Breast Cancer Identification using Convolutional Neural Network (April 2018)

Palak Agrawal Vinoth Punniyamoorthy Ayushi Sachdeva
{agrawal.pala, punniyamoorthy.v, sachdeva.ayu}@husky.neu.edu
INFO – 7390 Information Systems – College of Engineering
Northeastern University
Boston, MA

Abstract -

B*reast* cancer is the leading cause of cancer-induced mortality among women with 2.6 million and men with 2550 new cases of invasive and in-Situ diagnosed and 40,920 deaths per year. In past, a determination has been at first performed utilizing clinical screening took after by histopathological investigation. Robotized arrangement of growths utilizing histopathological pictures is a challenging undertaking of exact discovery of tumor sub-types. This procedure could be teamed up with deep learning approaches, which might be more dependable and conservative contrasted with conventional techniques. To demonstrate this rule, we applied fine-tuned pre-trained deep neural networks and retrained a model based on Inception V3.
To test the approach, we worked with 10, 239 mammographic images having DICOM format downloaded from Cancer Imaging Archives wiki ([1]). Our framework accurately detected on average 75% of the three results of breast cancer including benign, malignant, and benign without callback. With 4 layers of neural networks, we re-trained our Inception V3 model for the Breast Cancer category. Then, we used transfer learning technique to classify the images.
Availability: Source codes, guidelines and data sets are temporarily available on google drive upon request before moving to a permanent GitHub repository.

I. Introduction

Cancer is one of the main causes of death worldwide. Among the cancer types, breast cancer is second most common for women, excluding skin cancer. Routine mammography is standard for preventive care and detection of breast cancer before biopsy. Biopsy is the diagnostic procedure that can determine if a suspected region is cancerous.

Breast cancer are identifiable from mammograms thanks to the different X-ray absorption rates of normal and abnormal tissues. Mammography entails exposing a patient's breasts to low levels of X-ray radiation. Tumors can appear as masses, distortions or micro-calcifications on mammograms. However, it is still a manual process, prone to human error due to high variability in mass appearance.

Because of the dense breast tissue, the tumor mass may overlap with the dense tissue, creating masking effect and making mammography less sensitive. Manual classification by radiologists still incurs a high recall rate and requires years of experience on the part of the radiologist. This high recall rate results in an abundance of additional diagnostic tests, including biopsy, and thus contributes to increased health-care costs, pain, anxiety for the patients themselves. It is therefore necessary to improve diagnostic accuracy of mammography in terms of both sensitivity for early detection of breast cancer and specificity to keep the recall rate low.

Recent advances in image processing and machine learning techniques allow to build Computer-Aided Detection/Diagnosis (CAD), which are established for robust assessment of medical image-based examination, can assist to be more productive and consistent in diagnosis. Conventional classification approaches rely on extensive pre-processing and manual extraction of specific visual features before classification usually designed for a specific problem based on field-knowledge. To overcome the many difficulties of the feature-based approaches, deep learning methods are becoming important alternatives.

Among the different approaches, the Convolutional Neural Network (CNN) introduced by has been widely used to achieve results in different pattern recognition and image classification problems. CNNs allow to reduce the field-knowledge needed to design a

This research paper is about Breast Cancer identification using Convolutional Neural Network as well as using Transfer Learning concepts. It is submitted on April 21^{st,} 2018 for review to Prof. Nick Bear Brown. This work was a part of final project and research under course Advance Data Science (INFO 7390).

Palak Agrawal is a graduate student majoring in Information Systems at Northeastern University, Boston, MA 02115, USA (e-mail: agrawal.pala@husky.neu.edu)

Ayushi Sachdev is a graduate student majoring in Information Systems at Northeastern University, Boston, MA 02115, USA (e-mail: sachdeva.ay@husky.neu.edu)

Vinoth Punniyamoorthy is a graduate student majoring in Information Systems at Northeastern University, Boston, MA 02115, USA (e-mail: punniyamoorthy.v@husky.neu.edu)

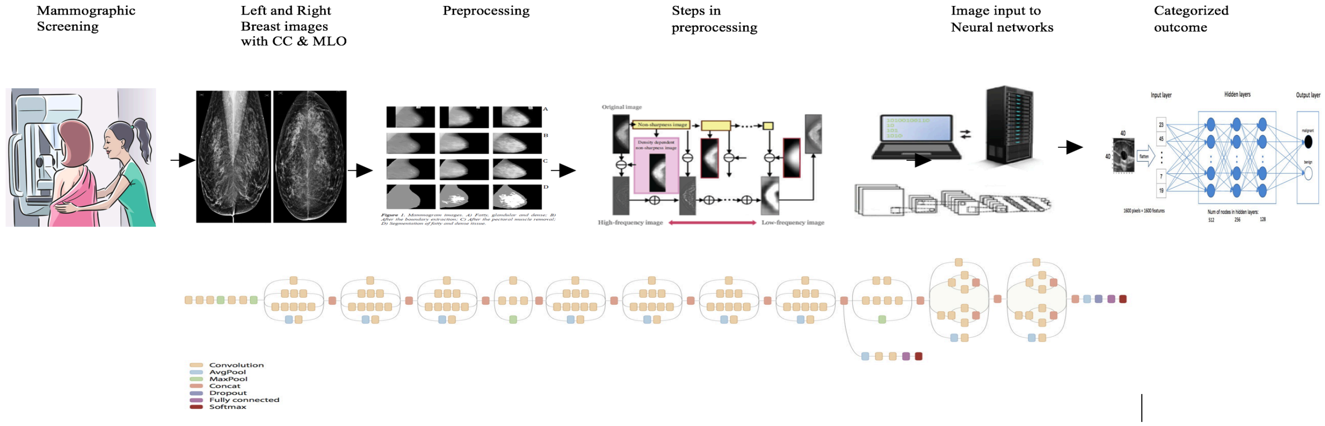


Figure 1. Project work-flow; Data gathering, Image capturing, and Deep learning approaches using Inception V3 model and Transfer Learning. Data gathering included downloaded image from Curated Breast Imaging Subset of Digital Database for Screening Mammography database converted in JPEG format. Preprocessing is the next step followed by deep learning techniques to extract unique presentation for each separated cancer input.

classification system. The convolutional process can simplify an image containing millions of pixels to a set of small feature maps, thereby reducing the dimension of input data while retaining the most-important differential features.

Transfer learning is the re-use of information obtained during the training phase of a previous project. Two popular transfer-learning methods involve (a) fine-tuning the parameters in certain layers of the trained CNN, or (b) using the trained CNN to calculate the feature maps of new types of data.

This study utilizes Inception v3 ([4]), trained with ImageNet. Considering the fact that mammograms differ dramatically from the images in the ImageNet dataset, the trained model was used only to obtain the feature maps.

II. Approach

In this research, we developed and presented an exact and dependable computer based strategy enabled with deep learning approaches to identify breast cancer from mammographic images in DICOM format downloaded from Curated Breast Imaging Subset of Digital Database for Screening Mammography database.

TABLE I
STATISTICS OF IMAGES

COLLECTION NAME	STATISTICS
Number of Patients	6671
Number of Studies	6775
Number of Series	6775
Number of Images	10239
Image Size (GB)	163.9

Our project contains below steps:

a) Image acquisition and conversion from DICOM to JPEG.
b) Deep learning Preprocessing to sanitize the images.
c) Transfer learning and fine-tuning pre-trained.
d) Hierarchical feature extraction and classification with Google's Inception V3.

All steps have been illustrated in Figure 1.

III. Data and Methods

A. DATASET

Dataset was collected from Curated Breast Imaging Subset of Digital Database for Screening Mammography database ([1]). It is a database of 2, 620 mammographic images which were decompressed and converted to DICOM format. DICOM is Digital Imaging and COmmunications in Medicine used for medical examined images. Resources on remote server were downloaded through Java Network Launch Protocol (JNLP). Database contains **normal**, **benign**, and **malignant** cases with verified pathology information shown in figure 2. This dataset has 10, 239 mammographic images with 64-bit optical density values collected from 6775 patients who were selected for breast cancer identification. This dataset is a collection of Left and Right breast mammograms taken with two image views

1. Craniocaudal (CC)
2. Mediolateral Oblique (MLO).

It comprises 1, 429 Benign cases, 628 Normal cases and 1, 457 Malign cases. The data were split for all mass cases and all calcification cases separately. It should be note that approximately 80% of available data were randomly chosen to construct the training set. The remaining 20% of the data were used for performance evaluation or as a test data.

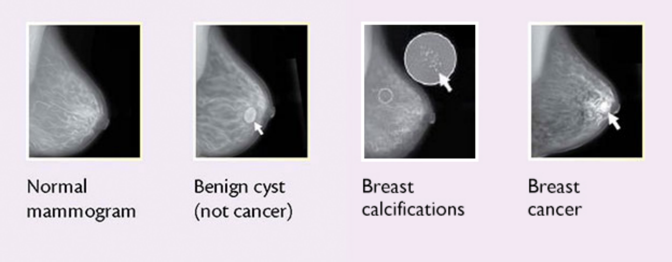
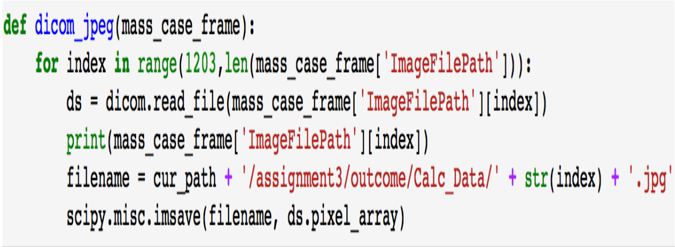


Figure 2

B. IMAGE CONVERSION

The MRI scanning facilities typically does not contain any image files in traditional formats that can be opened up by your default image viewing program. It contains a list of DICOM files, which can difficult to read sometimes. Thus, infamous Digital Imaging and Communications in Medicine (DICOM) standard the de-facto solution to storing and exchanging medical image-data can be converted to JPEG format using PYDICOM and SCIPY library methods.



```
def dicom_jpeg(mass_case_frame):
    for index in range(1203,len(mass_case_frame['ImageFilePath'])):
        ds = dicom.read_file(mass_case_frame['ImageFilePath'][index])
        print(mass_case_frame['ImageFilePath'][index])
        filename = cur_path + '/assignment3/outcome/Calc_Data/' + str(index) + '.jpg'
        scipy.misc.imsave(filename, ds.pixel_array)
```

Below two functions are used to convert DICOM to JPEG

a. dicom.read_file : It reads the complex files into natural pythonic structures for easy manipulation

b. scipy.misc.imsave : This function uses bytescale under the hood to rescale images to use the full (0, 255) range if mode is one of None, 'L', 'P', 'l'. It will also cast data for 2-D images to uint32 for mode = None

C. PRE-PROCESSING STEPS

Gathering data and converting DICOM to JPEG images is followed by Pre-processing the images for below reasons,

a. Extracting boundaries
b. Removing Pectoral muscle
c. Removing artifact number (LMLO, RCC, etc.)

OpenCV library is used to support a lot of algorithms related to Computer Vision and Machine Learning. OpenCV-Python is the Python API of OpenCV. All the OpenCV array structures are converted to-and-from Numpy arrays. So, whatever operations we can do in Numpy, can be combined with OpenCV.

Below are the methods followed as part of pre-processing mammographic images:

1. Converting image to single channel grey scaled image.

   cv2.IMREAD_GREYSCALE

2. Smoothing the image and removing the noise

   cv2.medianblur

3. Thresholding the image and suppressing the artifact number.

   cv2.Threshold-THRESH_BINARY

4. Enhancing the contrast and again thresholding the image to black and white.

   cv2.equalizeHist, cv2.Threshold - THRESH_BINARY

5. Using Watershed methods like **Erosion** and **Dilation** to remove the pectoral muscle feature.

   cv2.erode, cv2.dilate, cv2.morphologyEx - MORPH_OPEN, cv2.bitwise_and

D. CONVOLUTIONAL NEURAL NETWORK

For this project, we constructed our own Convolution Neural Network to identify the breast cancer is benign, normal or malign and compared the results obtained with those obtained from the pre-trained models.

Our CNN is stacked with the following layers. The functionality of each layer is briefly explained below:

a. Convolution Layer:

The primary purpose of Convolution layer is to extract features from the input image. Here, filters act as Feature detectors and the value of the filter, is in fact, not manually provided but the machine chooses the suitable value by training and changing its weights

b. Pooling Layer:

Pooling layer reduces the dimensionality of each feature map but retains the most important information. Pooling layer can be of different types: Max, Average, Sum etc.

c. Dense Layer:

Dense layer comprises of the traditional fully connected neural network which classifies the input to their respective classes

d. Batch Normalization Layer:

A batch normalization layer normalizes each input channel across a mini-batch. The layer first normalizes the activations of each channel by subtracting the mini-batch mean and dividing by the mini-batch standard deviation. Then, the layer shifts the input by a learnable offset β and scales it by a learnable scale factor γ. Using batch normalization layers between convolutional layers and nonlinearities, such as ReLU layers, speeds up training of convolutional neural networks and reduces the sensitivity to network initialization.

e. Fully Connected Layer:

Finally, after several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers. A fully connected layer takes all neurons in the previous layer and connects it to every single neuron it has.

Input layer: (128,128,3)
128 - refers to no. of pixels in each dimension
3 - refers to the RGB values (channels) for each pixel

Layer 1:
Conv-1: 16 filters of size 3 x 3 with ReLU activated
Pool-1: Max pool with pool_size 1
Batch-1

Layer 2:
Conv-2: 32 filters of size 3 x 3 with ReLU activated
Pool-2: Max pool with pool_size 2
Batch-2

Layer 3:
Conv-3: 64 filters of size 3 x 3 with ReLU activated
Pool-3: Max pool with pool_size 2
Batch-3

Layer 4:
Conv-4: 128 filters of size 3 x 3 with ReLU activated
Pool-4: Max pool with pool_size 1
Batch-4

Batch Size: 32
Dropout: 50%
FC-1: 128 neurons with activation function ReLU

E. Transfer Learning with Inception V3

Transfer learning is defined as exporting knowledge from previously learned source to a target task. Learning from clinical images from scratch is often not the most practical strategy due to its computational cost, convergence problem, and insufficient number of high-quality labeled samples. Pre-trained CNN alongside fine-tuning and transfer learning lead to faster convergence and outperform training from scratch.

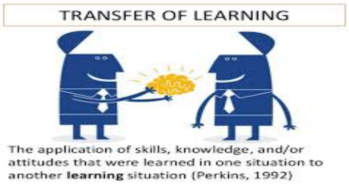

TRANSFER OF LEARNING
The application of skills, knowledge, and/or attitudes that were learned in one situation to another learning situation (Perkins, 1992)

Figure 3
Inception-v3 Model

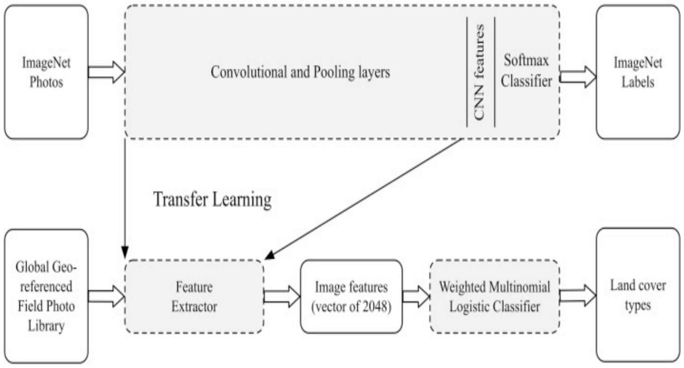

ImageNet Photos
Convolutional and Pooling layers
CNN features
Softmax Classifier
ImageNet Labels
Transfer Learning
Global Geo-referenced Field Photo Library
Feature Extractor
Image features (vector of 2048)
Weighted Multinomial Logistic Classifier
Land cover types

Figure 4

IV. Code with Documentation

https://github.com/palakagrawal912/agrawal_palak_spring18_ads/tree/master/project

V. Results

The results were divided into following parts,
a. Breast cancer classification using Convolutional Neural networks
b. Breast cancers classification using Transfer Learning

Breast cancer classification was defined using three classes, Benign, Benign without Callback and Malignant

The breast cancer data were categorized into malignant and benign groups. Using an 80% training set and 20% test set, the Convolutional Neural networks fine-tuned all layers, correctly classified malignant and benign cancer types with 72.17% confidence.
Using transfer Learning, correctly classified malignant and benign cancer types with 74.76% confidence.

VI. Discussion

In case of our CNN model, we used TensorFlow directly to construct the model and train them. We also created model checkpoints frequently which contained all the data related to the model at that instant and pre-trained the model. Henceforth, once the model is assessed the yield is printed which contains the precision of the model.
Before improvement, training accuracy reached to 100%. This was due to overfitting. Overfitting could be reduced by changing the probability and image augmentation. The test accuracy was 67.5%. So, we changed the dropout rate and induced randomization of the batch size.

In future, we can plan to implement more than 4 CNN layers to extract features for classification problem. We also plan on collecting more training samples from other sources of breast cancer mammograms, such that the model can be trained properly.

VII. References

[1] Wiki page of Cancer Imaging Archive collection for CBIS - DDSM

[2] Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks

[3] Whole mammogram image classification with convolutional neural networks

[4] Inception V3 using TensorFlow

[5] Convolution Neural Network – Wikipedia

[6] IEEE Library – Study on histopathology of cancer

[7] Transfer Learning by Sebastian Ruder

[8] Classification of breast cancer histology images using CNN

[9] Breast Cancer Statistics, 2017, Racial Disparity in Mortality by State

[10] Deep Learning Framework for Multi-class Breast Cancer Histology Image Classification