Expert identification of visual primitives used by CNNs during mammogram classification

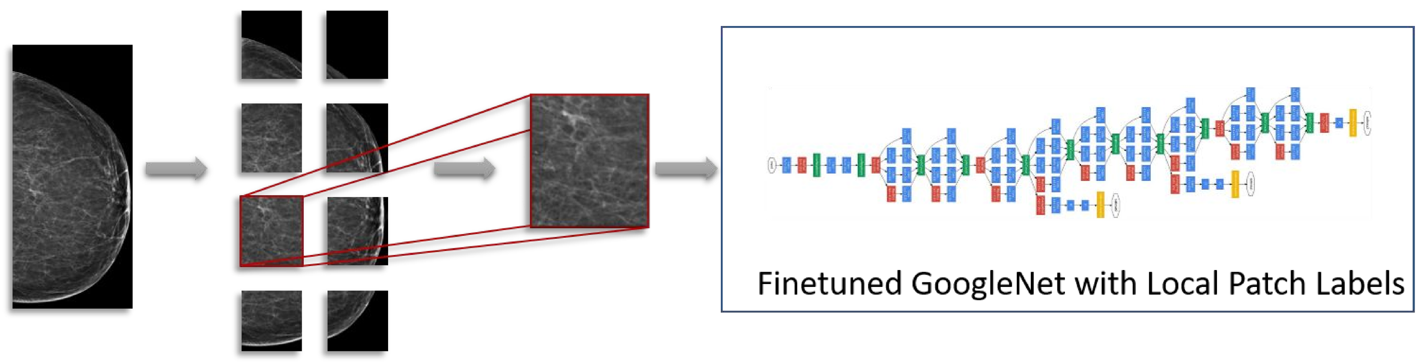
Jimmy Wu[a], Diondra Peck[b], Scott Hsieh[c], Vandana Dialani, MD[d], Constance D. Lehman, MD[e], Bolei Zhou[a], Vasilis Syrgkanis[f], Lester Mackey[f], and Genevieve Patterson[f]

[a]MIT, Cambridge, USA [b]Harvard University, Cambridge, USA [c]Department of Radiological Sciences, UCLA, Los Angeles, USA
[d]Beth Israel Deaconess Medical Center, Cambridge, USA [e]Massachusetts General Hospital, Cambridge, USA [f]Microsoft Research New England, Cambridge, USA

INTRODUCTION

This work interprets the internal representations of deep neural networks trained for classifying the diseased tissue in 2D mammograms. We propose an expert-in-the-loop interpretation method to label the behavior of the internal units of 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 calcified 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.

MODELS



GoogleNet Inception-v3 architecture fine-tuned with local image patches and their labels. Multiple overlapping patches are extracted from each image with a sliding window 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.

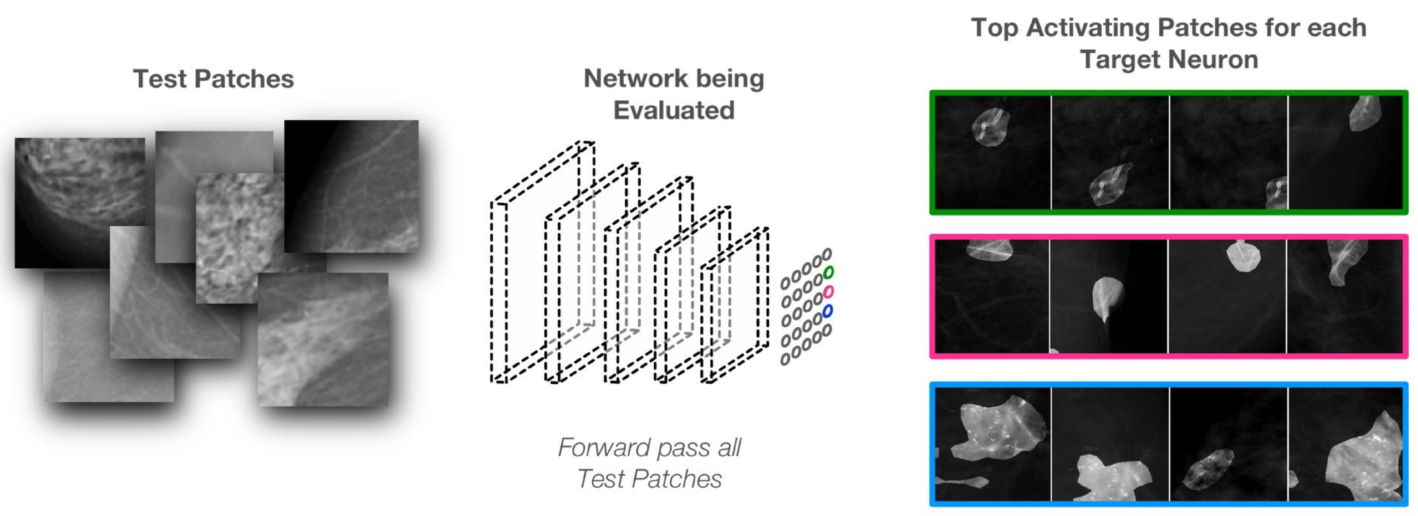
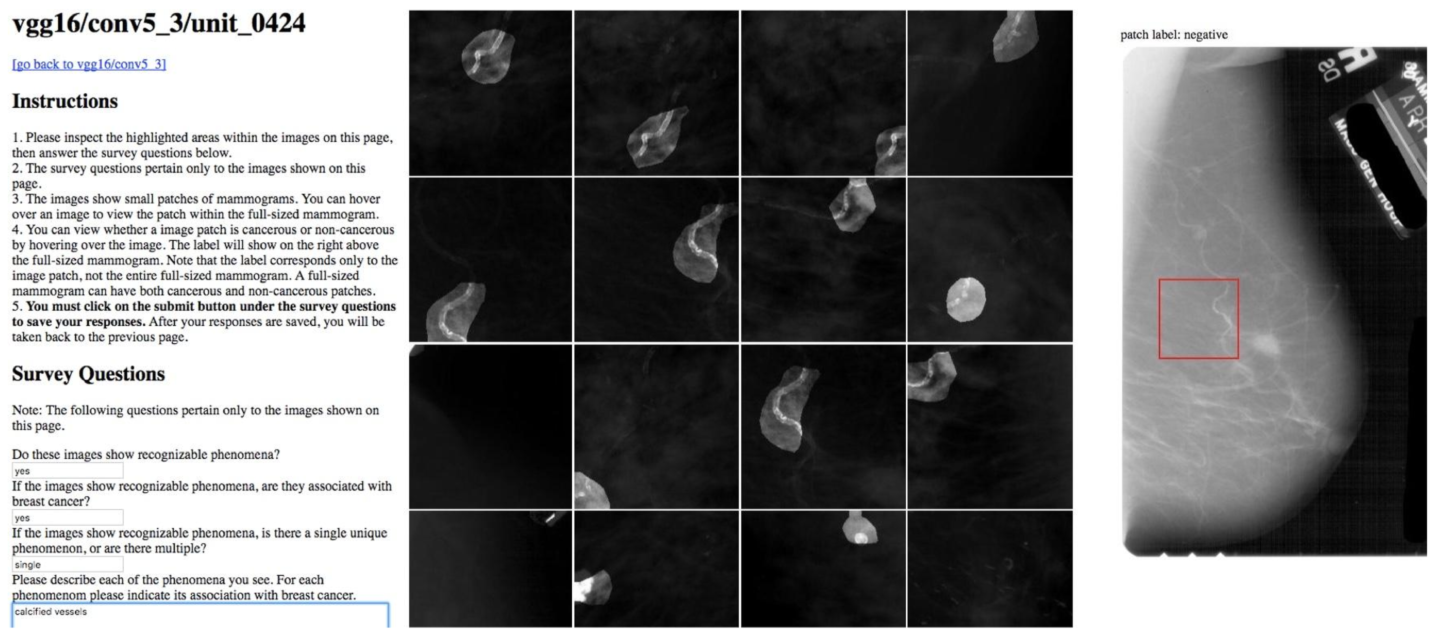


Illustration of how Network Dissection proceeds for all units of interest in a given convolutional layer. For each unit in the target layer, the convolution layer we were investigating, we recorded the unit's max activation value as the score and the Region of Interest (ROI) from the image patch that caused the measured activation. To visualize each unit (3 rows on the right side), 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.

METHODS



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." In the screenshot above, one expert has labeled the unit's phenomena as 'Calcified Vessels'.

Fine-tuned Model Performance

Architecture	Training Epochs	AUC
AlexNet	45	0.86
VGG 16	12	0.89
Inception v3	7	0.88
ResNet 152	5	0.87

CONCLUSIONS

1. **Many internal units of a deep network identify visual concepts used by radiologists (significant overlap with the BI-RADS lexicon)**
2. **Future Work: Investigate how to further disentangle and identify computationally discriminative medical visual phenomena using deep nets**
3. **Future Work: Use the unit labeling technique presented in this paper to generate natural language explanations of the predictions made by diagnosing neural networks.**

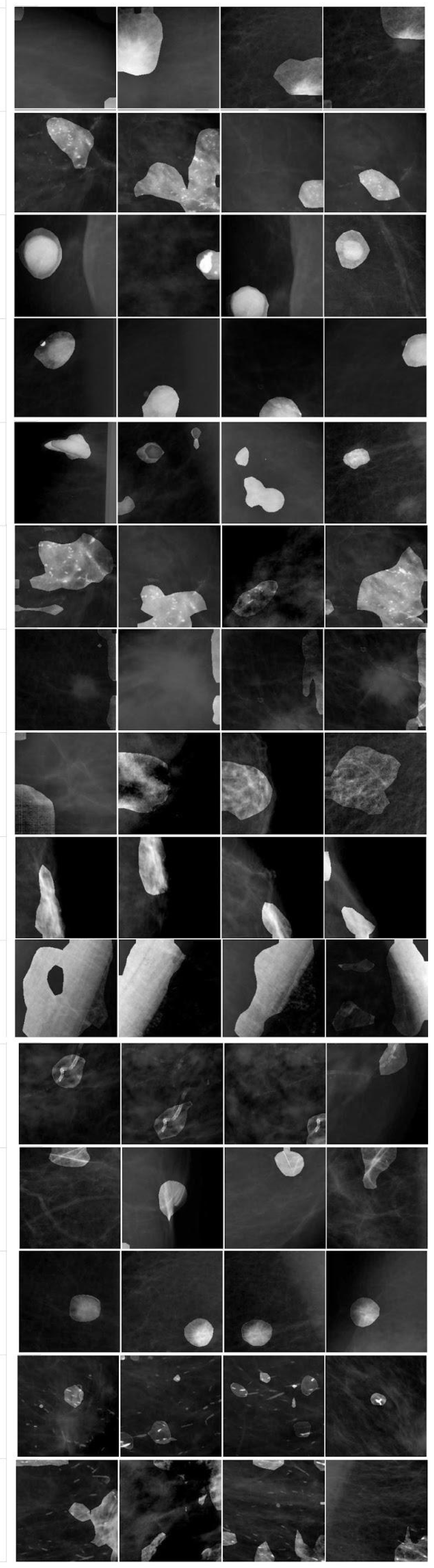
REFERENCES

1. Bolei, Z., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Object detectors emerge in deep scene CNNs," ICLR (2015).
2. Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A., "Network dissection:Quantifying interpretability of deep visual representations," CVPR (2017).
3. 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).

RESULTS

Unit Visualization: The table below 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 ID number.

BI-RADS Lexicon Category	*Neuron Annotation*	Network, Layer, Neuron
Mass - Margin	*masses with spiculated edge*	Inception v3 mixed_7a unit 0371
Calcification	*calcifications, innumerable*	VGG 16 conv5_3 unit 0063
Breast Composition	*high density area, large calcifications*	AlexNet conv5 unit 0014
Mass	*advanced cancers*	VGG-16 conv5_3 unit 0283
Associated Features	*architectural distortion*	*ResNet 152 layer 4 unit 0183*
Calcification, Associated Features	*calcifications, nearby tissue distortions*	VGG 16 conv5_3 unit0048
Calcification, Mass	*calcification adjacent to masses*	ResNet 152 layer 4 unit 0253
Breast Composition	*fatty breast texture*	ResNet 152 layer 4 unit 0005
Associated Features	*structure close to nipple*	AlexNet conv5 unit 0079
-	*pectoralis muscle*	*VGG 16 conv5_3 unit 0167*
Calcification	*calcified vessels*	VGG 16 conv5_3 unit0424
Calcification	*calcified vessels*	VGG 16 conv5_3 unit0195
Mass	*masses*	ResNet 152 layer 4 unit 0582
Calcification	*clustered microcalcifications*	ResNet 152 layer 3 unit 0235
Calcification	*linear calcification*	AlexNet conv5 unit 0048



Check out our paper at this link:

Corresponding authors **Jimmy Wu** *jimmywu@alum.mit.edu* and **Geneviève Patterson** *gen@microsoft.com*