2018 Kaggle Competition

Xinran Zhou(xz60), Peiguang Wang(pw15)
Rice University, Houston, Texas

Project Overview

Problem Description

In this competition, we are asked to create a computer model that can identify a range of nuclei across varied conditions. The program should automatically identify and extract the nuclei area.

Our Work

We first analyze the data and do data exploration. We tried to classified based on size, color, shape number of nuclei, and try to identify the potential noise. We also did statistical analysis on the data. After EDA, We use CV methods and U-Net models to solve the problem. CV methods is trying to solve the problem from the traditional computer vision point of view. U-NET is a modified version of fully connected network, which is considered one of standard architectures for image classification tasks. We train our network in Google "colab" platform and do prediction on test set locally.

Result

We submit 11 times in stage 1 and we submitted 1 time in stage 2. Stage 1 score(highest) is 0.344. Our stage 2 score is 0.44. Our Rank is 392/861.

Workflow

Data → Preprocessing → Model → Postprocessing → Output

Pre-processing: Do preprocessing to the raw data. This includes resizing the original images, image normalization, and transform the data into the desired representation. In U-Net model, transform the data into a matrix; in CV model, we also need to compute the histogram of the image.

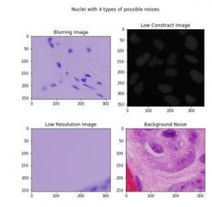
Model: We used CV model and U-Net network model.

Post-processing: Do postprocessing to the outcome from the model. This includes resizing the masks, augmenting the masks, and line encoding the results.

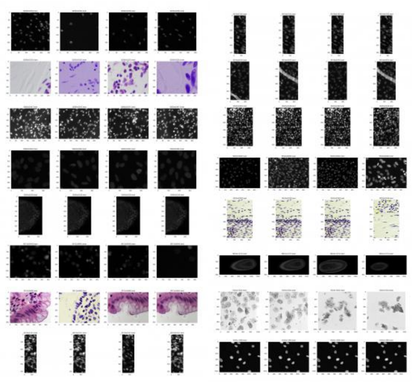
Data Analysis & preprocessing

4 important EDA results

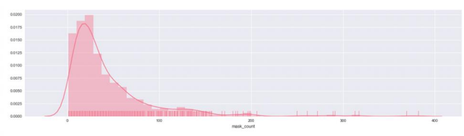
Potential Noise



Different Sizes

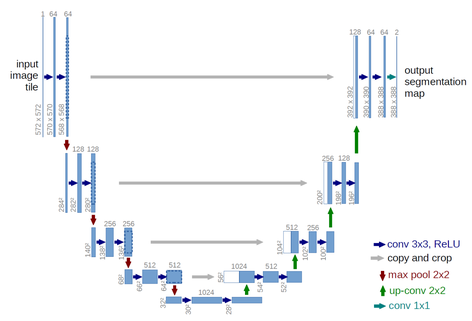


Nuclei number distribution



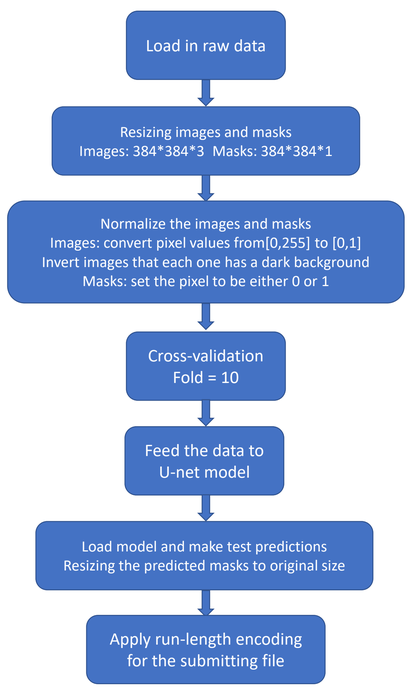
Height & Width distribution

U-Net Model



U-Net Introduction and Features

- U-NET is considered one of standard architectures for image classification tasks.
- In the up-sampling part, a large number of feature channels is added to the original channels. It allows the network to propagate context information to higher resolution layers.
- The network does not have any fully connected layers and only use the valid part of each convolution.



U-Net Workflow

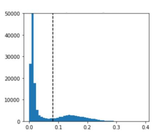
CV Methods

Overview

CV method is to setup a reasonable threshold and set pixels that above the threshold to 1. The CV method achieves 0.22 accuracy (stage 1 score).

How to choose a reasonable threshold

- Assumption (1): Assume that the background pixels are more than nuclei pixels and remove the background based on this.
- Assumption (2): Treats the image histogram as two additive distributions.
- Find threshold value based on Otsu's method.

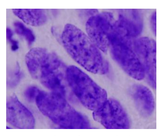


Grayscale Histogram

After Removing Background

Why it fails

- The first assumption can be wrong. Some special cases that CV method will fail. Image below shows a failure case.



Special Case for CV method

Improvements on U-Net

How to Overcome Overfit

- Add dropout in our model. This is implemented by only keeping a neuron active with some probability p or setting it to zero otherwise.
- "Early stop" trick. This trick saves the best model every epoch and stop if the accuracy in validation set does not decrease.
- In our final model, the dropout rate is set to 0.3.

Learning Rate Adjustment

- Reduce Learning periodically. Starting from learning rate of 0.0015, and after every 3 epochs, we reduce the learning rate of 40%.

Other Improvements

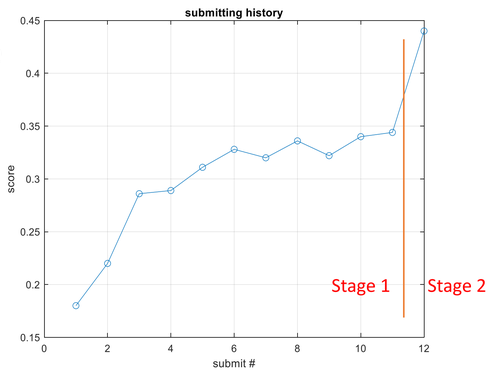
- We trained the model for 10 epochs and add early stop to save the best result
- Update score metric

Submit History & Result

Submitting History

We submit 11 times in stage 1 and we submitted 1 time in stage 2.

Submit #	Stage	Model	Score	Milestone
1	Stage 1	CV	0.18	
2	Stage 1	CV	0.22	Normalization
3	Stage 1	U-Net	0.286	
4	Stage 1	U-Net	0.289	Epochs
5	Stage 1	U-Net	0.311	Resize
6	Stage 1	U-Net	0.328	Dropout rate
7	Stage 1	U-Net	0.320	
8	Stage 1	U-Net	0.336	Learning rate
9	Stage 1	U-Net	0.322	
10	Stage 1	U-Net	0.340	
11	Stage 1	U-Net	0.344	
12	Stage 2	U-Net	0.44	



Result

Stage 1 score(highest) is 0.344
Stage 2 score is 0.44.
Rank 392/861.