Project Description

Metric Design and evaluation

The metrics we chose to focus on were accuracy, ROC, recall. In general, we have preferred a better recall over the precision.

- ROC - Higher value signifies how well our model is able to discriminate between the two classes. This is important as we want our model to identify what a tumor looks like and what a healthy cell looks like separately and properly

- Precision, recall and F1- Since we do run experiments with the skewed datasets too, these are good measure. We have given preference to a higher recall over precision. The main motivation for this is the motto - 'Better safe than sorry'. We feel it is more important for us to predict lesser false negatives than it is to predict lesser false positives. This is related to the potentially serious implications of cancerous tissues and the risk of a misdiagnosis for an actually affected person

- Accuracy - While not super significant while testing due to skewness of the dataset, we have made the dataset balanced while feeding it to the multi scale architecture and accuracy during training/validation phase gives us a good idea of how our model is learning.

Experiments conducted

1. Difference between transfer learning and Convolutional model approach: We fed same inputs to both these models both in multi scale and single scale architecture

2. Difference in performance with a skewed vs balanced dataset: Since tumors were relatively fewer, We used data augmentation/randomly dropped samples to bridge the discrepancy between healthy and tumor images in the dataset. To augment the dataset we did 3 transformations - flip left, flip right and flip both

3. Difference between single scale architecture and multi scale architecture: We fed levels 4 and 5 separately to the single architecture. We compared this with performance when we feed both level 4 and 5 to the multi scale architecture

4. Difference between two levels : 4 and 5 in single scale architecture

5. Difference between an averaged approach: running two models and merging result in the end vs making a single model with two inputs and single outpu

Analysis and Results

- In general, for single scale, accuracy, roc, precision and recall is better for zoomed in images
- We found that the multi scale model with levels 4 and 5 performs better than single scale model with one level
- The performance of convolution model is comparable or slightly better than the transfer learning one. In short, a huge network did not significantly increase performance
- Also, we have balanced the dataset by different techniques such as augmentation/truncating excess samples. We conducted experiments on level 5 with a single scale model. It was observed that with a more balanced dataset, the ROC, precision, Recall and accuracy improved. Also, visibly the heat Map became better
- Our created models generally have good accuracy, ROC score and decent recall score. We have fewer False negatives - this helps in cases of critical diseases such as Cancer, where it is far more serious to say that a person does not have cancer when he actually has it versus vice versa
- Since we have trained the multi-scale on an augmented, balanced dataset for levels 4,5, it has developed a good model to identify potential tumor areas. However, as we have observed, the model generally shows more cancerous than actually present as the test set is far more skewed (few tumors, mostly healthy) than the training data the model is trained on
- Since recall and precision have a tradeoff, our model does show poor precision scores and F1 scores in cases where there is no tumor
- We have observed that the predicted masks for our model do identify most of the cancerous regions in the original mask.

Utility

- It could be deployed to provide a second opinion to the doctor as to where the cancerous cells are present.

- Especially since the model predicts more false positives, it helps reduce missing out on cells which might be cancerous hence decreasing the ultimate negative impact although it requires a little more work on the clinicians end.

Future work:

- This project can be extended to employ attention mechanisms along with the current implementation of context in deep learning models which perhaps would help reduce the false positives.
- Secondly, it would be interesting to test the model on other similar large scale cancer datasets.