ADL Final Project

Wei Han (wh2365), Chengtian Xu (cx2168), Yang Gao (yg2499)

Relevant Links

Github
Youtube: Presentation Code Walkthrough

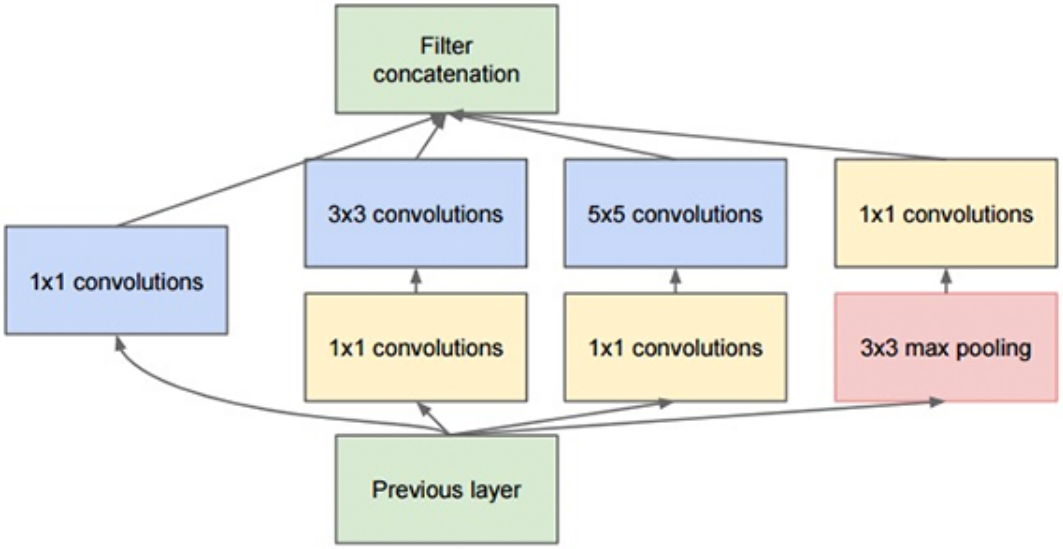
Model Architecture

Inception v3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. Also, the VGG16 model achieves 92.7% top-5 test accuracy in ImageNet competition. We considered that Inception v3 and VGG16 are both good practice to start with pre-trained model.

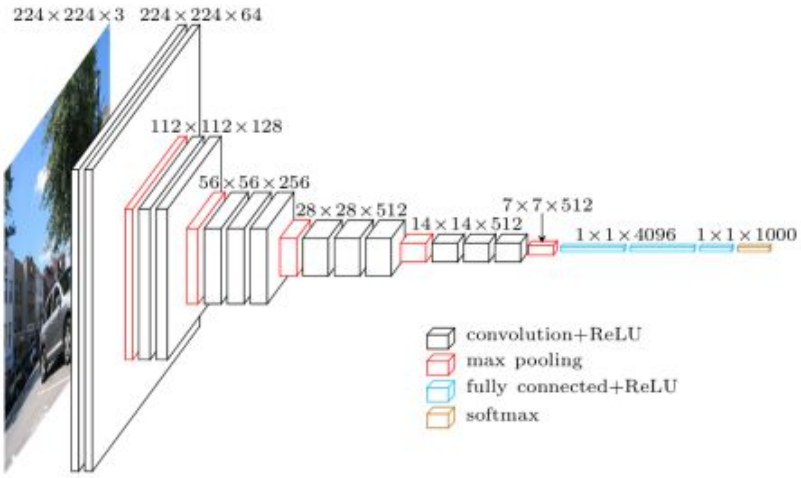
Since Inception V3 and VGG16 model are both highly efficient in solving large scale image classification problem, we decided to use these two models to make a prediction in tumor images (normal or has tumor).

Pre-trained base model

- InceptionV3
    - Inception Module: multi-level feature extractors by computing 1 x 1, 3 x 3, 5 x 5 convolutional layers within the same module.
    - Stack the channel dimensions on 1 x 1, 3 x 3, and 5 x 5 convolutions filters as the output of the module.
    - Each output module result saved as the next module input layer.
    - Advantages: modest computational cost, considering the large amount of training data



- VGG16
    - Stack of convolutional layers: Each batch of convolution layers contains 2 convolutional layers and 1 Max-Pooling layer.
    - Fully-Connected(FC) layers: 3 Fully-Connected layers follow a stack of convolutional layer.

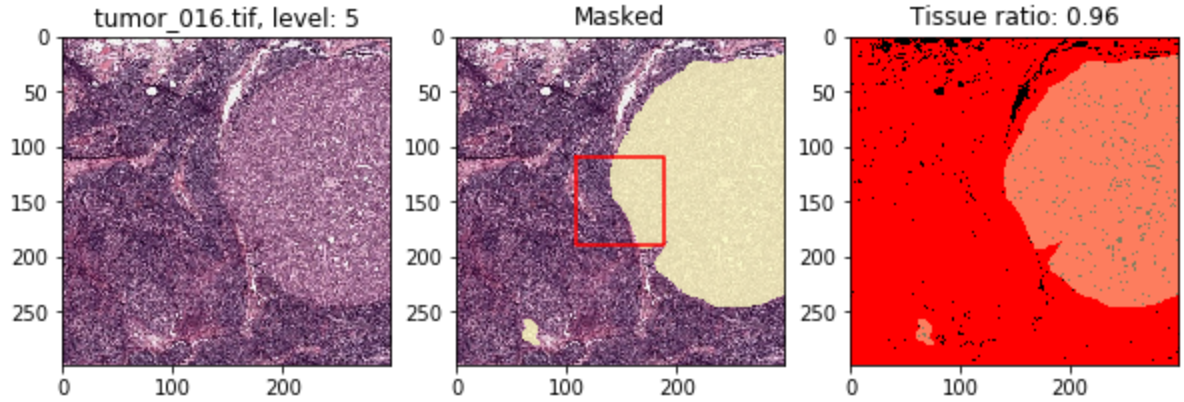


VGG16 deeply increase the depth of the layer and decrease the wright/height of the layer.

Project Customized Model :

- InceptionV3 + Dense Layer (ReLU) + Dense Layer (Sigmoid for 2 Classification)
- VGG16 + Dense Layer (ReLU) + Dense Layer (Sigmoid for 2 Classification)

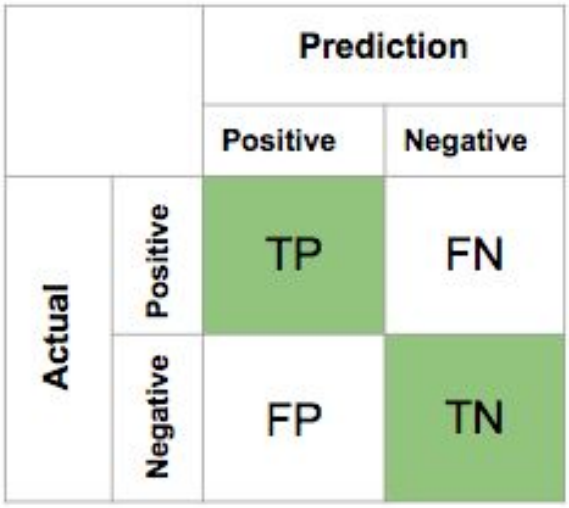
Data Preprocessing



- 300 sample patches per training slide
- 120 positive samples + 170 negative samples
    - Since the tumorous region is small and concentrated on most of the slides, more negative samples are extracted in comparison to the number of positive samples.
    - Positive samples are defined as the patches that contains tumor at the center 128 * 128 region.
- The patch size is default to 299 * 299
    - Experiments with a smaller patch size of the training patches 100 * 100 have shown that the resultant models are inclined to “overpredict” the testing patches, i.e. leading to low precision. Since the positive regions oftentimes tend to be darker (higher greyscale value), the resultant models tend to predict every single darker patches as positive. 100 * 100 may have been too small to capture the context of the slide for prediction.
    - Experiments with a larger patch size of the training patches 400 * 400 have shown that the resultant models are inclined to “underpredict” the testing patches, i.e. leading to low recall.
- Tissue Region > 50%
    - Only patches that have > 50% of pixels as tissues would be extracted as samples.

Evaluation of Model Performance (Tested on 110)

- Accuracy is irrelevant since the task of tumor tissue classification is class imbalanced
- Based on the following Confusion Matrix:



    - Recall = tp / (tp + fn)
        - How well the model is able to recognize the actually positive patches
    - Precision = tp / (tp + fp)
        - How many are actually positive patches among positive predictions
    - F1 score = 2 (precision * recall) / (precision + recall)
        - Harmonic average of recall and precision, when reaches 1 it means both recall and prediction are optimal
- Metric Measurements
    - InceptionV3 at level 5:
        - Recall: 0.78
        - Precision: 0.99
        - F-1 score: 0.87
    - InceptionV3 at level 6:
        - Recall: 0.88
        - Precision: 1.00
        - F-1 score: 0.93
    - VGG16 at level 5:
        - Recall: 0.88
        - Precision: 0.88
        - F-1 score: 0.88
    - VGG16 at level 6:
        - Recall: 0.79
        - Precision: 0.90
        - F-1 score: 0.84
- Performance Comparison Result:
    - Both VGG16 and InceptionV3 perform better when prediction are made on level 6 than level 5. However they have all reached a F-1 score greater than 0.80 on both levels, indicating accurate predictions are made.
- Further Thoughts on Evaluation Metrics

- Due to the goal of our project, which aims at ***facilitating*** rather than replacing human pathologists' diagnosis, it is important for the model to obtain the capability of recognizing "potentially cancerous regions," to assist the human pathologists. Therefore recall which measures how much the model recognizes true cancerous cells is a much more important metric.