E9 253: Neural Networks and Learning Systems **Fall Semester 2020**

Automatic Breast Cancer Detection in Mammograms

Course Instructor: Prof. Shayan Garani *Roshan Yadav, Bhushan Deo*

1.1 Motivation

Breast cancer is the second most common cancer in women both in the developed and the developing world. The incidence of breast cancer is increasing in the developing world due to increase life expectancy, increase urbanization and adoption of western lifestyles. According to the World Health Organization (WHO), the number of cancer cases expected in 2025 will be 19.3 million cases. In India, one woman is diagnosed with breast cancer every 4 minutes, while one woman dies of breast cancer, every 13 minutes.[1] Mammography is currently one of the important methods to detect breast cancer early. The magnetic resonance imaging (MRI) is the most attractive alternative to mammogram. However, the MRI test is done when the radiologists want to confirm about the existence of the tumor. The drawback of the MRI is that the patient could develop an allergic reaction to the contrasting agent, or that a skin infection could develop at the place of injection. It may cause claustrophobia.

1.2 Introduction

It is important to detect breast cancer as early as possible. Mammogram can serve as an elementary tool to detect breast cancer in early stages but analyzing mammogram is a nontrivial task, and decisions from investigation of these kinds of images always require specialized knowledge. So computer based automatic breast cancer detecting techniques can help the specialist (doctors and physicians) to make more reliable decisions.[2] In this report, a new methodology for classifying breast cancer using two different combination of deep learning model (hybrid deep convolutional neural network) and some image preprocessing techniques (Image Augmentation) are introduced. Here two-step process are used to detect breast cancer. First a well-known DCNN architecture named ResNet50 is used and is fine-tuned to identify the cancerous breast mammography images and second step involves a ROI(Region of Interest) based simple CNN model which used to detect the location of patch which contain benign and malignant mass tumors in cancerous mammography images. The following publicly available datasets are used to train our model (1) The mini-MIAS database of mammograms (2) the digital database for screening mammography (DDSM)[3]; and (3) the Curated Breast Imaging Subset of DDSM (CBIS-DDSM).[4] For any CNN model training on a large number of data gives high accuracy rate. Since, the biomedical datasets contain a relatively small number of samples due to limited patient volume. Accordingly, in this project data augmentation method is used for increasing the size of the input data by generating new data from the original input data. There are many forms for the data augmentation; the one used here is combination of rotation, affine transformations and resizing. First we tried simple DCNN full-mammogram architecture on full mammogram but accuracy was very poor (around 62 %) but after using DCNN architecture named ResNet50 on full mammogram accuracy achieved was 92% and when cropping the ROI manually from the mammogram accuracy achieved was 96% for the ROI samples obtained from Image Augmentation techniques.

1.2.1 Dataset

In this project we mainly deal with the following kinds of mammograms:

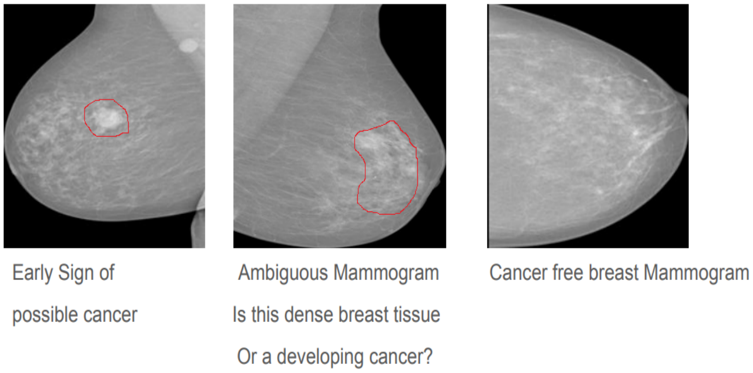


Figure 1.1: Typical Breast Cancer Mammograms

There are other indicators of breast cancer also but those are less significant mainly a mass can be either benign or malignant. The difference between benign and malignant tumors is that the benign tumors have round or oval shapes, while malignant tumors have a partially rounded shape with an irregular outline. In addition, the malignant mass will appear whiter than any tissue surrounding it.We formed the ROI dataset by manually cropping the regions of interest in the complete mammogram image. Both datasets were oversampled by standard affine transformations.

1.2.2 Methodology-ROI patch dataset

We started testing the dataset with a simple CNN model as below:

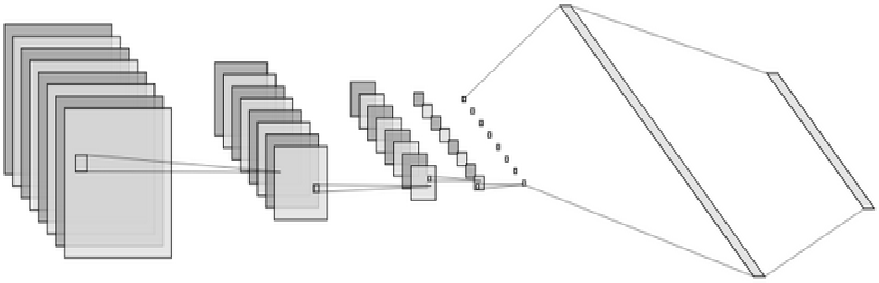


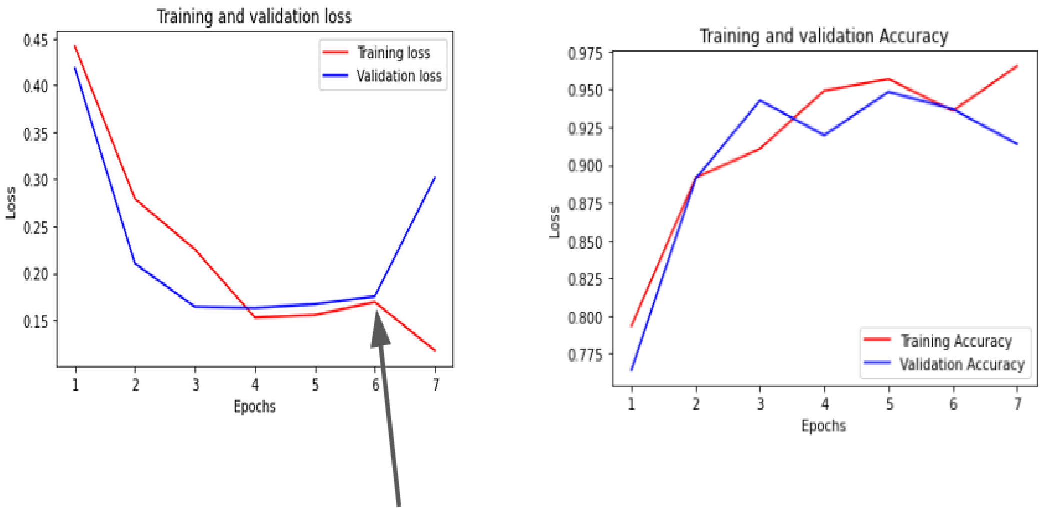
Figure 1.2: CNN model

Briefly our model is as follows:

- Conv->Max-Pool->Conv-¿Max-Pool->Conv->Conv->Max-Pool->Conv->Max-Pool->Conv->Flatten->Dense. Each Convolution layer has 3x3 masks and 64 depth. Each max-pool downsizes by 2.

- We have used adam optimizer, relu in all intermediate layers, softmax in the final error. Categorical cross-entropy is used as a loss metric.

Our model has only 189,570 parameters and hence results in fast computation of the response. We used Early stopping in order to prevent the model from overfitting, and we outline the error trajectories as follows:



Early stopping with best selection of weights ensured that these weights were selected and overfitting was reduced

Figure 1.3: Accuracy and Loss trajectories

The following is the confusion matrix for the ROI test dataset:

	Pred-NO	Pred-Yes
Act-NO	121	11
Act-Yes	4	125

As we can see we obtained more than 90% accuracy for both the training as well as validation data, so we decided to stick with this model for further evaluations.

1.2.3 Methodology-Complete Mammogram Dataset

Using the same simple CNN model we tried to test upon the complete mammogram dataset but the results were unsatisfactory:

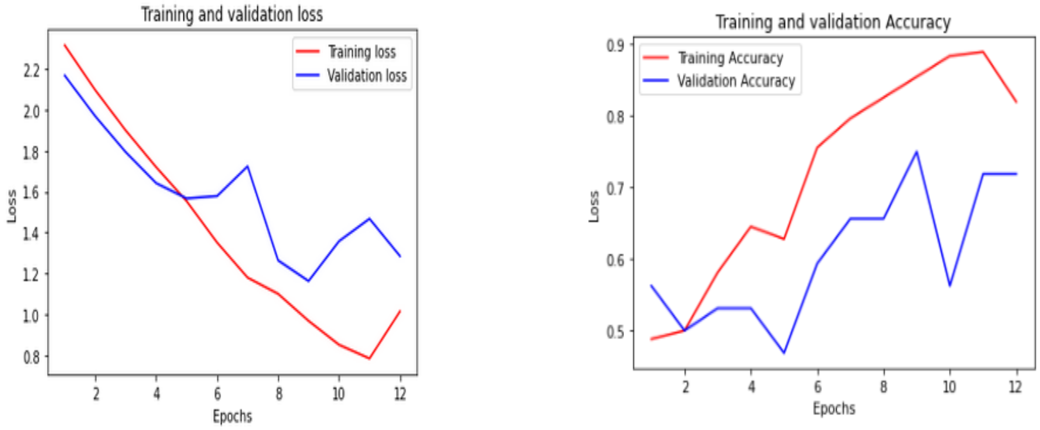


Figure 1.4: Accuracy and Loss trajectories

So we decided to use transfer learning in order to learn the features in the images and just train the last fully connected layer of the pretrained model. The model we used for the task was ResNet50 which can be found here.
The deep model has skip or residual connections which helps to better learn the features of the underlying image. We tuned the last fully connected layer (ANN) of the model according to our classification labels:

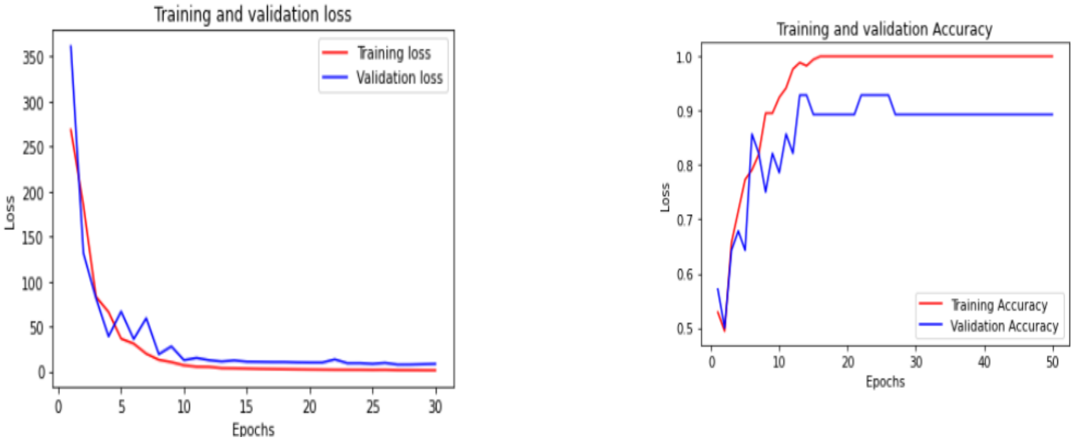


Figure 1.5: Accuracy and Loss trajectories

The dataset was found to be quite noisy but overall the ResNet50 model performed quite well. Performance depends upon various factors such as weight initialization, early stopping, order of the data,etc. The model was trained with patience-3 and best weights chosen. The confusion matrix is as indicated below:

	Pred-NO	Pred-Yes
Act-NO	17	5
Act-Yes	1	20

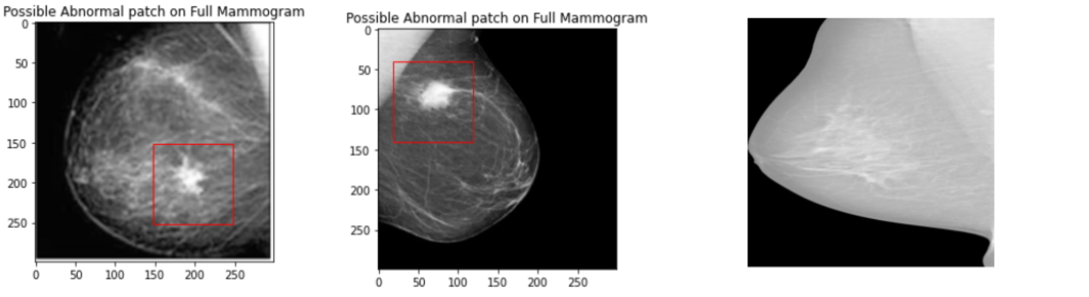
False Negatives are the least among misclassifications, is what we should ideally want in biomedical analysis.

1.3 Automatically Detecting Lesions

The procedure to determine lesions automatically given a mammogram image is as follows:

- Get test mammogram image
- Convert it to a GRAYSCALE IMAGE and save it
- Do Mean subtraction and standardization
- Replicate it to make number of channels as 3 (RGB)
- Then pass test image to the ResNet50 Model and observe predicted output
- If predicted value is greater than threshold
- Then select GRAYSCALE IMAGE of full mammogram (step 2) and extract 1000 patches
- Do Mean subtraction and standardization for each patch
- Then use ROI based simple CNN model to detect lesion automatically

We briefly outline some of the experimental results:


Possible Abnormal patch on Full Mammogram

Possible Abnormal patch on Full Mammogram

Lesion automatically

Detected by our model

[1.0000000e+00, 8.6999463e-26]

No Cancer in the submitted Mammogram

Figure 1.6: Automatically Detecting Lesions

1.4 Conclusion

We were able to automatically detect lesions in the complete mammogram with our 2-step strategy. The tuned model weights are present here: https://drive.google.com/drive/folders/1ykpp0FZnlXN2pT3YYbodlEer4ndsFhnz?usp=sharing

Compact code to automatically detect lesions is present here:
https://colab.research.google.com/drive/1LU6ydPMY$_{m}ovFdR4iikwKyi1m3Db7srW?usp = sharing$

References

[1] BREAST CANCER INDIA , https://www.oncostem.com/blog/alarming-facts-about-breast-cancer-in-india/.

[2] SHEN, L. , Shen, L., Margolies, L.R., Rothstein, J.H. et al. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep 9, 12495 (2019). https://doi.org/10.1038/s41598-019-48995-4

[3] KAGGLE DDSM , https://www.kaggle.com/skooch/ddsm-mammography

[4] BREAST CANCER DATASET ,https://drive.google.com/drive/folders/0B-7-8LLwONIZZm1pQWdyak5Od28