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.

RESULTS

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’.

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