Breast Cancer Classification

Nhung Le (nhl256), B V Nithish Addepalli (bva212), Ravi Choudhary (rc3620)

May 8, 2019

Abstract

The first step in identifying breast cancer requires inspection of mammogram scans to find existence of lesion and its pathology as Benign/Malignant. With a lot of women having high probability of breast cancer it will be really helpful to radiologists if we can help speed up this process of inspecting mammograms to find lesions and their nature. Use of deep learning methods to identify lesions can significantly help and support the current radiologists. We use various deep learning methods with residual connections and transfer learning approaches for this classification task. In addition to some very popularly known architectures, we've experimented with our very own custom deep CNN architecture which performs better than a few of them. We also show heatmaps with bounding box highlighting the Region of Interest ROI.

1 PROBLEM MOTIVATION

Breast cancer is the second most common cancer in women worldwide. About 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime. The five year survival rates for stage 0 or stage 1 breast cancers are close to 100%, but the rates go down dramatically for later stages: 93% for stage II, 72% for stage III and 22% for stage IV. Human recall for identifying lesions is estimated to be between 0.75 and 0.92, which means that as many as 25% of abnormalities may initially go undetected. Mammogram screening is one of the most common diagnosis method for breast cancer. Multiple traditional machine learning and newly developed deep learning models have been developed to assist radiologists in accurately diagnose breast cancers early on. We propose a custom model that leverages the residual learning framework and transfer learning method to further improve models' capability to predict breast cancer.

2 DATA

2.1 DATA SOURCE

The data set CBIS-DDSM [1] [2](Curated Breast Imaging Subset of Digital Database for Screening Mammography) is an updated and standardized version of the DDSM, which has 2,620 scanned film mammography studies. The images went through a thorough preparation process including removal of questionable cases, image decompression, image processing, image cropping, and mass segmentation.

The dataset has 1,541 calcification studies and 1,318 mass studies. Each study has full mammogram image, crops of abnormalities (i.e., abnormalities were cropped by determining the bounding rectangle of the abnormality with respect to its ROI), and ROI images. Note that the lesion segmentation algorithm was applied to ROI masses images to identify more accurate ROIs for masses. The size of the ROI images in more than 10% of the cases did not match with the mammogram images which reduced the reliability of the ROI and prevented us from using them in our model. Mammogram images have different width and length, their width ranges from 2500 to 5500 pixels and the length varies from 4000 pixels to 7500 pixels. The dataset also has CSV files with the description of calcification and masses. These were not used for the prediction, since these descriptions were obtained with the help of a radiologist and would not be available during actual test time. The scans consisted of both CC and MLO views of patient's mammogram. Most of the patients only had mammograms of one breast. Given below are some of the statistics of the data.

Metric	Value	Train	Val	Test
# Patients	1,645	1,013	270	349
# Scans	3,563	2,031	562	645

2.2 DATA PREPARATION

We combined all the patient studies into one file and sampled the train, validation and test data sets from it. The sampling was done based on patient id to avoid any overlap that would result in leakage. The data was divided into 70, 20 and 10% for training, validation and test set respectively. Since the images were of different sizes, different techniques were used for batching the images. For training with batch size = 1, we used the images without any resizing. For training with batch size > 1, images were either padded to the same size $7500 * 5500$ i.e., maximum length and width of images or they were resized using transforms. The data was augmented by flipping the scans about the horizontal/vertical axis randomly. The images were also normalized using training sample mean and variance. The data was normalized to $0 - 1$ range by dividing with the 65535 since this is the maximum number for the data type of the pixels of the image - $uint16$.

	Train	Val	Test
# Malignant	888	243	260
# Benign	1,143	319	385

3 MODEL BUILDING PROCESS

3.1 EXPERIMENT SETUP

Since we had four different types of labels - Malignant Masses, Benign Masses, Malignant Calcification and Benign Calcification, we decided to work on two sets of model, one for binary-classification model with labels being Malignant (1) vs. Benign (0), and the other one for the original 4 classes.

3.2 LOSS FUNCTION

We're using Cross-Entropy Loss as the task in hand is a classification problem with the non-independent classes and balanced dataset. Since, the classes are dependent among each other we didn't use Binary Cross-Entropy loss.

3.3 EVALUATION METRICS

We evaluated our models primarily in terms of Accuracy and AUC (Area Under the ROC curve) for classification tasks on the image level. We evaluated micro AUC in case of 4 classes. Because of balanced dataset, micro and macro AUC of all of our models were very close.

3.4 HYPER-PARAMETERS SELECTION

- **Learning Rate:** We applied manual method and learning scheduler from Pytorch to get the best learning rate for each model. Learning rates in general range from 0.001 to 0.00005. 3 Learning Schedulers were tried - StepLR, CosineAnnealingLR and LRonPlateau. StepLR gave much simpler controls and better results.

- **Mini-batch Size:** Given the memory constraints, we could try batch size of up to 4 for original image size. We also tried training with mixed precision[3] to accommodate bigger batch size. This resulted in worse results. Comparison of the same network with training on batch size 1 image without padding vs batch size 4 with padded images showed better performance on 4.

- **Optimization:** Adam outperformed Stochastic Gradient Descent and Root Mean Square Propagation (RMS Prop). This is because Adam calculates an exponential moving average of the gradient and the squared gradient, while controlling the day rates of these moving averages, thus taking the benefits of both AdaGrad and RMSProp.

- **Activation:** Given small sizes of mass or calcification lesions, we worried that random initialization might result in exacerbating the problems associated with ReLU i.e., dying ReLU. Thus, we tried ELU [4] - combination of Batch Norm with ReLU while solving the dying ReLU problem, Leaky ReLU, and PReLU - which introduces extra parameters and has been shown to perform better than ReLU [5]. PReLU proved to have a lot of learning capacity as it overfits on the sample data set. Performance on PReLU based network is also shown below.

- **Regularization:** In order to get better generalization performance with our model, we used weight decay and dropout. We also used early stopping based on validation performance. Most of our models use Batch Norm which in itself provides regularization.

3.5 DEEP LEARNING MODELS

3.5.1 TRANSFER LEARNING MODELS

We started with Res-Net 18, Res-Net 34 and Res-Net 50 [6] given their popularity in image classification task [7]. In order to take advantage of transfer learning, we ran the Res-Net models pre-trained on ImageNet data [8], using same training and validation transformation (e.g., image resized to $2048 * 2048$). The second model was a Res-Net 18 designed to handle the images at the full resolution. We also trained Inception Resnet v2 on our data.

In order to improve the performance from ResNet architecture, we decided to try a few changes to model architecture. To test the change in learning capacity of the new model, the model was first tested on a sample dataset. Instead of using 1x1 convolution kernel with a stride of 2 for down sampling in residual connection, we used concatenated max and average pooling layers which we believed to be a better way to pass Identity connection. In addition, we tried to switch the position of Batch Normalization to after the activation since the presence of ReLU activation after Batch Normalization could have resulted in discarding of more features than required. We had also tried to do $1x1$ convolution after max pooling. None of the models failed to overfit on the sample data.

We attempted to try Densenet architecture as the final block with Resnet architecture before the classifier, assuming that this would increase the number of features with representation of the masses/calcification. This model too failed to overfit on the sample data, proving lack of learning capacity.

3.5.2 CUSTOM CNN

The model architecture was designed to work on the non-resized images through aggressive down sampling in the initial layer. In order to draw maximum representation power from the model and to keep the largest batch size we used p40 GPU for training as it has 22 GB of usable memory. The basic block for the model is shown below.

Due to the memory constraint, number of convolution kernels in the first layer cannot exceed 16. Since all the later convolution kernels depend on the features captured in the initial layer, we had to ensure that the kernels remained linearly independent from each other. One way was to try CReLU [9], which uses concatenated versions of negative and positive activation of feature maps. Memory constraints made it infeasible to incorporate CReLU in our model.

Regarding the initial convolution kernel size, both sizes of 5x5 and 7x7 did not outperform the kernel size 3x3. The ultimate choice of 16 kernels of 3x3, like the choice of 64 kernels of 7x7 in Res-Net, would have reduced the number of linearly dependent filters but would have still allowed a few to be linearly dependent.

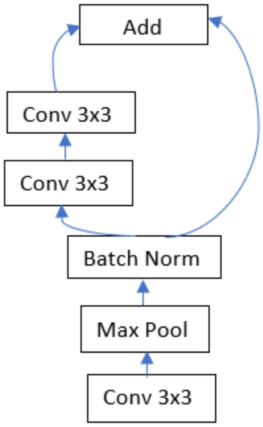


Figure 1: Custom CNN Block

We decided to first train the custom model on the Chest X-ray dataset for predicting Pneumothorax and Cardiomegaly. The weight from the derived model was used for weight initialization for the Mammograms dataset as mammograms are X-rays too, thus giving a more relevant transfer learning approach.

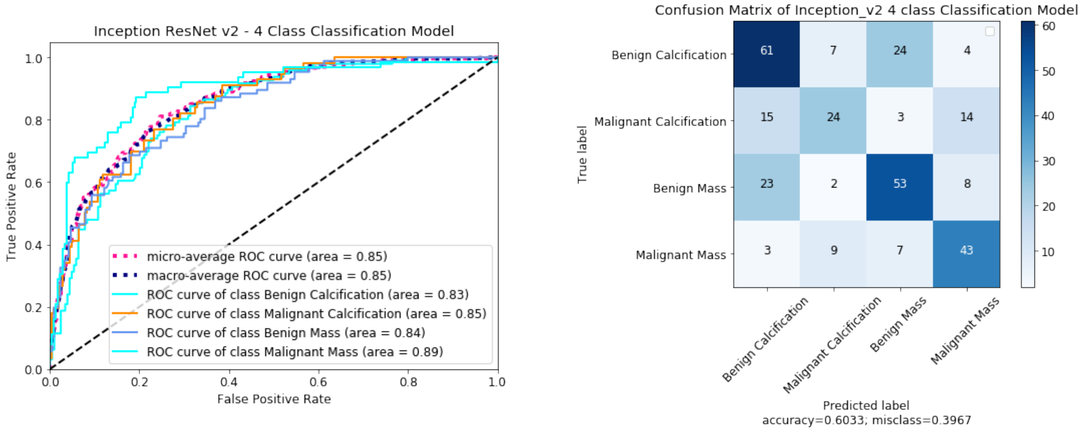
4 EXPERIMENTS AND RESULTS

Resnet based models with pretraining on Imagenet seem to work well on the dataset. The best models came out to be Inception Resnet v2 and Resnet50. These models were trained on images resized to 1024x1024. This seems to indicate that models with very high learning capacity either through depth or width are much better for this problem given the low signal to noise ratio. The table below shows performance of all the models on test dataset.

Architecture	Transfer Learning	2 Classes		4 Classes	
		AUC	Accuracy	*AUC	Accuracy
ResNet18 (full image)	-	-	-	0.63	0.36
Custom CNN with PReLU	-	-	-	0.64	0.35
ResNet18	Imagenet	0.71	0.65	0.53	0.32
ResNet34	Imagenet	0.67	0.65	-	-
Custom CNN	Chest X-Ray	0.64	0.56	0.73	0.43
ResNet50	Imagenet	0.82	0.72	-	-
Inception Resnet V2	Imagenet	0.84	0.74	0.85	0.60

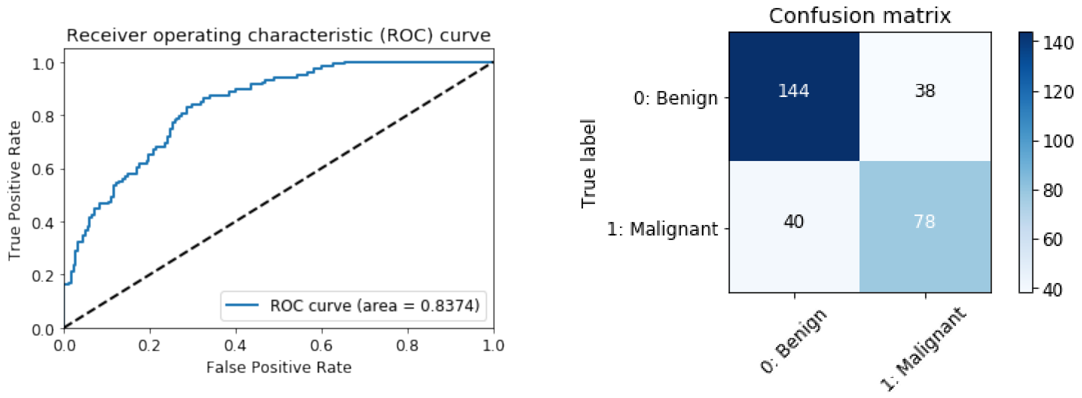
* AUC presented for 4 class is Micro-AUC

Inception Resnet v2 Performance metrics: Below are the confusion matrices and the plots showing the class wise AUC score for both 2 & 4 class classification models, along with Macro and Micro AUC score for 4 class classification model.



True label \ Predicted label	Benign Calcification	Malignant Calcification	Benign Mass	Malignant Mass
Benign Calcification	61	7	24	4
Malignant Calcification	15	24	3	14
Benign Mass	23	2	53	8
Malignant Mass	3	9	7	43

Figure 2: Inception ResNet v2 4 class classification model AUC plot and Confusion Matrix



True label \ Predicted	0: Benign	1: Malignant
0: Benign	144	38
1: Malignant	40	78

Figure 3: Inception ResNet v2 2 class classification model AUC plot and Confusion Matrix

5 DISCUSSION AND FUTURE WORK

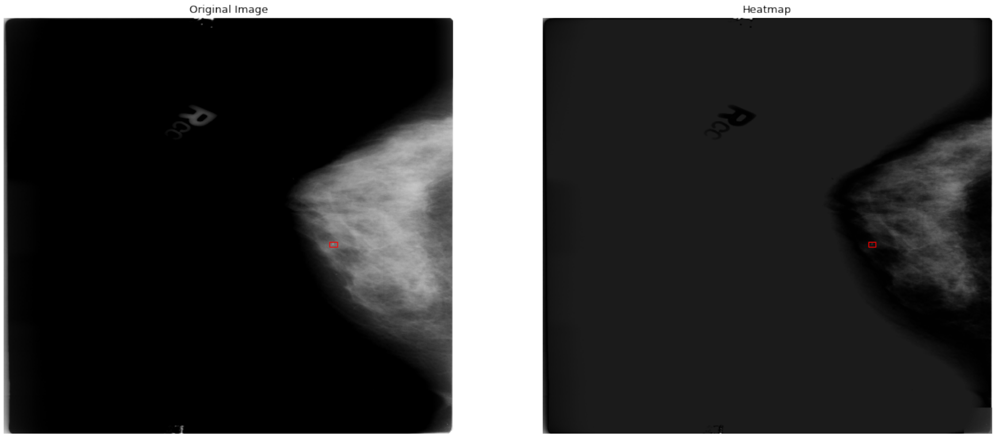


Figure 4: Heat Map showing the model's incorrect prediction

We've generated heatmaps to further identify the cases where our model is failing to classify correctly. In the heatmap below (red bounding boxes highlight the actual ROI), we can see that the model is failing to identify the lesion correctly. This may be because the size of the lesion is so small compared to the image that the signal-to-noise ratio becomes very low and hence our model fails. We believe that a patch based classification technique used in [7] might be a better way to identify the lesion and classify these images correctly.

References

[1] *Data Source: CBIS-DDSM (Curated Breast Imaging Subset of Digital Database for Screening Mammography).* https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#.

[2] *A curated mammography data set for use in computer-aided detection and diagnosis research.* https://www.nature.com/articles/sdata2017177.

[3] Paulius Micikevicius et al. *Mixed Precision Training*. 2017. arXiv: 1710.03740.

[4] Djork-ArnÃ© Clevert, Thomas Unterthiner, and Sepp Hochreiter. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. 2015. arXiv: 1511.07289.

[5] Kaiming He et al. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015. arXiv: 1502.01852.

[6] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385.

[7] Nan Wu et al. *Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening*. 2019. arXiv: 1903.08297.

[8] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009.

[9] Wenling Shang et al. *Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units*. 2016. arXiv: 1603.05201.