Machine Learning Engineer Nanodegree

Capstone Proposal

Tanush Goel

June 21st, 2019

Proposal

Domain Background

Invasive ductal carcinoma (IDC), also known as infiltrating ductal carcinoma, is cancer that began growing in a milk duct and has invaded the fibrous or fatty tissue of the breast outside of the duct. *Invasive* means that the cancer has “invaded” or spread to the surrounding breast tissues. *Ductal* means that the cancer began in the milk ducts, which are the “pipes” that carry milk from the milk-producing lobules to the nipple. *Carcinoma* refers to any cancer that begins in the skin or other tissues that cover internal organs, such as breast tissue. All together, “invasive ductal carcinoma” refers to cancer that has broken through the walls of the milk ducts and begun to invade the tissues of the breast. Over time, invasive ductal carcinoma can spread to the lymph nodes and possibly to other areas of the body. According to the American Cancer Society, about two-thirds of women are 55 or older when they are diagnosed with invasive breast cancer. Invasive ductal carcinoma also affects men.

Problem Statement

Invasive ductal carcinoma is the most common form of invasive breast cancer and represents 80 percent of breast cancer cases/diagnoses. As with any breast cancer, there may be no signs or symptoms. A mammogram may reveal a suspicious mass, which will lead to further testing. The average cost of a mammogram is $100 but may range from $75 to $250, which is a large sum of money for poorer families without healthcare or health insurance. More women are diagnosed with breast cancer than any other cancer, besides skin cancer. This year, an estimated 268,600 women in the United States will be diagnosed with invasive breast cancer and 2,670 men in the

United States will be diagnosed with breast cancer. It is estimated that 42,260 deaths (41,760 women and 500 men) from breast cancer will occur this year.

Datasets and Inputs

I would like to use the 1GB breast histopathology image dataset provided by Kaggle: https://www.kaggle.com/paultimothymooney/breast-histopathology-images

The original dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive).

These pictures could be fed into an algorithm which would be able to classify whether invasive ductal carcinoma is present in an image.

Solution Statement

The proposed solution to this problem is to apply a Convolutional Neural Network (CNN) as it has been proven to have great performance in image classification problems. The CNN would be trained as a binary classification problem on the thousands of images being either IDC positive or negative.

Benchmark Model

Most other models trained on this data have about 85% testing accuracy on about 15,000 images. The models are simple keras CNN's of 1-4 convolutional layers, some with some forms of data augmentation.

Evaluation Metrics

The model can be measured by simply the accuracy (total correctly predicted over total predicted) or f1-scores of both classes (average of precision and recall). It could also be measured by the sensitivity (true positives/(true positives + false negatives)) and specificity (true negatives/(true negatives + false positives)) of both classes, but the positive class would matter more, especially since it is the underrepresented class in terms of the number of images.

Project Design

1. Unzip the file of images
2. Separate the images into two subfolders as positive and negative classes
3. Make a dataframe of all the image id's, patient id's, and class
4. Split the dataframe into train, valid, and test images
5. Make train, valid, and test folders
6. Make augmented training images of underrepresented class until both classes are almost equal in number
7. Make train, test, and valid batches using datagen.flow_from_directory
8. Possibly update class weights to make positive class more sensitive
9. Build functional CNN
10. Train CNN on train batches using a checkpointer and valid batches as validation
11. Make predictions on test batches
12. Show accuracy of test predictions and use confusion matrix/classification report

References

Breast Cancer - Statistics
https://www.cancer.net/cancer-types/breast-cancer/statistics

Breast Histopathology Images - Paul Mooney
https://www.kaggle.com/paultimothymooney/breast-histopathology-images

Information and Resources About For Cancer: Breast, Colon, Lung, Prostate, Skin
https://www.cancer.org/

Invasive Ductal Carcinoma: Diagnosis, Treatment, and More
https://www.breastcancer.org/symptoms/types/idc

Invasive Ductal Carcinoma (idc) Breast Cancer: Johns Hopkins Breast Center - Ken Brown
https://www.hopkinsmedicine.org/breast_center/breast_cancers_other_conditions/invasive_ductal_carcinoma.html