arXiv:1805.12323v1 [cs.CV] 31 May 2018

DeepMiner: Discovering Interpretable Representations for Mammogram Classification and Explanation

Jimmy Wu[1], Bolei Zhou[1], Diondra Peck[2], Scott Hsieh[3], Vandana Dialani, MD[4]
Lester Mackey[5], and Genevieve Patterson[5]

[1] MIT CSAIL, Cambridge, USA
[2] Harvard University, Cambridge, USA
[3] Department of Radiological Sciences, UCLA, Los Angeles, USA
[4] Beth Israel Deaconess Medical Center, Cambridge, USA
[5] Microsoft Research New England, Cambridge, USA

Abstract. We propose DeepMiner, a framework to discover interpretable representations in deep neural networks and to build explanations for medical predictions. By probing convolutional neural networks (CNNs) trained to classify cancer in mammograms, we show that many individual units in the final convolutional layer of a CNN respond strongly to diseased tissue concepts specified by the BI-RADS lexicon. After expert annotation of the interpretable units, our proposed method is able to generate explanations for CNN mammogram classification that are correlated with ground truth radiology reports on the DDSM dataset. We show that DeepMiner not only enables better understanding of the nuances of CNN classification decisions, but also possibly discovers new visual knowledge relevant to medical diagnosis.

Keywords: deep learning, interpretability, human-in-the-loop machine learning

1 Introduction

Deep convolutional neural networks (CNNs) have made great progress in visual recognition challenges such as object classification [1] and scene recognition [2], even reaching human-level image understanding in some cases [3]. Recently, CNNs have been widely used in medical image understanding and diagnosis [4,5,6]. However, with millions of model parameters, CNNs are often treated as 'black-box' classifiers, depriving researchers of the opportunity to investigate what is learned inside the network and explain the predictions being made. Especially in the domain of automated medical diagnosis, it is crucial to have interpretable and explainable machine learning models.

Several visualization methods have previously been proposed for investigating the internal representations of CNNs. For example, internal units of a CNN can be represented by reverse-mapping features to the input image regions that activate them most [7] or by using backpropagation to identify the most salient regions of an image [8,9]. Our work is inspired by recent work that visualizes and annotates interpretable units of a CNN using Network Dissection [10].

Meanwhile, recent work in automated diagnosis methods has shown promising progress towards interpreting models and explaining model predictions. Wu et al. [11] show that CNN internal units learn to detect medical concepts which match the vocabulary used by practicing radiologists. Rajpurkar et al. [4] and Wang et al. [6] use the class activation map defined in [12] to explain informative regions relevant to final predictions. Zhang et al. propose a hybrid CNN and LSTM (long short-term memory) network capable of diagnosing bladder pathology images and generating radiological reports if trained on sufficiently large image and diagnostic report datasets [13]. However, their method requires training on full medical reports. In contrast, our approach can be used to discover informative visual phenomena spontaneously with only coarse training labels. Jing et al. [14] successful created a visual and semantic network that directly generates long-form radiological reports for chest X-rays after training on a dataset of X-ray images and associated ground truth reports. However, even with these successes, many challenges remain. Wu et al. only show that interpretable internal units are correlated with medical events without exploring ways to explain the final prediction. The heatmaps generated in [4,6] qualitatively tell *where* is important in an image but fails to identify *specific concepts*. Jing et al. train their models on large-scale medical report datasets; however, large text corpora associated with medical images are not easily available in other scenarios. Additionally, Zhang et al. acknowledge that their current classification model produces false alarms that it cannot yet self-correct from.

In this paper, we propose a general framework called *DeepMiner* for discovering medical phenomena in coarsely labeled data and generating explanations for final predictions, with the help of a few human expert annotations. We apply our framework to mammogram classification, an already well-characterized domain, in order to provide confidence in the capabilities of deep neural networks for discovery, classification, and explanation.

To the best of our knowledge, our work is the first automated diagnosis CNN that can both discover discriminative visual phenomena for breast cancer classification and generate interpretable, radiologist-collaborative explanations for its decision-making. Our main contribution is two-fold: (1) we propose a human-in-the-loop framework to enable medical practitioners to explore the behavior of CNN models and annotate the visual phenomena discovered by the models, and (2) we leverage the internal representations of CNN models to explain their decision making, without the use of external large-scale report corporaora.

2 The DeepMiner Framework

The DeepMiner framework consists of three phases, as illustrated in Fig. 1. In the first phase, we train standard neural networks for classification on patches cropped from full mammograms. Then, in the second phase, we invite human experts to annotate the top class-specific internal units of the trained networks. Finally, in the third phase, we use the trained network to generate explainable predictions by ranking the contributions of individual units to each prediction.

In this work, we select mammogram classification as the testing task for our DeepMiner framework. The classification task for the network is to correctly classify mammogram patches as normal (containing no findings of interest), benign (containing only non-cancerous findings), or malignant (containing cancerous findings). Our framework can be further generalized to other medical image classification tasks. Note that we refer to convolutional filters in our CNNs as 'units', as opposed to 'neurons', to avoid conflation with the biological entities.

2.1 Dataset and Training

We choose ResNet-152 pretrained on ImageNet [15] as our reference network due to its outstanding performance on object classification. We fine-tune the pretrained model to classify mammogram patches containing normal, benign, or malignant findings from the Digital Database for Screening Mammography (DDSM) [16]. DDSM is a dataset compiled to facilitate research in computer-aided breast cancer screening. It consists

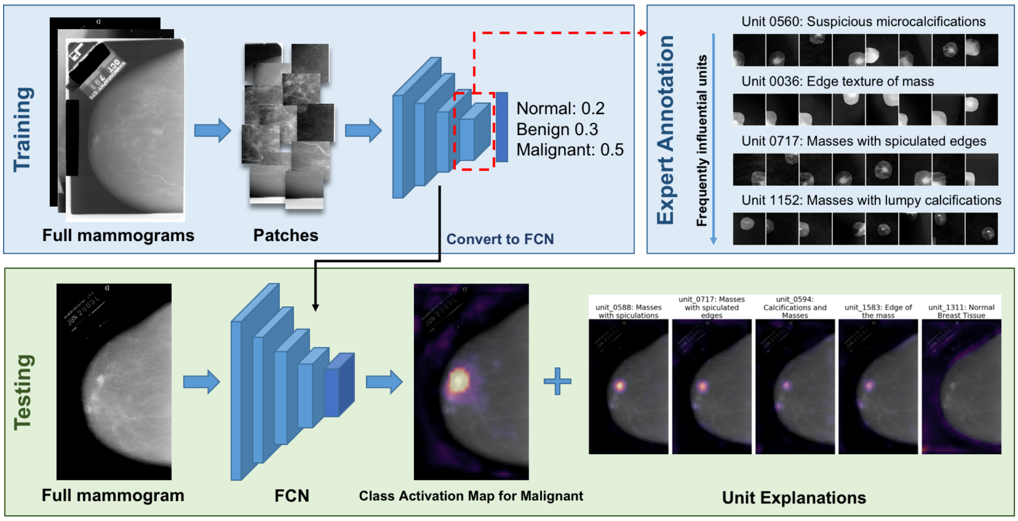

Training
Full mammograms
Patches
Normal: 0.2
Benign 0.3
Malignant: 0.5
Convert to FCN
Expert Annotation
Frequently influential units
Unit 0560: Suspicious microcalcifications
Unit 0036: Edge texture of mass
Unit 0717: Masses with spiculated edges
Unit 1152: Masses with lumpy calcifications
Testing
Full mammogram
FCN
Class Activation Map for Malignant
Unit Explanations
unit_0588: Masses with spiculations
unit_0717: Masses with spiculated edges
unit_0594: Calcifications and Masses
unit_1583: Edge of the mass
unit_1311: Normal Breast Tissue

Fig. 1: Illustration of the DeepMiner framework for mammogram classification and explanation.

of 2,500 studies, each including two images of each breast, a BI-RADS rating of 0-5 for cancer risk, a radiologist's subjective subtlety rating for each finding, and a BI-RADS keyword description of abnormalities. Labels include image-wide designations (e.g., malignant, benign, and normal) and pixel-wise segmentations of lesions [16].

For the following experiments, we partition the DDSM dataset scans into 80% train, 10% test, and 10% additional hold-out scans. CNN performance metrics are reported on the test partition. DeepMiner explanations are evaluated on the additional hold-out partition. All images belonging to a unique patient are in the same partition, to prevent training and testing on different views of the same breast.

To increase the number of training examples for fine-tuning, we break up mammograms into smaller image patches in a sliding window fashion. The dimensions of each image patch are 25% of the width of the original mammogram, and overlapping patches are extracted using a stride of 50% of the patch width. Any patch containing less than 50% breast tissue was excluded from our training patches. We create three class labels (normal, benign, malignant) for each image patch based on (1) whether at least 30% of the patch contains benign or malignant tissue, and (2) whether at least 30% of a benign or malignant finding is located in that patch.

We fine-tune our reference network using stochastic gradient descent (SGD) with learning rate 0.0001, momentum 0.9, weight decay 0.0001, and batch size 32. Performance metrics for tissue classification are shown in Sec. 3.1.

2.2 Human Annotation of Visual Primitives Used by CNNs

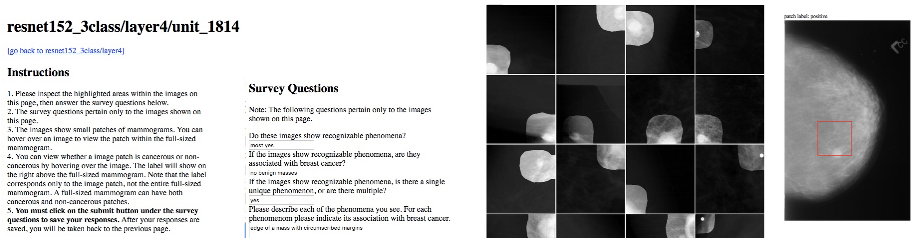
We use our test split of DDSM to create visualizations for units in the final convolutional layer of our fine-tuned ResNet-152. We choose the final layer since it is most likely to contain high-level semantic concepts due to the hierarchical structure of CNNs.

It would be infeasible to annotate all 2048 units in the last convolutional layer. Instead, we select a subset of the units deemed most frequently 'influential' to classification decisions. Given a classification decision for an image, we define the influence of a unit towards that decision as the unit's maximum activation score on that image multiplied by the weight of that unit for a given output class in the final fully connected layer.

For each of the three classes, we selected the twenty most frequently influential units (60 total) and asked human experts to annotate them. For the normal tissue class, if the twenty units we selected were annotated, those annotations would account for 59.27% of the per-image top eight units over all of the test set images. The corresponding amount for the benign class is 69.77%, and for the malignant class is 75.82%.

We create visualizations for each individual unit by passing every image patch from all mammograms in our test set through our classification network. For each unit in the final convolutional layer, we record the unit's maximum activation value as well as the receptive field from the image patch that caused the measured activation. To visualize each unit (see Figs. 2 and 3), we display the top activating image patches sorted by their activation score and further segmented by the binarized and upsampled response map of that unit.

A radiologist and a medical physicist specializing in mammography annotated the 60 most frequently influential units we selected. We compare the named phenomena detected by these units to the BI-RADS lexicon [17]. The experts used the annotation interface shown in Fig. 2. Our survey displays a table of dozens of the top scoring image patches for the unit being visualized. When the expert mouses over a given image patch, the mammogram that the patch came from is displayed on the right with the patch outlined in red. This gives the expert some additional context. From this unit preview, experts are able to formulate an initial hypothesis of what phenomena a unit detects.



resnet152_3class/layer4/unit_1814

[go back to resnet152_3class/layer4]

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?
most yes
If the images show recognizable phenomena, are they associated with breast cancer?
no benign masses
If the images show recognizable phenomena, is there a single unique phenomenon, or are there multiple?
yes
Please describe each of the phenomena you see. For each phenomenom please indicate its association with breast cancer.
edge of a mass with circumscribed margins

Fig. 2: *The interface of the DeepMiner survey:* Experts used this survey form to label influential units. The survey asks 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, the radiologist who was our expert-in-the-loop has labeled the unit's phenomena as 'edge of mass with circumscribed margins'.

Of the 60 units selected, 45 were labeled by at least one expert as detecting a nameable medical phenomena. Fig. 3 shows five of the annotated units. In this figure, each row illustrates a different unit. The table lists the unit ID number, the BI-RADS category for the concept the unit is detecting, the expert-provided unit annotation, and a visual representation of the unit. We visualize each unit by displaying the top four activating image patches from the test set. The unit ID number is listed to uniquely identify each labeled unit in the network used in this paper, which will be made publicly available upon publication.

Fig. 3 demonstrates that the DeepMiner framework discovers significant medical phenomena, relevant to mammogram-based diagnosis. Because breast cancer is a well-characterized disease, we are able to show the extent to which discovered unit detectors overlap with phenomena deemed to be important by the radiological community. For diseases less well understood than breast cancer, DeepMiner could be a useful method for revealing unknown discriminative visual features helpful in diagnosis and treatment planning.

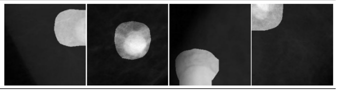
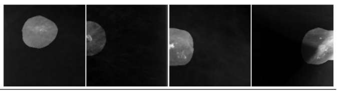
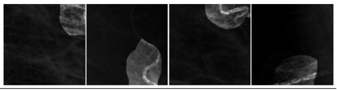
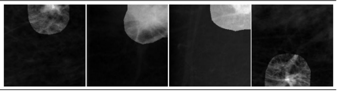
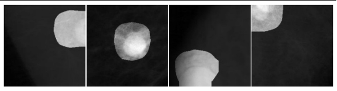
Unit ID	BI-RADS Lexicon	Expert Annotation	Top Activated Images
1814	Mass, Associated Features	Edge of a mass with circumscribed margins	
1152	Calcification	Malignant pleomorphic calcifications	
860	Calcification	Benign vascular calcifications	
1299	Associated Features	Spiculation	
1468	Mass	Masses with smooth edges	

Fig. 3: *Interpretable units discovered by DeepMiner:* The table above illustrates five annotated units from the last convolutional layer of our reference network. Even though the CNN presented in this paper was only trained to classify normal, benign, and malignant tissue, these internal units detect a variety of recognizable visual events. Both benign and malignant calcifications are identified, as well as features related to the margins of masses. These details are significant factors in planning interventions for breast cancer. Please refer to the supplement for a full table of annotated units.

2.3 Explaining Network Decisions

We further use the annotated units to build an explanation for single image prediction. We first convert our trained network into a fully convolutional network (FCN) using the method described in [18] and remove the global average pooling layer. The resulting network is able to take in full mammogram images and output probability maps aligned with the input images.

As illustrated in Fig. 1, given an input mammogram, we output a classification as well as the Class Activation Map (CAM) proposed in [12]. We additionally extract the activation maps for the units most influential towards the classification decision. By looking up the corresponding expert annotations for those units, we are able to see which nameable visual phenomena contributed to the network's final classification. For examples of the DeepMiner explanations, please refer to Sec. 3.2.

3 Results

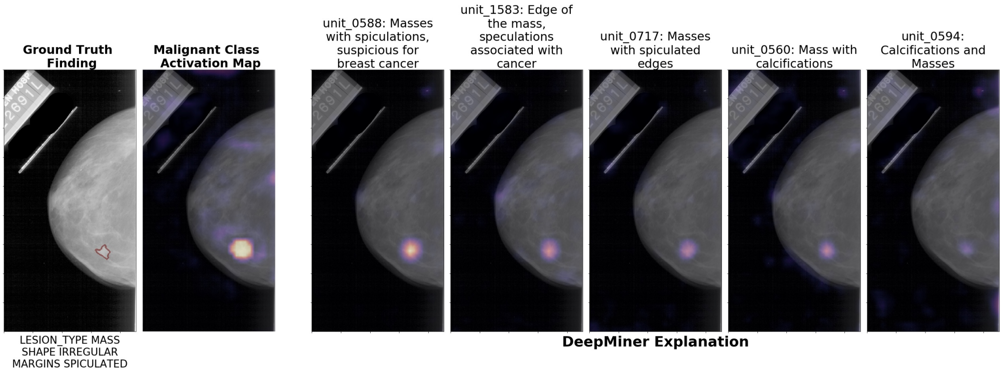
3.1 Classifying the Mammogram

We benchmark our reference network on the test set patches using the area under the ROC curve (AUC) score. Our network achieves AUCs of 0.838 for the normal class (pAUC @ TPR of 0.8 was 0.133), 0.802 for the benign class (pAUC of 0.121), and 0.872 for the malignant class (pAUC of 0.144). This performance is comparable to the state-of-the-art AUC score of 0.88 [19] for single network malignancy on DDSM. For comparison, positive detection rates of human radiologists range from 0.745 to 0.923 [20]. Note that achieving state-of-the-art performance for mammogram classification is not the focus of this work.

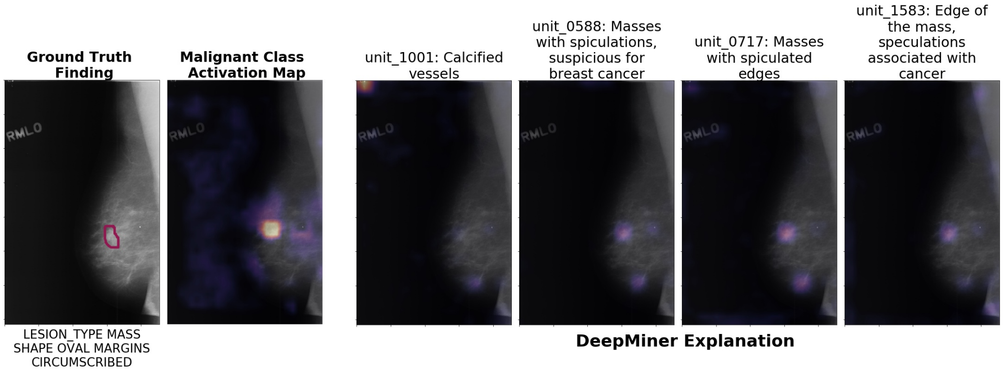
3.2 Explanation for Predictions

Using the DeepMiner framework, we create explanations for the classifications of our reference network on the hold-out set. Figs. 4 and 5 show sample DeepMiner explanations for malignant and benign classifications, respectively. In these figures, the left-most image is the original mammogram with the benign or malignant lesion outlined in maroon. The ground truth radiologist's report from the DDSM dataset is printed beneath each mammogram. The heatmap directly on the right of the original mammogram is the class activation map for the detected class.

In Figs. 4 and 5, the four or five images on the right-hand side show the activation maps of the units most influential to the prediction. In all explanations, the DeepMiner explanation units are among the top eight most influential units overall, but we only print up to five units that have been annotated as part of the explanation.


Ground Truth Finding
Malignant Class Activation Map
unit_0588: Masses with spiculations, suspicious for breast cancer
unit_1583: Edge of the mass, speculations associated with cancer
unit_0717: Masses with spiculated edges
unit_0560: Mass with calcifications
unit_0594: Calcifications and Masses
LESION_TYPE MASS SHAPE IRREGULAR MARGINS SPICULATED
DeepMiner Explanation

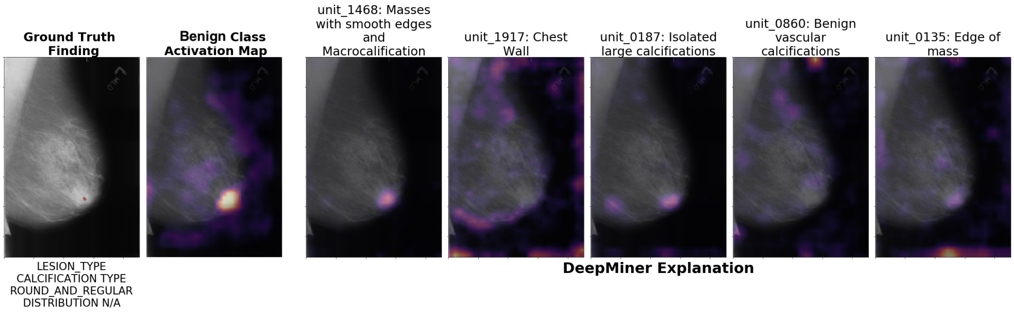
(a) The mammogram above is labeled BI-RADS assessment 4 (high risk), DDSM subtlety 2 (not obvious). Our network correctly classifies the mammogram as containing malignancy. Then, DeepMiner shows the most influential units for that classification, which correctly identify the finding as a mass with spiculations.


Ground Truth Finding
Malignant Class Activation Map
unit_1001: Calcified vessels
unit_0588: Masses with spiculations, suspicious for breast cancer
unit_0717: Masses with spiculated edges
unit_1583: Edge of the mass, speculations associated with cancer
RMLO
LESION_TYPE MASS SHAPE OVAL MARGINS CIRCUMSCRIBED
DeepMiner Explanation

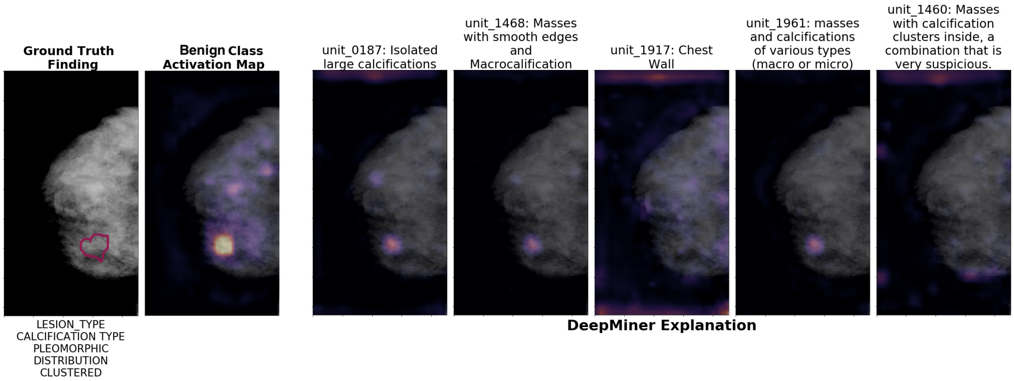
(b) This mammogram is falsely classified by our network as containing a malignant mass, when it in fact contains a benign mass. However, the DeepMiner explanation lists the most influential unit as detecting calcified vessels, a benign finding, in the same location as the malignant class activation map. The most influential units shown here help explain how the network both identifies a benign event and misclassifies it as a malignant event.

Fig. 4: Sample DeepMiner explanations of mammograms classified as malignant. Best viewed in color.

In these examples, the DeepMiner explanation gives context and depth to the final classification. For the true positive classifications in Figs. 4a and 5a, the explanation


Ground Truth Finding
Benign Class Activation Map
unit_1468: Masses with smooth edges and Macrocalification
unit_1917: Chest Wall
unit_0187: Isolated large calcifications
unit_0860: Benign vascular calcifications
unit_0135: Edge of mass
LESION_TYPE CALCIFICATION TYPE ROUND_AND_REGULAR DISTRIBUTION N/A
DeepMiner Explanation

(a) The above image sequence explains a true positive classification of a benign mammogram. The benign mass is quite small, but several unit detectors identify the location of the true finding as ‘mass with smooth edges’ (likely benign) and ‘large isolated calcification’.


Ground Truth Finding
Benign Class Activation Map
unit_0187: Isolated large calcifications
unit_1468: Masses with smooth edges and Macrocalification
unit_1917: Chest Wall
unit_1961: masses and calcifications of various types (macro or micro)
unit_1460: Masses with calcification clusters inside, a combination that is very suspicious.
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC DISTRIBUTION CLUSTERED
DeepMiner Explanation

(b) The above image sequence shows a false positive for benign classification. The mammogram actually contains a malignant calcification. However, the 5th most influential unit detected a ‘mass with calcification clusters inside [...] very suspicious’ just below the location of the ground truth finding.

Fig. 5: Sample DeepMiner explanations of mammograms classified as benign. Best viewed in color.

further describes the finding in a manner consistent with a detailed BI-RADS report. For the false positive cases in Figs. 4b and 5b, the explanation helps to identify why the network is confused or what conflicting evidence there was for the final classification.

To understand how strongly the DeepMiner explanations overlap with the ground truth DDSM annotations, we use the Greedy Matching Score defined by Sharma et al. [21]. Greedy matching calculates a similarity score between a candidate and reference sentence by calculating an average of the cosine similarity between the words in each sentence (as characterized by their embedding in the Word2Vec word embedding space [22]). Averaged over all images in the hold-out set, the Greedy Matching Score for the similarity between the ground truth report and the single most influential unit for each image was 0.627. When the top eight units for each image were considered, the average score was 0.533. Ultimately, the goal of the DeepMiner framework is to discover fine-grained phenomena that may not be part of canonical medical reporting. However, these overlap scores suggest that for well-understood conditions such as breast cancer, our framework is able to discover similar phenomena to those already identified by radiologists.

4 Conclusion

We proposed the DeepMiner framework, which uncovers interpretable representations in deep neural networks and builds explanation for deep network predictions. We trained a network for mammogram classification and showed with human expert annotation that interpretable units emerge to detect different types of medical phenomena even though the network is trained using only coarse labels. We further use the expert annotations to automatically build explanations for final network predictions. We believe our proposed framework is applicable to many other domains, potentially enabling discovery of previously unknown discriminative visual features relevant to medical diagnosis.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
2. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in neural information processing systems. (2014) 487–495
3. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. (2015) 1026–1034
4. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
5. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639) (2017) 115–118
6. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 3462–3471
7. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer (2014) 818–833
8. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR **abs/1312.6034** (2013)
9. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. CoRR **abs/1412.0035** (2014)
10. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. CVPR (2017)
11. Wu, J., Peck, D., Hsieh, S., Dialani MD, V., Lehman MD, C.D., Zhou, B., Syrgkanis, V., Mackey, L., Patterson, G.: Expert identification of visual primitives used by cnns during mammogram classification. In: SPIE Medical Imaging. (2018)
12. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE (2016) 2921–2929
13. Zhang, Z., Xie, Y., Xing, F., Mcgough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
14. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
16. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The digital database for screening mammography. In: Proceedings of the 5th international workshop on digital mammography, Medical Physics Publishing (2000) 212–218
17. Reporting, B.I.: Data system (bi-rads). Reston VA: American College of Radiology (1998)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of CVPR. (2015)
19. Shen, L.: End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv preprint arXiv:1708.09427 (2017)

20. 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) (2009) 641–651
21. Sharma, S., El Asri, L., Schulz, H., Zumer, J.: Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. CoRR **abs/1706.09799** (2017)
22. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA (May 2010) 45–50 `http://is.muni.cz/publication/884893/en`.