Visualizing the differences in visual primitives used by CNNs and radiologists

Diondra Peck[a], Genevieve Patterson[b], Lester Mackey[b], and Vasilis Syrgkanis[b]

[a]Harvard University, Cambridge, USA
[b]Microsoft Research New England, Cambridge, USA

ABSTRACT

We propose a method to compare the visual primitives used by radiologists and convolutional neural networks (CNNs) for the identification of diseased tissue in mammograms. Using the latest advances in neural network interpretability, we investigate the visual primitives used by two different CNNs, a baseline model and our own high-performance model, to perform automatic medical diagnosis. By visualizing the phenomena identified by diagnosis CNNs and interviewing specialists about the CNNs' behavior, we show the difference between visual factors used by radiologists and automatic systems. Our proposed method yields insights into improving accuracy and interpretability in automatic diagnosis systems.

1. PURPOSE

Many state-of-the-art CNNs can now match and even supersede human performance on generic recognition tasks;[1] however, these significant advances in discriminative ability have been achieved in part by deepening nets which compounds computational obscurity.[2] Due to the depth and complexity of CNN, they are often criticized as black boxes. Lack of intepretability prevents CNNs from being used for scientific exploration of medical phenomena. Medical applications are also heavily dependent on a sense of trust between patient and practitioner, which can impeded by monolithic automatic diagnosis systems. Thus, automated image diagnosis provides a compelling opportunity to investigate both the discriminatory ability of state-of-the-art CNNs as well as the difference between factors used by humans and automatic systems.

Using canonical network activation visualization techniques,[3] we investigate the potential of CNNs to provide insight into image-based diagnostics of tissue lesions. Given the level of trust needed in medical applications, we also examine how visualization techniques may help us confirm that neural networks can "diagnose" in a human-interpretable fashion. Understanding the classification behavior of CNNs may reveal similarities and differences between the diagnostic processes of human radiologists and automated systems, potentially yielding insights that help both do their job more accurately and responsibly.

2. METHODS

To gain a richer understanding of which visual primitives CNNs use to automate diagnosis, we passed images from the Digital Database for Screening Mammography (DDSM) through two different models and recorded their discrimination performance.[4] We then evaluated the visual primitives that emerged for each model using Network Dissection, a novel visualization technique.[3]

Further author information: Send correspondence to Diondra Peck `diondrapeck@college.harvard.edu`; This paper is being submitted solely to SPIE for publication and presentation.

2.1 Dataset

In order to ascertain alignment of image labels with both image-wide observations and local events, we used images from the Digital Database for Screening Mammography, a dataset constructed to facilitate research in computer-aided breast cancer screening.[4] 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.[4] Labels include image-wide designations (e.g., cancerous, benign, and normal) and pixel-wise segmentations of lesions.[4]

2.2 System Architectures

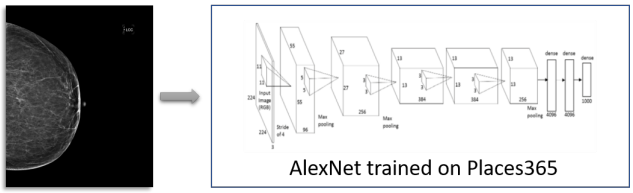

Figure 1: AlexNet trained on a completely different dataset of Places365 for scene recognition, applied to the whole mammogram image.

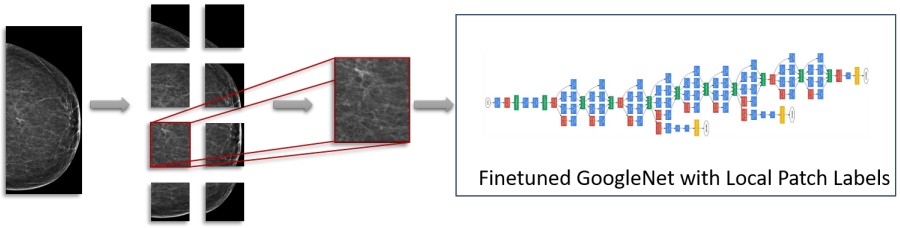

Figure 2: GoogleNet fine-tuned with local labels on patches of an image. Image is dissected in multiple patches (overlapping with a sliding window) and then passed through a copy of a CNN with the local label determined by the lesion masks from DDSM. After finetuning GoogleNet we tested performance on the task of classifying whether the patch contains a malignant lesion. Performance on out-of-sample prediction was as follows: Area Under Curve (AUC) 0.8627, 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[5,6]), by incorrectly predicting as positive 2% of the negative examples.

2.3 Network Dissection

Network Dissection is a novel method for assessing how well a visual concept is disentangled within CNNs.[3] Disentanglement is a measure of how clear a visual concept is in terms of human perception. Thus, Network Dissection defines and quantifies intepretability as a measure of how well network visualizations align with sets of human-interpretable concepts. The process of applying Network Dissection to a CNN is inspired by the methodology used in neuroscience to investigate analogous problems in the human brain.[7]

To evaluate our networks using Network Dissection, we first collected a diverse set of human-interpretable concepts used by radiologists to diagnose breast abnormalities. We

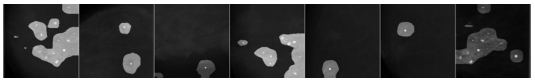

Figure 3: Several calcification detectors emerge in layer conv5 of the AlexNet.[3]

then passed test images from DDSM forward through our CNNs and recorded which receptive fields caused strong activations in which units in order to quantify their alignment. In this manner, we can visualize the visual primitives associated with each unit in the network by looking at which ROIs from the test set activated each unit.

3. RESULTS

3.1 Visual Primitives

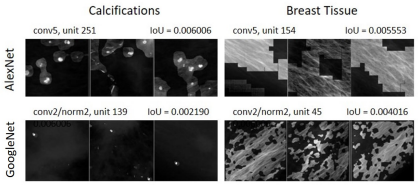

Figure 4: Calcifications and breast tissue were the top ranked visual primitives used by both the AlexNet and GoogleNet to perform a binary classification task, deciding if mammograms were cancerous or not. For each set of visualizations, the unit layer and number are indicated on the top left, and the intersection over union (IoU) is indicated on the top right.

We investigated visual primitives used by two different CNNs, a baseline model and our own high-performance model, to perform automatic cancer diagnosis. For a baseline model, we used an AlexNet that was trained on the Places365 dataset for scene recognition.[8] Despite the AlexNet having never been trained on medical images, detectors for many medical phenomena, including breast masses, calcifications, normal tissue, and blood vessels, emerged.

Using the same dataset, we also observed detectors emerging in our GoogleNet which was fine-tuned for automatic diagnosis. However, we observed fewer visual primitives used by the GoogleNet compared to the AlexNet. Further, of the visual primitives that we observed, the ROIs were significantly smaller than those of the AlexNet.

Interestingly, despite smaller ROIs, the GoogleNet performs very well on automated diagnostic tasks. This may indicate that subtle features in medical images can inform diagnoses with a level of accuracy similar to more prominent features. We plan to investigate

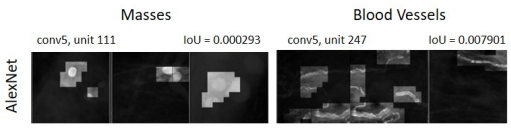

Figure 5: Masses and blood vessels emerge as visual primitives in the AlexNet. Interestingly, these did not emerge in the fine-tuned GoogleNet.

how these region sizes compare to those used by human radiologists in order to determine if humans and CNNs develop divergent diagnostic processes.

3.2 Human Evaluation of Network Activations

The final version of this work will compare the visual primitives identified by CNNs with the diagnostic visual features used by radiologists. We will draw upon both the fine-grained radiologist annotations present in the DDSM dataset and the results of a survey conducted with practicing radiologists. Representative questions from our survey are shown below.

1. Consider the highlighted regions in each of the following mammogram patches:
2. How would you describe the phenomena highlighted in these images?
3. Please describe any relationship between the highlighted phenomena and breast cancer.

4. CONCLUSION

By comparing the neuron activations of a baseline CNN and a CNN fine-tuned for the specific task of cancerous lesion classification, we observe different sets of visual primitives being used for discrimination. For the baseline, Places365-trained AlexNet, the discovered visual primitives were useful, nameable medical phenomena. For the GoogleNet fine-tuned on image patches from the DDSM dataset this was also true, but the visual elements occupied a smaller visual extent. This confirms the usefulness of hyper-local visual events, like microcalcificationns and small-scale tissue texture, in diagnosing cancerous lesions. In the full-length version of this work, we will also compare the CNN-identified visual primitives to important discriminative visual events identified by radiologists.

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] Bolei, Z., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Object detectors emerge in deep scene cnns," *ICLR* (2015).

[3] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A., "Network dissection: Quantifying interpretability of deep visual representations," *CVPR* (2017).

[4] 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).

[5] Lévy, D. and Jain, A., "Breast mass classification from mammograms using deep convolutional neural networks," *arXiv preprint arXiv:1612.00542* (2016).

[6] 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).

[7] Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I., "Invariant visual representation by single neurons in the human brain.," *Nature* **435**(5), 1102–7 (2005).

[8] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A., "Places: A 10 million image database for scene recognition," *IEEE PAMI* (2017).