patch was covered by a lesion; all other patches were assigned a negative label. Lesion locations were determined from the lesion segmentation masks of DDSM.

2.3 Network Dissection

Network Dissection (Net Dissect) is a recent method proposed for assessing how well a visual concept is disentangled within CNNs.¹² Network Dissection defines and quantifies the interpretability as a measure of how well individual units align with sets of human-interpretable concepts.



(a) Illustration of how Network Dissection proceeds for a single instance. Above, one unit is probed to display the Region of Interest (ROI) in the evaluated image responsible for that unit’s activation value. ROIs may not line up as directly as shown in this figure, please see Bau et al.¹² for a complete description of this process.



(b) Illustration of how Network Dissection proceeds for all units of interest in a given convolutional layer. All images from a test set are processed in the manner of Fig. 2a. The top activating test images for each unit are recorded to create the visualization of a unit’s top activated visual phenomena. Each top activating image is segmented by the upsampled and binarized feature map of that unit.

Figure 2: Illustration of Network Dissection for identifying exploited visual phenomena by a CNN of interest.

Figure 2 demonstrates at a high level how the Net Dissect works to interpret the units at the target layer of the network. We employed our validation split of DDSM to create visualizations for each unit in the final convolutional layer of each evaluated network. For our ResNet experiments we also evaluated the second to last convolutional layer due to the depth of that network. Because of the hierarchical structure of CNNs, the final convolutional layer is the layer that will contain the high-level semantic concepts, which are more likely to be aligned with the visual taxonomy used by radiologists than low-level gradient features. The Net Dissect approach to the unit visualization only applies to convolutional network layers due to their maintenance of spatial information.

Figure 2b shows how we created the unit visualizations for our analysis in Sections 2.4 and 3. We passed every image patch from all mammograms in our test set through each of the four networks. For each unit in the target layer, the convolution layer we were investigating, we recorded the unit’s max activation value as the score and the ROI from the image patch that caused the measured activation. To visualize each unit (Figs. 3 and 4), we display the top activating image patches in order sorted by their score for that unit. Each top activating image is further segmented by the upsampled and binarized feature map of that unit to highlight the highly activated image region.

2.4 Human Evaluation of Visual Primitives used by CNNs

To further verify the visual primitive discovered by the networks, we created a web-based survey tool to solicit input from expert readers. The expert readers consisted of two radiologists specialized in breast imaging and one medical physicist. A screenshot from the survey tool is shown in Figure 3. The survey provided a list of 40 to 50 units culled usually from the final layer of the neural network. The neural network often had many more units, too many for exhaustive analysis with the limited user population. Thus the units that were selected were composed partly of the top activating patches that all or mostly contained cancer and partly from a random selection of other patches.

The readers were able to see a preview of each unit, which consisted of several image patches that highlighted the region of interest that triggered the unit to activate most strongly. From this preview, the readers were able to formulate an initial hypothesis of what the unit was associated with. The readers could click through each preview to select units that could be interpreted, and they were brought to a second page dedicated specifically to the unit, which showed additional patches as well as the context of the entire mammogram, as shown in Figure 3. On this page, users could then comment on the unit in a structured report, indicating if there was a distinct phenomenon associated with the unit and its relationship to breast cancer. The web-based survey saved results after each unit and could be accessed over multiple sessions to avoid reader fatigue.

Some of the units shown had no clear connection with breast cancer and would appear to be spurious. Still other units presented what appeared to be entangled events, such as mixtures of mass and calcification, that were associated with malignancy but in a clearly identifiable way. However, many of the units shown appeared to be a clean representation of a single phenomenon known to be associated with breast cancer.

3. RESULTS

We compared the expert-annotated contents of 134 units from four networks to the lexicon of the BI-RADS taxonomy.^{7,11} This qualitative evaluation was designed to estimate the

vgg16/conv5_3/unit_0424

[\[go back to vgg16/conv5_3\]](#)

Instructions

1. Please inspect the highlighted areas within the images on this page, then answer the survey questions below.
2. The survey questions pertain only to the images shown on this page.
3. The images show small patches of mammograms. You can hover over an image to view the patch within the full-sized mammogram.
4. You can view whether a image patch is cancerous or non-cancerous by hovering over the image. The label will show on the right above the full-sized mammogram. Note that the label corresponds only to the image patch, not the entire full-sized mammogram. A full-sized mammogram can have both cancerous and non-cancerous patches.
5. You must click on the submit button under the survey questions to save your responses. After your responses are saved, you will be taken back to the previous page.

Survey Questions

Note: The following questions pertain only to the images shown on this page.

Do these images show recognizable phenomena?

yes

If the images show recognizable phenomena, are they associated with breast cancer?

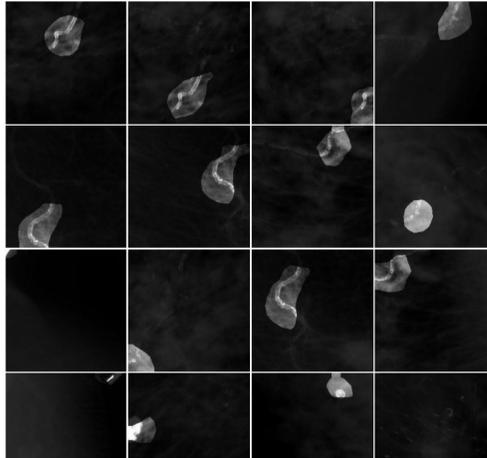
yes

If the images show recognizable phenomena, is there a single unique phenomenon, or are there multiple?

single

Please describe each of the phenomena you see. For each phenomenon please indicate its association with breast cancer.

calcified vessels



patch label: negative

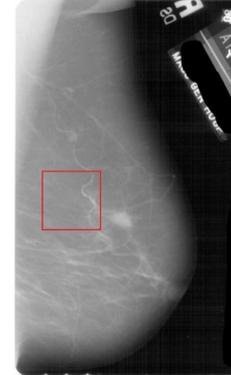


Figure 3: Web-based Survey Tool: This user interface was used to ask the expert readers about the units of interest. The survey asked questions such as, “Do these images show recognizable phenomena?” and, “Please describe each of the phenomena you see. For each phenomenon please indicate its association with breast cancer. In the screenshot above, one expert has labeled the unit’s phenomena as ‘Calcified Vessels’.

overlap between the standard system radiologists use to diagnose breast cancer and the visual primitives used by the trained CNNs.

Direct classification of BI-RADS entities has long been a topic of interest in machine learning for mammography.¹⁹ Our experiments differ from direct classification because our training set was constructed with simple positive/ negative labels instead of detail BI-RADS categories. In this work we chose a well-understood medical event, the presence of cancer in mammograms, to evaluate if unit visualization is a promising avenue for discovering important visual phenomena in less well-understood applications. Our results, shown in Fig. 4, show that networks trained to recognize cancer end up using a large percentage of the BI-RADS categories even though the training labels were simply cancer/ no cancer.

Units in all networks identify advanced cancers, large benign masses, and several kinds of obvious calcifications. Encouragingly, many units also identify important associated features such as spiculation, breast density, architectural distortions, and the state of tissue near the nipple. Several units in Fig. 4 show that the CNNs use breast density and parenchymal patterns to make predictions. This network behavior could be used to find a new computational perspective on the relationship between breast density, tissue characteristics, and cancer risk, which has been a popular research topic for the last 25 years.^{20–22}

4. CONCLUSION

In this exploratory study, we were able to show that many internal units of a deep network identify visual concepts used by radiologists. Indeed, Fig. 4 shows significant overlap with the BI-RADS lexicon. However, some units had no identified connection with breast cancer, and yet other units identified entangled events. In future work, we will investigate both the units with nameable phenomena and those which appear to be spurious to identify if there

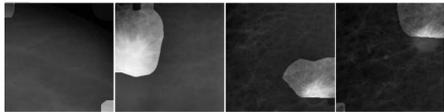
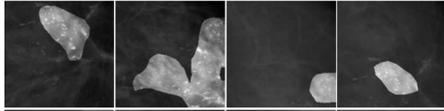
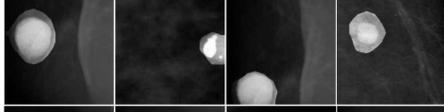
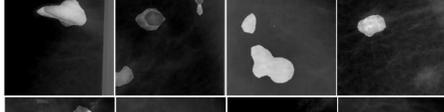
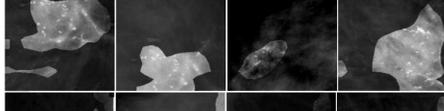
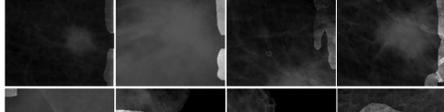
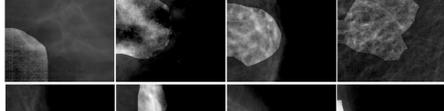
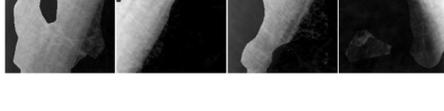
BI-RADS Lexicon Category	Neuron Annotation	Network, Layer, Neuron	
Mass - Margin	<i>masses with spiculated edge</i>	Inception v3 mixed_7a unit 0371	
Calcification	<i>calcifications, innumerable</i>	VGG 16 conv5_3 unit 0063	
Breast Composition	<i>high density area, large calcifications</i>	AlexNet conv5 unit 0014	
Mass	<i>advanced cancers</i>	VGG-16 conv5_3 unit 0283	
Associated Features	<i>architectural distortion</i>	ResNet 152 layer 4 unit 0183	
Calcification, Associated Features	<i>calcifications, nearby tissue distortions</i>	VGG 16 conv5_3 unit0048	
Calcification, Mass	<i>calcification adjacent to masses</i>	ResNet 152 layer 4 unit 0253	
Breast Composition	<i>fatty breast texture</i>	ResNet 152 layer 4 unit 0005	
Associated Features	<i>structure close to nipple</i>	AlexNet conv5 unit 0079	
-	<i>pectoralis muscle</i>	VGG 16 conv5_3 unit 0167	

Figure 4: The table above shows some of the labeled units and their interpretations. The far-left column lists the general BI-RADS category associated with the units visualized in the far-right column. The second-left column displays the expert annotation of the visual event identified by each unit, summarized for length. The third-left column lists the network, convolutional layer, and unit's unit ID number.

are medically relevant visual events predictive of cancerous lesions that are not used by clinicians. We will also explore how to use the unit labeling technique presented in this paper to generate natural language explanations of the predictions made by diagnosing neural networks.

REFERENCES

- [1] He, K., Zhang, X., Ren, S., and Sun, J., “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in [*ICCV*], 1026–1034 (2015).
- [2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature* **542**(7639), 115–118 (2017).
- [3] Bolei, Z., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., “Object detectors emerge in deep scene cnns,” *ICLR* (2015).
- [4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., “Going deeper with convolutions,” in [*Computer Vision and Pattern Recognition (CVPR)*], (2015).
- [5] He, K., Zhang, X., Ren, S., and Sun, J., “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385* (2015).
- [6] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K., “Aggregated residual transformations for deep neural networks,” in [*Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*], 5987–5995, IEEE (2017).
- [7] Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D’orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., et al., “Influence of computer-aided detection on performance of screening mammography,” *New England Journal of Medicine* **356**(14), 1399–1409 (2007).
- [8] Song, L., Hsu, W., Xu, J., and Van Der Schaar, M., “Using contextual learning to improve diagnostic accuracy: Application in breast cancer screening,” *IEEE journal of biomedical and health informatics* **20**(3), 902–914 (2016).
- [9] Bionetworks, S., “Digital mammography dream challenge,” (2016).
- [10] Heath, M., Bowyer, K., Kopans, D., Moore, R., and Kegelmeyer, W. P., “The digital database for screening mammography,” in [*Proceedings of the 5th international workshop on digital mammography*], 212–218, Medical Physics Publishing (2000).
- [11] Reporting, B. I., “Data system (bi-rads),” *Reston VA: American College of Radiology* (1998).
- [12] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A., “Network dissection: Quantifying interpretability of deep visual representations,” *CVPR* (2017).
- [13] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., “Imagenet: A large-scale hierarchical image database,” in [*Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*], 248–255, IEEE (2009).
- [14] Paszke, A., Gross, S., Chintala, S., and Chanan, G., “Pytorch,” (2017).
- [15] Krizhevsky, A., Sutskever, I., and Hinton, G. E., “Imagenet classification with deep convolutional neural networks,” in [*Advances in neural information processing systems*], 1097–1105 (2012).
- [16] Simonyan, K. and Zisserman, A., “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556* (2014).

- [17] Lévy, D. and Jain, A., “Breast mass classification from mammograms using deep convolutional neural networks,” *arXiv preprint arXiv:1612.00542* (2016).
- [18] Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., Yankaskas, B. C., Kerlikowske, K., Onega, T., Rosenberg, R. D., et al., “Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy,” *Radiology* **253**(3), 641–651 (2009).
- [19] Orel, S. G., Kay, N., Reynolds, C., and Sullivan, D. C., “Bi-rads categorization as a predictor of malignancy,” *Radiology* **211**(3), 845–850 (1999).
- [20] Oza, A. M. and Boyd, N. F., “Mammographic parenchymal patterns: a marker of breast cancer risk,” *Epidemiologic reviews* **15**(1), 196–208 (1993).
- [21] Petroudi, S., Kadir, T., and Brady, M., “Automatic classification of mammographic parenchymal patterns: A statistical approach,” in [*Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*], **1**, 798–801, IEEE (2003).
- [22] McCormack, V. A. and dos Santos Silva, I., “Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis,” *Cancer Epidemiology and Prevention Biomarkers* **15**(6), 1159–1169 (2006).