**Capstone 2: Final Report**

**Problem Statement**

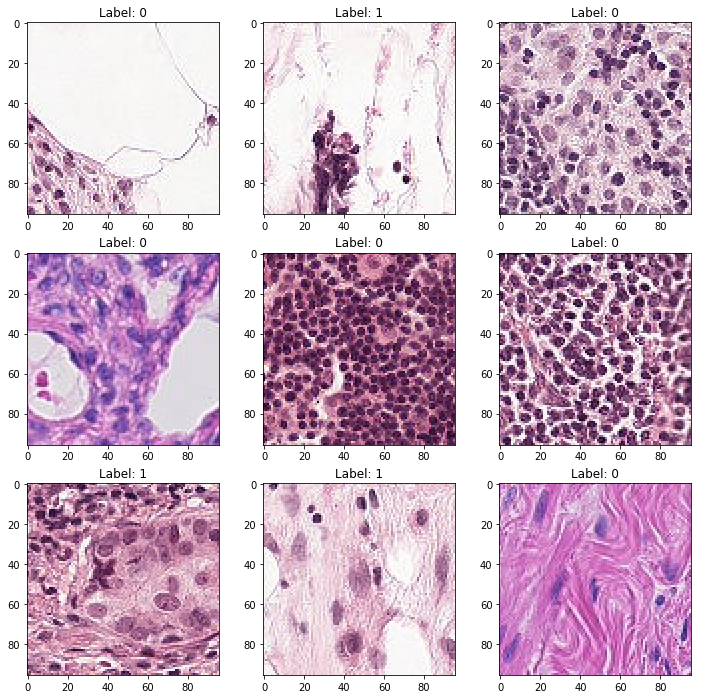
With the increases in cases of cancer, early detection and treatment is of vital importance as it relates to reduction in mortality rate among certain groups in the population. This project focuses on identifying metastatic tissue in histopathologic scans of lymph node sections.

**Dataset Overview**

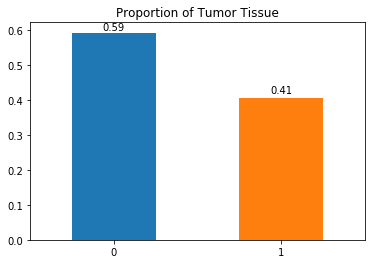
Data was obtained from the Kaggle competition, Histopathologic Cancer Detection. The dataset can be obtained here, www.kaggle.com/c/histopathologic-cancer-detection

The folder contains 277,483 images (7.41 GB) of small pathology images of which 220,025 are for training and 57,458 for testing. Each image has a resolution of 96 x 96 px, however only the center 32x32 px region is used for this binary classification as having at least one pixel of tumor tissue.

Each image varies in cell density, shape, size, and staining.



Along with the images the folder contains a csv file that contains the labels for the training data. Simple visualization reveals that 90,210 (41%) of the training images contain tumor cells in the center 32x32 px region.



**Methodology**

This project presents a large dataset of images which will results in a long training time. Great effort must be expended in pre-processing and system design to feed the images to the network.

The aim of the project is to classify images into two classes, therefore this task requires the use of Deep Convolutional Neural Networks.

To effectively train the model we will use a large number of epochs using GPU to increase the processing speed. Different hyperparameter values will be applied to the model along with the addition of multiple layers. The train set images will be divided into a training set (75%) and a validation set (25%).

The images are all stored in one folder with a csv file containing the names and labels. The images will be feed to the model using the flow\_from\_dataframe() function.

**Model Selection**

The number of nodes were selected to prevent two unwanted situations; underfitting and overfitting. High training errors (underfitting) can be prevented by not having a simple architecture with too few nodes for this project that has a complex predictive factor. On the other hand, having too many nodes can result in overfitting, model performing well on training data but lacks generalization.

Image classification problems aren’t linearly separable and therefore a multi-layered architecture with respectable number of nodes is required. Experimentation and intuition are two ways to determine the number of layers and nodes to use. This is the first image classification I have performed; therefore, a lack of experience would not allow me to use intuition. Experimentation will will translate into intuition as more classification problems are undertaken. It was noted that training and validation accuracy didn’t translate into a high accuracy (82%) in test images. After obtaining the accuracy of a few modified architectures I have selected the model as is presented below. This model satisfies both depth and width.

**Model**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

conv2d\_23 (Conv2D) (None, 94, 94, 64) 1792

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conv2d\_24 (Conv2D) (None, 92, 92, 64) 36928

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conv2d\_25 (Conv2D) (None, 90, 90, 64) 36928

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

max\_pooling2d\_9 (MaxPooling2 (None, 45, 45, 64) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conv2d\_26 (Conv2D) (None, 43, 43, 128) 73856

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conv2d\_27 (Conv2D) (None, 41, 41, 128) 147584

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conv2d\_28 (Conv2D) (None, 39, 39, 128) 147584

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

max\_pooling2d\_10 (MaxPooling (None, 19, 19, 128) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

flatten\_5 (Flatten) (None, 46208) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_5 (Dense) (None, 1024) 47318016

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dropout\_7 (Dropout) (None, 1024) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

pred (Dense) (None, 1) 1025

=================================================================

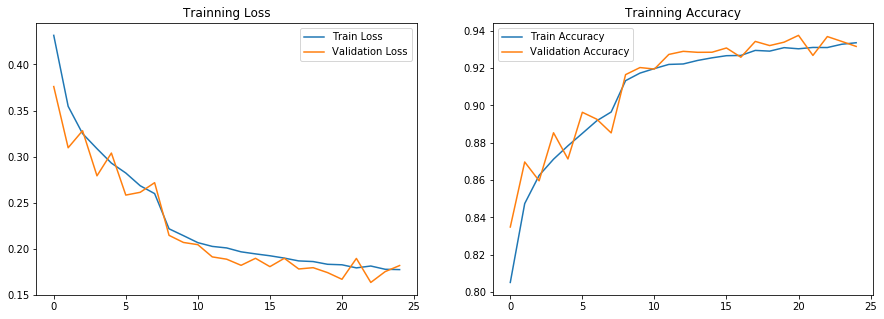
Total params: 47,763,713

Trainable params: 47,763,713

Non-trainable params: 0

**Results**

In the image augmentation step multiple image transformations were applied to the training set images to effectively identify the target feature.



**Discussion**

At the beginning of the process CPU was used to train and evaluate the images and each epoch took approximately 5 hrs to complete. After tapping into the gaming GPU on the computer the time was reduced to ~30 mins per epoch. Callbacks where effect in increasing the accuracy and decreasing the loss, if there was no improvement in two successive epoch the learning rate was adjusted, and the subsequent epoch showed improvement. The test images were classified and given an accuracy of 82% on submission. It would be interesting to see how additional image manipulations affect accuracy.