Project Proposal

Regis University - MSDS696 - Data Science Practicum II

Student: Mary Crawley

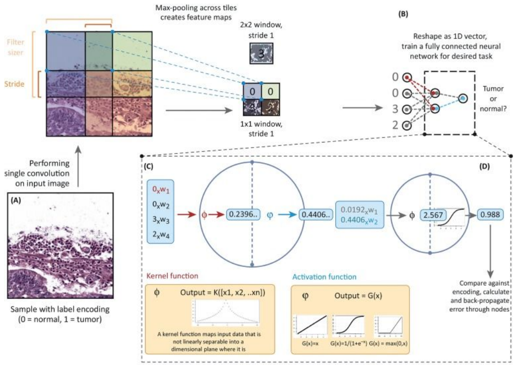
Summer 2022

Contents

1. Problem or Situation
2. Research Question
3. The Dataset
4. Methodology
5. Project Timeline

Problem or Situation

The most common breast cancer is called Invasive Ductal Carcinoma (IDC). Assigning an aggressiveness grade to a whole mount sample, pathologists normally focus on the areas which contain the IDC. This results in a common pre-processing step that aim for an automatic aggressiveness grading by delineate the exact regions of IDC inside of a whole mount slide.

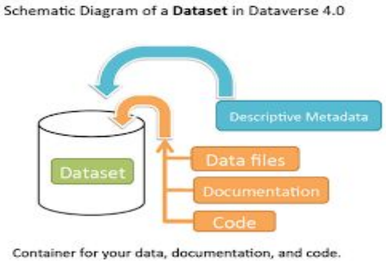

Max-pooling across tiles creates feature maps
Filter sizer
Stride
2x2 window, stride 1
3
0
0
1x1 window, stride 1
(B)
Reshape as 1D vector, train a fully connected neural network for desired task
0
0
3
2
Tumor or normal?
Performing single convolution on input image
(A)
Sample with label encoding (0 = normal, 1 = tumor)
(C)
(D)
$0_x w_1$
$0_x w_2$
$3_x w_3$
$2_x w_4$
ϕ
0.2396..
φ
0.4406..
$0.0192_x w_1$
$0.4406_x w_2$
ϕ
2.567
0.988
Kernel function
ϕ Output = K([x1, x2, ..xn])
A kernel function maps input data that is not linearly separable into a dimensional plane where it is
Activation function
φ Output = G(x)
G(x)=x
$G(x)=1/(1+e^{-x})$
G(x) = max(0,x)
Compare against encoding, calculate and back-propagate error through nodes

Trends in Cancer

Research Question

Can a Deep Learning classifier model help increase the accuracy and reduce time in determining whether a histology image is benign or malignant for potential Breast Cancer patients?

The Dataset


Schematic Diagram of a Dataset in Dataverse 4.0
Dataset
Descriptive Metadata
Data files
Documentation
Code
Container for your data, documentation, and code.

Data used: Kaggle-Breast Cancer Prediction Dataset

We will be using the IDC_regular dataset (breast cancer histology image dataset) from Kaggle. There is approximately 277,524 patches that are the size of 50x50 extracted from 162 whole mount slide images of breast cancer specimens that are scanned at 40x. The data holds 198,738 test negative and 78,786 test positive with IDC. The dataset is available for the public and can download it here. The size of the data is a minimum of 3.02GB of disk space for this project.

Filenames in this dataset look like this:

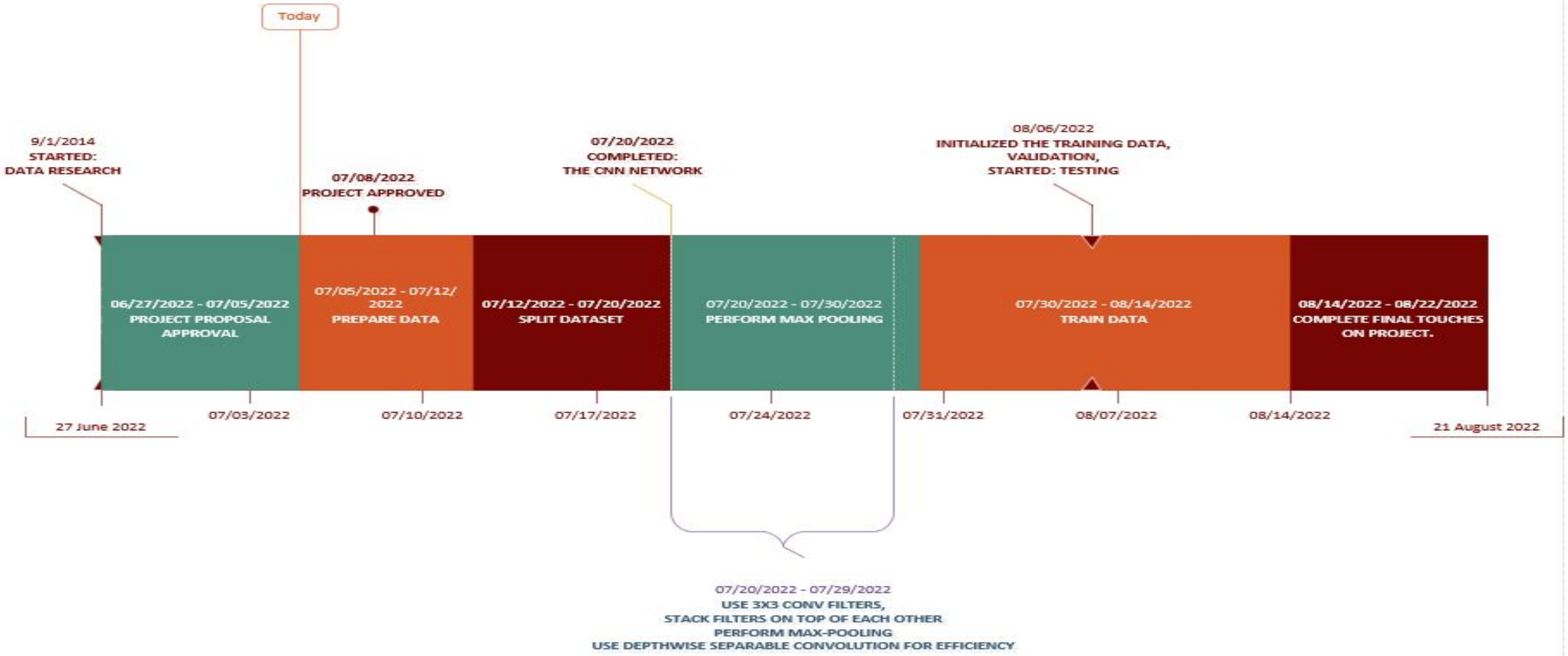
8863_idx5_x451_y1451_classo

Here, 8863_idx5 is the patient ID, 451 and 1451 are the x- and y- coordinates of the crop, and 0 is the class label (0 denotes absence of IDC).

Methodology

This project will include the use of Keras that will define a CNN (Convolutional Neural Network) to train it on our images. There will be a directory for each patient ID that holds 0 and 1 directories for images. This will be used to split our dataset into training, validation, and testing sets in an 80% for training with 10% of that reserved for validation, and 20% for testing. Using tuples (Python list that can hold a sequence of items) for information about the training, validation, and testing sets.

Project Timeline



Today

9/1/2014
STARTED:
DATA RESEARCH

07/08/2022
PROJECT APPROVED

07/20/2022
COMPLETED:
THE CNN NETWORK

08/06/2022
INITIALIZED THE TRAINING DATA,
VALIDATION,
STARTED: TESTING

06/27/2022 - 07/05/2022
PROJECT PROPOSAL
APPROVAL

07/05/2022 - 07/12/
2022
PREPARE DATA

07/12/2022 - 07/20/2022
SPLIT DATASET

07/20/2022 - 07/30/2022
PERFORM MAX POOLING

07/30/2022 - 08/14/2022
TRAIN DATA

08/14/2022 - 08/22/2022
COMPLETE FINAL TOUCHES
ON PROJECT.

27 June 2022

07/03/2022

07/10/2022

07/17/2022

07/24/2022

07/31/2022

08/07/2022

08/14/2022

21 August 2022

07/20/2022 - 07/29/2022
USE 3X3 CONV FILTERS,
STACK FILTERS ON TOP OF EACH OTHER
PERFORM MAX-POOLING
USE DEPTHWISE SEPARABLE CONVOLUTION FOR EFFICIENCY