Equivariant Convolution for Histopathology Segmentation

Josef Liem

1/31/2025

Background & Problem

Annotation and grading of whole slide images has proven to be a very time consuming and costly task, given the amount of investment it takes to train a pathologist. Nevertheless, this task is a critical component of the diagnosis of certain diseases such as cancers. For this reason, there is substantial interest in developing deep learning systems capable of understanding tissue and cellular morphologies in a segmentation setting, to assist and improve diagnostic accuracy. When designing such a system, several key constraints will influence the selection of appropriate architectures.

Firstly, a robust machine learning approach to this problem should make classification or segmentation decisions using as much context about the sample as possible; when grading cells, a trained pathologist should consider the context of the surrounding cells or tissue at varying magnifications.

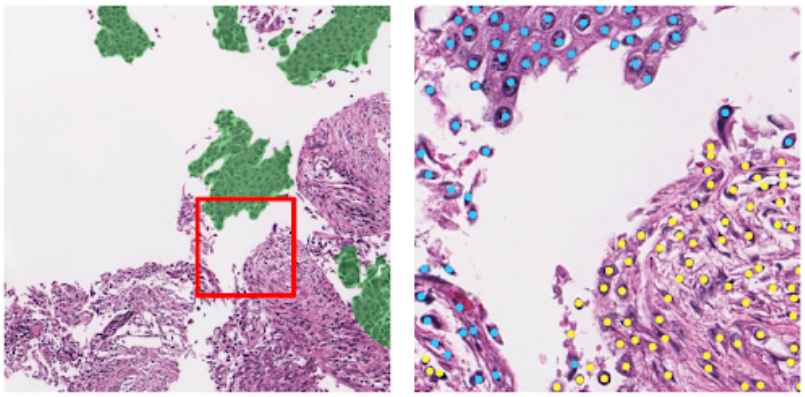

Figure 1: A tissue (left) and cell (right) pair from the Overlapped Cell On Tissue Dataset for Histopathology challenge. Tissue level segmentation should ideally factor into the classification decisions at the cellular level.

Secondly, due to the sensitivity of medical data, the time and cost required for annotating WSIs, and the diverse cellular morphologies of certain diseases, the available training sample size is often relatively small. Researchers will need to maximize how effectively they can use their data either through substantial data augmentation, or other advanced techniques such as self-supervised, transfer, or federated learning. Given that many types of medical images can present with symmetries, it would be nice to be more efficient on data during training by baking in these symmetries into the model; the property of equivariance.

Formally, *equivariance* is defined as $f(T(x)) = T(f(x))$ for a given x. That is, the transformation on the input of a function behaves in a consistent manner with the transformation on the output of that function. Moreover, invariance is the special case of equivariance with just the identity: $f(T(x)) = f(x)$. That does not mean, however, that if something is considered "equivariant", we can simply use any transformation on the input without careful consideration. We usually describe a function as equivariant under a specific set of *group actions* (translations, reflections, rotations...), and the dimensionality of the *space* on which we are applying these transformations. We can achieve equivariance through a number of *representations*, which describe

how we can apply a set of group actions in a way that respects the property of equivariance. A few of the natural consequences of equivariance include:

1. Models are more resilient to transformations on the input data, allowing for greater generalizability.
2. When our model is being trained, we do not need to augment data to have the model understand other transformations of that data.
3. Often a reduction in the number of parameters as is evident in convolution and translational equivariance versus fully connected networks.
4. Lower scaling in dataset size requirements when moving to $\mathbb{R}^3$ and beyond, with more group actions.

While convolutional neural networks are translationally equivariant, they are not equivariant under other group actions. Given that cells are rich in rotational symmetries, it would be nice to extend the property of equivariance in convolution to rotations.

Thus, the major long term goal of this project was to utilize rotational symmetries and equivariance on cell-level cancer imaging data, while factoring in large FOV tissue level context, to improve our ability to detect and segment malicious cells given limited training data.

Methods

Given the time constraints, we focused for the moment on cell level segmentation with rotational equivariance using U-Net - an autoencoder-like architecture. At every layer of convolution, we used a particular kind of rotationally equivariant kernel $K(x) = f(|x|)$ - the trivial representation involving scalar to scalar mappings. In this context, our model was equivariant under the cyclic group $C_4 = \{R_0, R_1, R_2, R_3\}$ (multiples of 90 degree rotations).

In the decoder portion of the model, we swapped out the transposed convolutions for nearest-neighbor upsampling, as transposed convolution still

remained somewhat ill-defined under this framework. Additionally, we used single convolution between each down-sampling as opposed to double convolution in more mainstream variants of U-Net.

We then compared the performance against a non-equivariant equivalent of our model that did not use our special kernel but instead with a substantial amount of data augmentations such as flips, rotations, elastic transformations, and slight changes to noise and color. We used 110 scalar channels in the first layer moving from the 3 channel RGB input in the equivariant model, and 64 channels in the non-equivariant model to allow for approximately the same number of trainable parameters for fair comparison.

We trained both models on the Adam optimizer and with Dice-Cross-Entropy loss, which is a weighted combination of Dice and Cross Entropy loss. Dice loss is defined as Dice Loss $= 1 - \frac{2\sum_i P_i G_i}{\sum_i P_i + \sum_i G_i}$ where P_i and G_i are binary 0/1 values for given predicted and ground truth pixels. We used dice as it was nice for ensuring imbalance in classes was handled properly.

For our dataset, we used the Overlapped Cell On Tissue Dataset for Histopathology (OCELOT), which consisted of pairs of tissue and cell images sourced from WSIs.

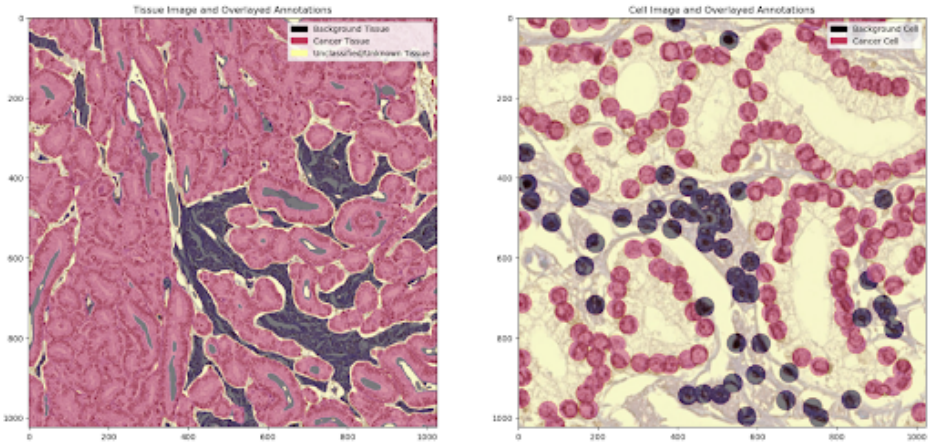

Figure 2: Tissue mask overlayed on image (left) and cells with mask (right)

There were 400 pairs of 1024 x 1024 images to train on, 137 to validate

on, and 130 to test on. These images were tiles extracted from larger WSIs from patients diagnosed with bladder, endometrium, head-and-neck, kidney, prostate, and stomach cancer.

The actual samples themselves were formatted with several important pieces of information: the tissue image, the tissue mask, the cell image, the csv file containing coordinates and classes of cells, and metadata describing the relative positions of cell patches in tissue images. To convert the cell annotations into a format compatible with segmentation, for each cell coordinate, we assigned a circle with a radius of 30 pixels and produced a corresponding mask.

We trained on Middlebury College's Ada cluster for a total of six hours, saving the best performing equivariant (without augmentation) and non-equivariant (with augmentation) models on validation loss over the course of 200 and 1000 epochs at a learning rate of 0.0001. During training, we tracked losses and other metrics in real-time through the Comet-ml API. Given more time we would perform a search for hyper-parameter tuning.

Results

Upon training both models, we noticed that the loss for the rotationally equivariant model initially started out extremely high, but after ten epochs generally settled to about the same value as the non-equivariant model.

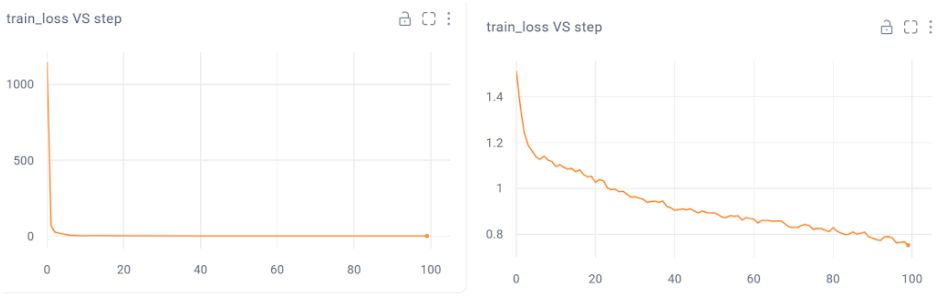

Figure 3: Equivariant training loss (left) and non-equivariant training loss (right). The equivariant model initially starts out extremely high but settles to about the same value as the non-equivariant model.

As for validation loss, the same trend was observed between equivariant and non-equivariant model losses. It is important to note that the training and validation losses should not be compared, as the scaling of the losses during calculation was not identical as a result of the data-loader.

Figure 4: Equivariant validation loss (left) and non-equivariant loss (right).

Inference on testing data showed that our equivariant model achieved a loss of 1.16, while the non-equivariant model was 2.29. However, when calculating the raw dice coefficient alone, the equivariant model coefficient was 0.62 while the non-equivariant version was 0.68. Both of these coefficients are relatively moderate, suggesting noticeable errors were present, especially in some fine detail. This discrepancy between the loss and the Dice coefficient suggests that the equivariant model may still be struggling with class imbalance, potentially due to overconfident background predictions dominating the outputs.

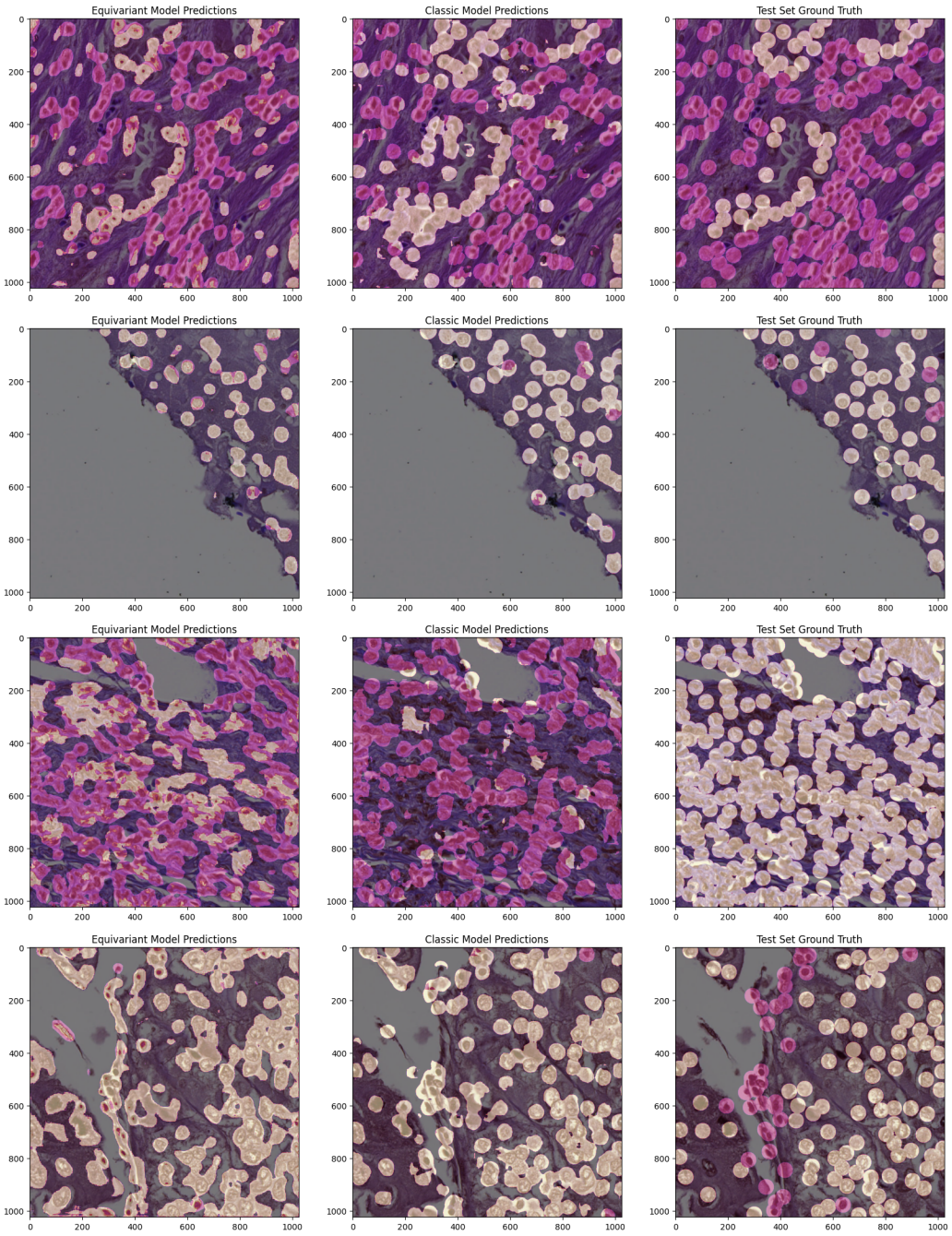

Figure 5: Equivariant segmentations (left column), non-equivariant segmentations (middle column), and ground truth (right column) with examples of good segmentations (top rows), and bad segmentations (bottom rows)

In Figure 5, the top two columns show that both models were capable of adequately segmenting instances of cells and making some differentiation between cancerous and background types in easy scenarios. However, on some samples, the non-equivariant model had a tendency of over-predicting cancer whereas the equivariant model was uncertain, blending both classes for a given cell instance.

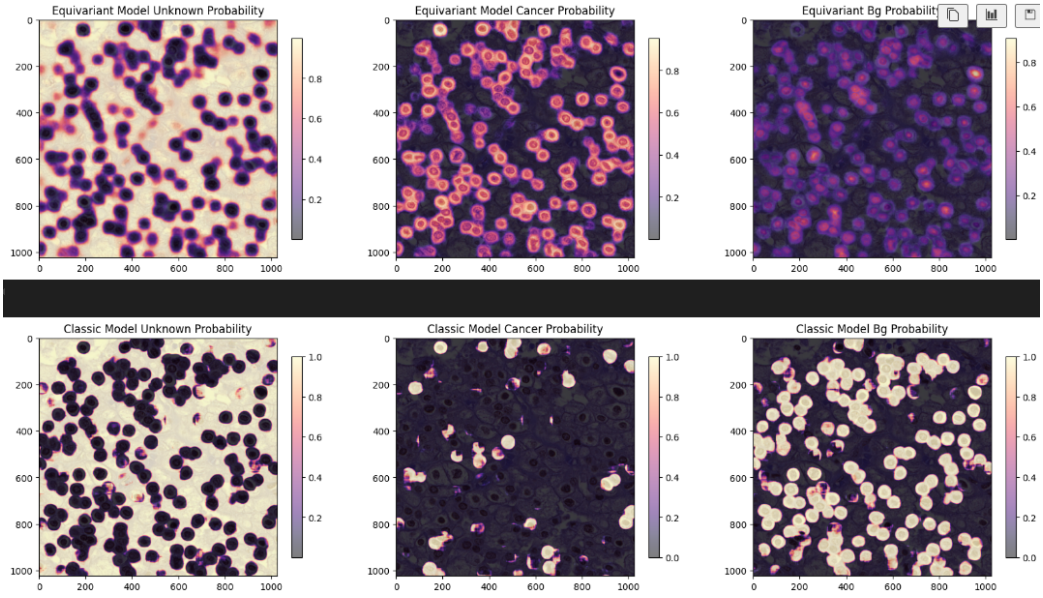

Figure 6: Probability maps of predictions on equivariant (top) and non-equivariant (bottom) models

In Figure 6, both equivariant and non-equivariant models were able to capture the distinction between cell instances and background. However, unlike the non-equivariant model, the equivariant model struggled to make confident pixel-wise classifications of either cancer or background cell. This is perhaps why on bad predictions, the non-equivariant model had a tendency of predicting either one cell type or another for a given cell, whereas the equivariant model is more uncertain, blending cell types per segmentation.

Conclusion

Through the course of this project, we were able to build an equivariant segmentation model on histopathology WSI data, and compare it against a non-equivariant model with data augmentation. In designing and modifying this architecture, it has become apparent that the design of many types of deep learning models can be somewhat arbitrary. There are many alternative choices of architectures out there, DeepLab and Mask-RCNN among them. We also built a greater appreciation for the useful properties of convolution, as well as their limitations in particular domains and tasks.

Moving forward, we would like to work on the second part of this project, which includes factoring in high level tissue context into the final cell segmentation model. We propose that using Mask-RCNN at the cellular level would allow us to take full advantage of rotational equivariance, as this architecture would ideally first perform a region proposal on a given cell and then apply segmentation. By applying segmentation on a smaller cell ROI, we would be able to fully leverage the rotational symmetries of the proposed region, rather than an entire patch of cells. Prior to that, we would also use a tissue segmentation model, and feed the tissue output together with cell images into Mask-RCNN (instead of DeepLab like in the SoftCTM approach).

We would also like to expand on the metrics we report such as mean Intersection Over Union (mIOU). We would also like to produce an F1 score by treating each cell segmentation as an instance of classification to benchmark against other OCELOT approaches.