Hodgkin's Lymphoma
Cell Classification Report
draft 2

Lihan Yao

Introduction

We analyze the dataset of an oncology study centered around Hodgkin's Lymphoma. The dataset format consists of high-resolution microscopy images of tumor tissues under various states of reactivity to the same medical treatment. These tumor tissues and images were procured by the Ingo Mellinghoff Lab at Memorial Sloan Kettering Cancer Center as part of on-going oncology research. A natural machine learning task may be formulated as: given a multi-channel cell image where individual channels correspond to experimentally introduced biological markers, classify the cell between tumor/T-cell types, and if possible, the specific T-cell type.

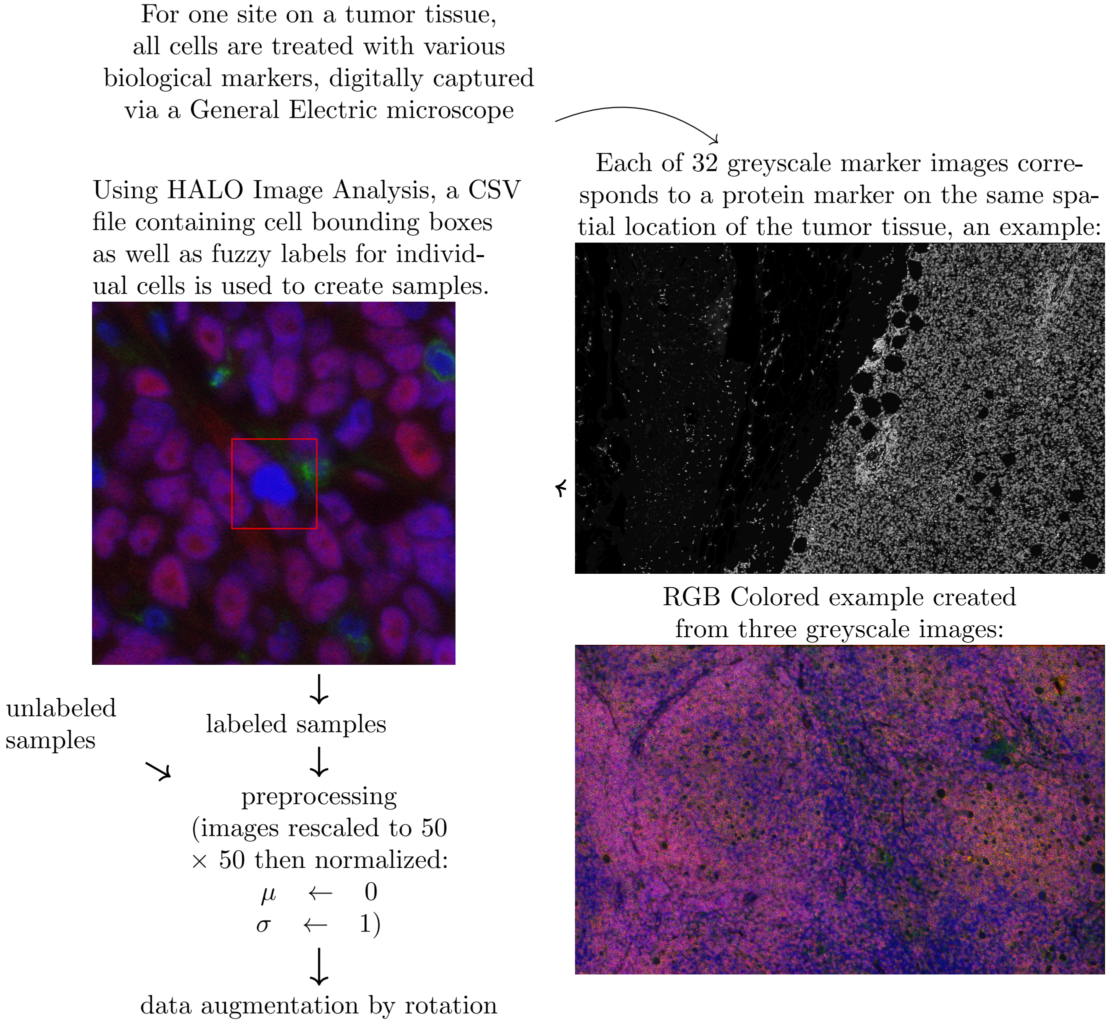
A central difficulty to this task is the lack of a labeled dataset from which supervised learning traditionally proceeds. With help from MSK oncology researchers, we aim to generate a labeled dataset of 80 samples. We maximize the usage of this human expertise by 1. prioritizing labeling of samples that optimally improved an initial model and 2. applying data augmentation and model architectures specialized in utilizing unlabeled data.

Data

Due to the exotic nature of tumor cell image data, a central obstacle to training and evaluation has been access to a reliable labeled dataset. In addition, the acquisition of good data has traditionally been the most resource exhaustive component of data-centered projects. Special care has been taken to maximize the effort of human experts.

Prior to the creation of a labeled dataset, we utilize HALO Image Analysis software for both cell detection (in the form of bounding boxes) and fuzzy labeling. By carefully viewing the tissue site through various biological markers, for a marker m, the researcher selected an appropriate threshold θ_m. If a cell's average pixel intensity is above θ_m, it is 'labeled' positive for m. This is repeated for 32 biological markers. In this sense, a cell's class can be represented as a 32 dimensional zero-one vector, and fuzzy, due to this threshold model's inherent variance. To support the on-going oncological research, we concern ourselves with states of four markers in particular. Moreover, in this subset of $2^4 = 16$ class labels, a majority are not biologically possible. When the threshold model generates such labels, we treat the corresponding sample as unlabeled. For these two reasons, a cell may belong to one of five potential classes. See appendix for an explanation of interpreting class labels.

Figure 1: **Pipeline for cell sample generation from raw images of the study**



For one site on a tumor tissue, all cells are treated with various biological markers, digitally captured via a General Electric microscope

Each of 32 greyscale marker images corresponds to a protein marker on the same spatial location of the tumor tissue, an example:

RGB Colored example created from three greyscale images:

Using HALO Image Analysis, a CSV file containing cell bounding boxes as well as fuzzy labels for individual cells is used to create samples.

labeled samples

unlabeled samples

preprocessing (images rescaled to 50 × 50 then normalized: $\mu \leftarrow 0$ $\sigma \leftarrow 1$)

data augmentation by rotation

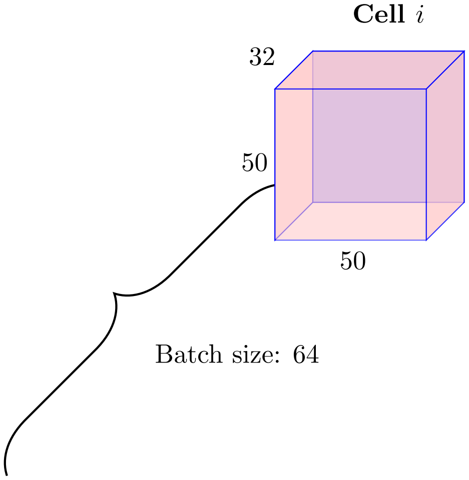
First the ladder network is trained on a large datasets with fuzzy labels generated by the threshold model. This is to distinguish a few hundred difficult samples to present to oncologists. We may then use these labeled samples to evaluate both models and begin a human-labeled dataset. The training set at this stage has 104,500 samples, with test set containing 41,000 and unlabeled set containing 260,800 samples, all after augmentation.

Model

Model selection is determined by three observations:

1. image samples have high number of channels and high resolution

Figure 2: Dimensions of a cell's resultant representation, in a batch



2. human labeled dataset is a few hundred in size while unlabeled dataset is orders of magnitude larger.

3. augmenting cell image data by rotation leads to new data which, to the human eye, is indistinguishable from original data

To take advantage of the last point, we quadruple labeled data by applying right angle rotations to individual samples. This practice was noted in Gao et al. (in the task of HEp-2 Cell Image Classification) to improve the model's rotational invariance. The second point hints at intricate nonlinear relationships which may arise from cell features. This suggests the complexity of a deep net is appropriate for the circumstance. In the area of image classification, ConvNets have been standard.

The first point, mainly that an ideal model should take advantage of the semi-supervised format, has put forth the Ladder Network as most fitting for our task, since it operates an unsupervised denoising task for representation learning, in addition to the original supervised task. For an architecture diagram and short summary of its operations, see first section of appendix.

Parameters of the encoder component (the feedforward component) is detailed below. Each convolutional layer introduces gaussian noise drawn from $\mathcal{N}(\mu, 0.1)$ for the decoder to denoise. Each layer of the encoder learns a representation of the noisy images. During model evaluation, test set data is passed through these same layers without noise. The decoder reverses the convolutional layers of the encoder by deconvolution and mirrors the encoder configuration.

Evaluation

At the time of writing, the ladder network has not been trained on human-labeled data. We find that our human-labeled dataset is only sufficiently large for model evaluation, even after data augmentation. The natural baseline to our ladder network is the threshold model whose parameters have been set by researchers via HALO.

Table 1: Encoder component of the model

Layer Type	Kernel	Output	Stride	Activation	Notes
Conv	3X3	10	1	elu	Batchnorm prior to activation
Conv	3X3	20	1	elu	BN
Conv	3X3	40	1	elu	BN
Conv	3X3	80	1	elu	BN
MaxPool	2X2		2		
FullyConn		5			

Though a labeled dataset is absent, we may evaluate both models via a biophysical technology called Flow Cytometry : an experimental process which breaksdown the tumor tissue to tally cell types cell-by-cell. Assuming a model draws samples i.i.d from the population, deviation from flow cytometry percentages indicates misclassification is occuring. The comparison is tabulated in the table below. Note that a CD4 positive, a CD8 positive, or a double positive cell must also be CD3 positive.

Table 2: T-Cell types by percentage, computed experimentally and algorithmically

	CD3 %	CD4 %	CD8 %	Total Cells Considered
Threshold	91.7%	74.5%	13.6%	433327
Ladder Network	76%	67.1%	8.4%	49920
flow trial III	85%	74.6%	9.1%	436585
flow trial II	80.8%	71.1%	N/A	434822

In another evaluation approach, we select specific cell examples where the two models disagreed, and evaluate them individually by an oncologist. [evaluation set comparison to be inserted here]

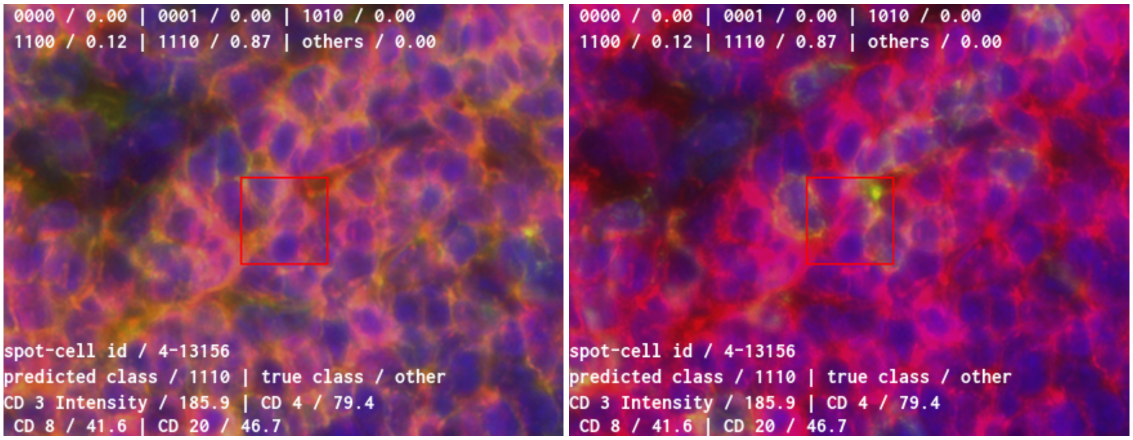
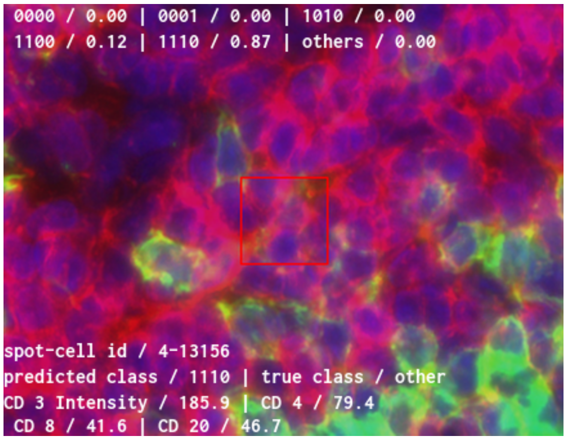

0000 / 0.00 | 0001 / 0.00 | 1010 / 0.00
1100 / 0.12 | 1110 / 0.87 | others / 0.00
spot-cell id / 4-13156
predicted class / 1110 | true class / other
CD 3 Intensity / 185.9 | CD 4 / 79.4
CD 8 / 41.6 | CD 20 / 46.7
0000 / 0.00 | 0001 / 0.00 | 1010 / 0.00
1100 / 0.12 | 1110 / 0.87 | others / 0.00
spot-cell id / 4-13156
predicted class / 1110 | true class / other
CD 3 Intensity / 185.9 | CD 4 / 79.4
CD 8 / 41.6 | CD 20 / 46.7

Figure 3: Three channel views of a captioned example where the models disagreed, with Ladder Network's class probabilities listed above



0000 / 0.00 | 0001 / 0.00 | 1010 / 0.00
1100 / 0.12 | 1110 / 0.87 | others / 0.00

spot-cell id / 4-13156
predicted class / 1110 | true class / other
CD 3 Intensity / 185.9 | CD 4 / 79.4
CD 8 / 41.6 | CD 20 / 46.7

Discussion and Future Work

We showed initial progress towards a semi-supervised procedure using a dataset with limited label information. The approach employs data augmentation, utilization of vast amounts of unlabeled data, and sample priortization to optimize human labeling. Results of this deep learning model, as evaluated by flow cytometry and human cell-by-cell classifcation, show improvements over the threshold baseline. There are a variety of directions for future work.

Methods for decreasing the cost of domain-expertise labeling is much needed. As an auxiliary effort, we developed a cell-labeling program complete with keyboard shortcuts, channel view cycling, and image zoom. However, we find that more advanced features such as dynamic image manipulation, i.e. features commonly found in proprietary microscopy image analysis software such as HALO, to be highly helpful in data labeling.

Data augmentation methods beyond rotation should be tried. In the area of cell images, rotated synthetic data satisfies the important property of being indistinguishable from original data to the human eye. Future data augmentation should follow this principle to prevent model performance degradation.

In comparison to a convolutional neural network, the ladder network has at least tripled the computation time for the same number of batches processed. A common operation throughout the network is batch normalization followed by nonlinear activation. An interesting experiment would be to substitute these two steps with a nonlinear activation function that can also implicitly normalize the activations. Very recent developments such as ELU (implemented in the current model alongside BN) and SELU claim to have this property with proper initializations.

Appendix

Ladder Network

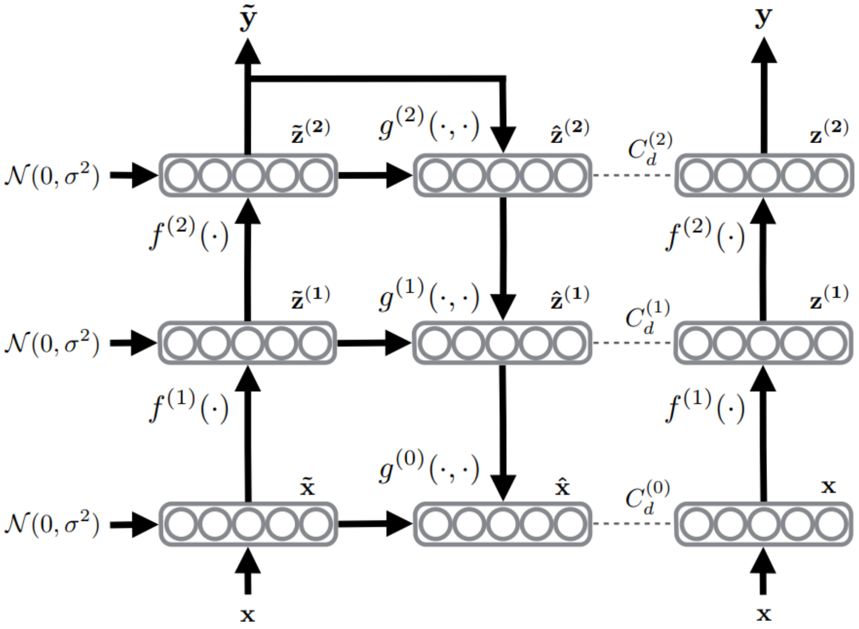


Figure 2: A conceptual illustration of the Ladder network when $L = 2$. The feedforward path ($\mathbf{x} \to \mathbf{z}^{(1)} \to \mathbf{z}^{(2)} \to \mathbf{y}$) shares the mappings $f^{(l)}$ with the corrupted feedforward path, or encoder ($\mathbf{x} \to \tilde{\mathbf{z}}^{(1)} \to \tilde{\mathbf{z}}^{(2)} \to \tilde{\mathbf{y}}$). The decoder ($\tilde{\mathbf{z}}^{(l)} \to \hat{\mathbf{z}}^{(l)} \to \hat{\mathbf{x}}$) consists of the denoising functions $g^{(l)}$ and has cost functions $C_d^{(l)}$ on each layer trying to minimize the difference between $\hat{\mathbf{z}}^{(l)}$ and $\mathbf{z}^{(l)}$. The output $\tilde{\mathbf{y}}$ of the encoder can also be trained to match available labels $t(n)$.

The leftmost path is the so-called corrupted feedforward path. Noise is injected into the sample at every layer of this path to denoise. To summarize why this works at a high level: as the model aims to decrease its denoising cost $\sum_l C_d^{(l)}$, it is compelled to train robust denoising functions $g^{(l)}$ at every layer. In order to train $g(\cdot,\cdot)^{(l)}$ well at layer l, the feedforward layers above l must find features that generate latent representations close to the posterior distribution. Such features are 'cleaner' in the sense that a corresponding latent variable has higher probability values.

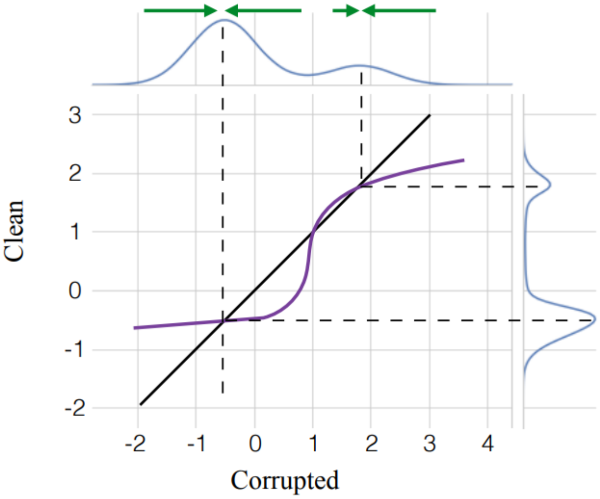


Figure 1: A depiction of an optimal denoising function for a bimodal distribution. The input for the function is the corrupted value (x axis) and the target is the clean value (y axis). The denoising function moves values towards higher probabilities as show by the green arrows.

Model Parameter Choices

The first ladder network was worse computationally and performance-wise. The encoder component is detailed below:

Layer Type	**Kernel**	**Output**	**Stride**	**Activation**	**Notes**
Conv	3X3	20	1	elu	Batchnorm prior to activation
Conv	3X3	40	1	elu	BN
Conv	3X3	80	1	elu	BN
MaxPool	2X2		2		
fully-connected		5			

Given the complexity of cell images and well known practices from image classification literature, it is better to have a deep network with thin layers. Prior to the development of a ladder network, a three layered vanilla ConvNet with batchnorm and dropout was developed (source code found in vanilla_convnet). With threshold labeled data, ConvNet attained 86% accuracy. However the large size disparity between human labeled and unlabeled datasets warranted the added complexity of the ladder network.

Cell segmentation, a common practice in microscopy image analysis, was also tried as a preprocessing step. The procedure zeros out low values in the cell tensor according to a special marker channel which contours the spatial geometry of cell bodies. The overall effect of this procedure is that most pixels outside of the cell membrane have been removed, so as to crop out the central cell. However, this was not used as a preprocessing step because the model suffered additional misclassification following this procedure.

Lastly, an important parameter is the gaussian noise introduced at every encoder layer. As noise increases, the encoder is forced to find more efficient or 'cleaner' representations, since the latent representation passed onto the decoder contains less signal. Experiments show that relative to the MNIST dataset, the tumor cell dataset is much more sensitive to noise. Noise drawn from gaussian distributions with larger variance than $\mathcal{N}(\mu,\, 0.1)$ not only negatively affects the denoising task, but also supervised classification.

Interpreting Labels

Some combinations of CD3, CD4, CD8 and CD20 positivity are impossible, but at least one instance of every combination was found in the HALO CSV file. A combination of labels are encoded into binary as follows:

$$\text{CD3 +, CD4 +, CD8 -, CD20 -} \xrightarrow{\text{binary}} \text{`}1100'$$

The first position always correspond to CD3 $+/-$, the second to CD4, and so on.

In confusion matrices, summing over a column such as '0000' yields the number of times our model predicted CD3 -, CD4 -, CD8 -, CD20 -. Summing over a row such as the first one, also corresponding to '000', yields the true number of CD3 -, CD4 -, CD8 -, CD20 - cells.