Adapting Current CNN Model to Predict Breast Cancer Screening
CS 523 SU 21
Kyle Murray, Vella(Yiting) Liu

Summary of the project

1. Motivation and Rationale

Breast cancer is the second leading cancer-related cause of death for women in the States. Although mammography is the only imaging test that has reduced breast cancer mortality, it has some limitations (e.g. low specificity, associated false positive biopsies, and inconclusive mammograms). Additionally, traditional computer-aided detection (CAD) in mammography doesn't necessarily improve radiologists' diagnostic performances[1]. In addition, CAD is inefficient, as radiologists need to use handcrafted features to mark abnormal sites on a mammogram, and then decide whether these sites are clinically significant or not.

However, deep convolutional neural networks (CNN) can potentially solve these pain points by aiding to evaluate mammography, which improves accuracy of radiologists' diagnoses. CNN can efficiently process a huge number of mammograms, which is cost-effective.

2. Project Overview

Our main reference of work is "Deep Neural Networks Improves Radiologists' Performance in Breast Cancer Screening"[2]. This paper makes several contributions. Primarily, the authors came up with two models. The first model is image-only model, which predicts benign/malignant at breast-level. The second model is image-and-heatmaps model, which highlights the location of biopsied malignant and benign findings at pixel-level. The results demonstrated that the image-and-heatmaps model has a better accuracy than the image-only model. Additionally , the image-and-heatmaps model combined with the reader group, a group consisting of 12 radiologists & 2 medical students that manually predicted breast cancer screenings, outperforms both methods.

Our project is "Adapting Current CNN Model to Predict Breast Cancer Screening". First, we went through the paper and code from https://github.com/nyukat/breast_cancer_classifier in order to understand the question and authors' rationales and methodology. Secondly, we tried to reproduce the same results using sample data from the original dataset. Lastly, we adapted the algorithm and model to a new breast cancer dataset. O ur adapted model and results based on the new dataset are available at https://github.com/kpmurray80/cs512-project-liu-murray.

3. Methodology to predict probability of benign/malignant

We used four sample exams and applied them to both image-only model and image-and-heatmaps model. First, we pre-processed the images by cropping the breast and removing the background and any artifacts (Figure.1).

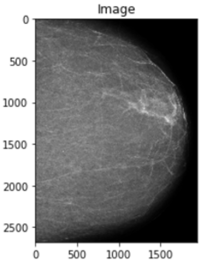


Fig.1. An example of mammogram after pre-processing

Then, we used the model to find the optimal center, which has the largest number of non-zero pixels (most informative area) with a fixed window size. We only focused on the optimal center and its neighboring area instead of the entire image, so this eased the burden on the GPU memory.

After calculating the optimal center, we generated heatmaps visualizing the predictions of probabilities for all patches of images (Figure.2). A brighter area means a higher probability a region is malignant/benign.

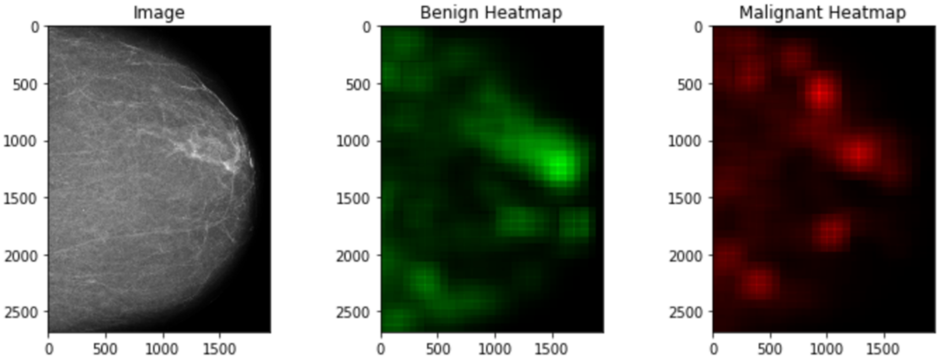
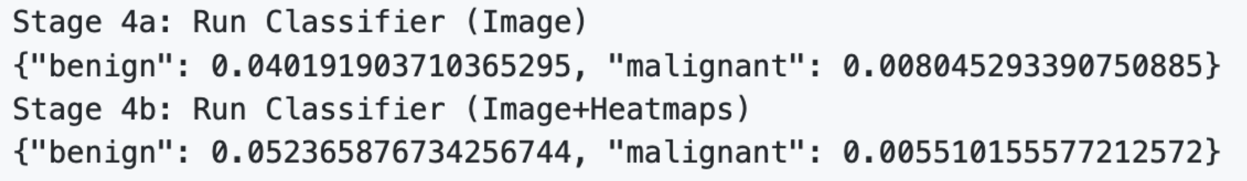


Fig.2. The cropped image, the benign heatmap and the malignant heatmap

Lastly, we ran image-only model and image-and-heatmaps model classifiers, and we got the probabilities of the mammogram being benign and malignant (Figure.3). Although we used the same dataset, the results from both models are slightly different, as the results from the auxiliary task of generating heatmaps are used as input for the model. This improves the prediction accuracy, giving the image-and-heatmaps model closer predictions to the ground truth labels.

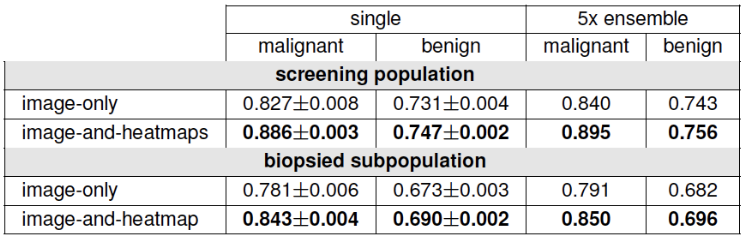


```
Stage 4a: Run Classifier (Image)
{"benign": 0.040191903710365295, "malignant": 0.008045293390750885}
Stage 4b: Run Classifier (Image+Heatmaps)
{"benign": 0.052365876734256744, "malignant": 0.005510155577212572}
```

Fig.3. An example of output for both image-only model and image-and-heatmaps model

4. Evaluate the model in a new dataset

We applied the model to a new breast cancer mammograms dataset. We followed the same approach to pre-process the data and then inputted the processed data into the two models. Then, we evaluated the new dataset using AUC and ROC curves (Figure 4), which are the same evaluation approaches used in the paper. O ur results (Table 2) range from .13 to .20 lower than the results shown in Table 1 from the paper.



	single		5x ensemble	
	malignant	benign	malignant	benign
screening population				
image-only	0.827±0.008	0.731±0.004	0.840	0.743
image-and-heatmaps	**0.886±0.003**	**0.747±0.002**	**0.895**	**0.756**
biopsied subpopulation				
image-only	0.781±0.006	0.673±0.003	0.791	0.682
image-and-heatmap	**0.843±0.004**	**0.690±0.002**	**0.850**	**0.696**

Table.1. AUCs of models on screening and biopsied populations (from reference paper)

	Single	
	malignant	benign
	new dataset	
image-only	0.6206	0.6057
image-and-heatmaps	0.7525	0.5738

Table.2. AUCs of 2 models on the new dataset

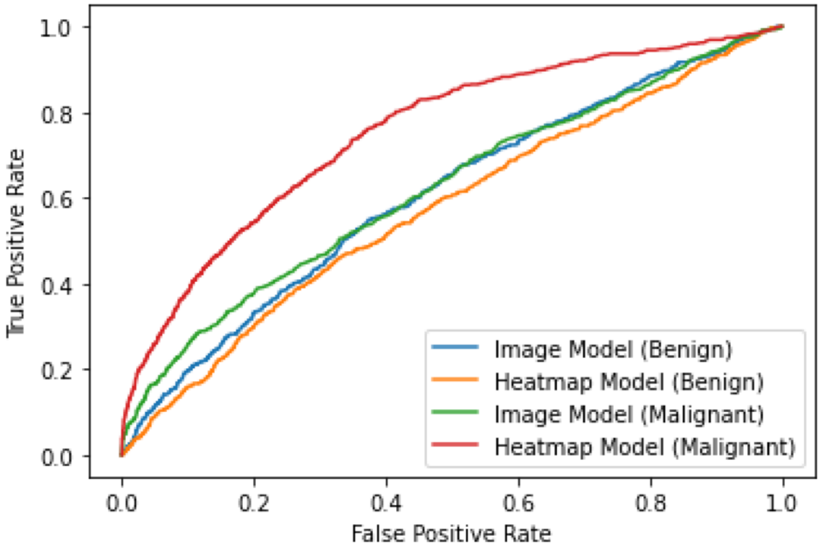


Fig.4. ROC curves on the new dataset.

There are several reasons potentially responsible for the lower AUC results. Firstly, the resolution of the new dataset is lower quality than the original dataset. The dataset in the paper is 16 bits while the new dataset is 8 bits. Additionally, the resolution of images in the original dataset is higher than those images in the new dataset. The decreased quality of images may lead to less accurate predictions.

Secondly, for the image-and-heatmaps model, the researchers sampled 10,000 patches in each epoch, and they classified data into 4 classes: malignant, benign, outside (a patch doesn’t overlap with any malignant/benign findings), and negative (a patch inside a normal breast image). However, the new dataset only has 3 classes: malignant, benign and negative. This class mismatch could cause prediction inaccuracy.

In addition, the new dataset is close to evenly balanced among 3 classes, while the original dataset is extremely unbalanced. In the original dataset, among sampled 10,000 patches, there are 20 malignant patches, 35 benign patches, 5000 outside patches and 4945 negative patches. The new dataset has a total of 7808 mammograms with benign (2684), malignant (2716) and normal (2408) images. The class weight and loss function may not be applicable in the new dataset, which can lead to less ideal prediction results.

Lastly, many images in the new dataset have artifacts, even after preprocessing (Figure 5). We can still see the month and an area of whitened out sensitive information. The addition of these artifacts in the new dataset could have thrown off predictions.

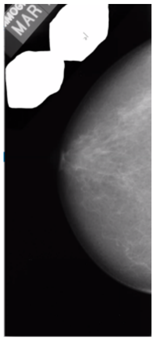


Fig.5. An example of a mammogram after pre-processing

5. Conclusion

In conclusion, we first implemented and visualized four exams from the original dataset in image-only model and image-and-heatmaps model. Then, we adapted the model to a new dataset. Although the AUC and ROC curves results were lower than the results in the paper, our results are still explainable due to lower quality of images, different class distribution and remaining artifacts after pre-processing in the new dataset.

Reference

1. Lehman CD, et al. (2015) Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine* 175(11).
2. N. Wu *et al.*, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," in *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184-1194, April 2020, doi: 10.1109/TMI.2019.2945514.