Identifying Cell Nuclei in Divergent Images

1 Background

Pathologists use immunohistochemistry (IHC) to detect tumours by identifying and quantifying the presence of important biomarkers expressed in cell nuclei. However, manual identification is time-consuming, thus it is highly desirable to develop an automated, high-accuracy method for isolating and analyzing nuclei in different kinds of IHC images.

2 Data Description & Challenge

The dataset is challenging because of high volume and dimensionality. Our data is divided into a training set (670 images, each containing between 1 to 100 masks for distinct nuclei) and test set (65 images). The images vary in size (total pixels) and were collected from many different cell types under a variety of imaging conditions (magnification, modality, etc). To achieve success, we will have to work with all the given data to develop a robust method for cell nucleus identification.

3 Hypotheses & Goals

Goal 0. One-index the pixels in each image from top to bottom, left to right.

Goal 1. Normalize across set of images.
The variety of cell type, staining and imaging condition all complicate cell identification. Pre-processing the data will ensure comparison across uniform images.

Goal 2. Identify all objects in each image.
We will separate objects from background, categorizing each pixel as ground or non-ground.

Goal 3. Separate individual cell nuclei. Once we have distinguished all objects as distinct from ground, the next step is to determine individual cells. For each image, we will return a set of masks, each mask only covering one nucleus with no overlap between any of the masks.

Goal 4. Maximize average precision.
Precision is the number of true positives divided by the sum of the number of true positives, false positives and false negatives. This is defined as a function of an intersection over union (IoU) threshold t between a set of predicted pixels A and a set of true object pixels B. For each image, we want to maximize average precision for $t \in [0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95]$.

4 Definition of Success

Success is a defined as a workflow that consists of pre-processing to normalize variation in imaging condition, separating objects from background in each image, and distinguishing individual cell nuclei. The differences in low, expected and high success are based on model accuracy as follows.

Low: Accomplish Goals 0-2. For Goal 3, predict nuclei regardless of accuracy.

Expected: Accomplish Goals 0-3. For Goal 4, achieve

- >65% accuracy for threshold values 0.5, 0.55, 0.6, 0.65
- >75% accuracy for threshold values 0.7, 0.75, 0.8, 0.85
- >85% accuracy for threshold values 0.9, 0.95

High: Accomplish Goals 0-3. For Goal 4, achieve

- >80% accuracy for threshold values 0.5, 0.55, 0.6, 0.65, 0.7
- >90% accuracy for threshold values 0.75, 0.8, 0.85, 0.9, 0.95

5 Deliverables

The key deliverable will be a Jupyter notebook containing:

- Code to convert images to one-indexed pixels and normalize the data from different imaging conditions
- Code to separate objects from the background
- Code to identify individual nuclei
- Documentation of the inputs and outputs of all functions
- Quantitative assessment of model accuracy
- Written critical analysis of successes/failures of the model

6 Schematic