Joshua Paik

Master’s in Math

Doing data science

1st stage top 1%, 2nd stage top 5% = 250th.

Problems of image masking.

Identify cell nuclei from microscopy data.

Why? To advance research on DNA of cell.

How? Use existing neural network architecture platforms and image processing techniques.

Original image = black and white.

Mask = colored. White became yellow, black became purple.

Count cells in each cell

We want the union of cell is 80% (goal). We can start using in medical research.

Right now it is 65%.

Kaggle Data Science Bowl 2018:

Exploratory Data Analysis:

Distribution of the width, height and area of the masks.

*There was a lot of variance*

Distribution of number of nuclei in each images.

*1 to 400 nuclei variation*

Smallest nuclei problem

*When you feed an image in a neural network, you want the image to be the same size.*

*It becomes dependent on scaling, the image will not be visible then.*

Overlapping cell problem

And.. there are a lot of overlapping cells – main issue

K means classification Fluorescent: 81.5% (purplish – polarized light)

K means classification Histological: 16.1% (black and white dots, shoot the right laser you don’t see the body of the nucleus)

K means classification Bright-field: 2.4% (like x-ray, highlighting the black dots)

*Different magnification*

If you can create an algorithm with purple color, easier?

*Limited data set: 600 images.*

*Create image mask*

*Pullout green channel*

*def parametric\_pipeline(img\_green, invert\_thresh\_pd = 10*

Non-machine learning approach (above).

What type of problem is the Kaggle DSBowl

Semantic Segmentation -> capture the outline and filled with color

Grass, Cat, Tree, Sky => no objects, just pixels

Classification + Localization -> captures the object

Cat => single object

Object Detection -> captures objects.

Dog, Dog, Cat => Multiple Object

Instance Segmentation -> capture objects’ outline and filled with color.

Dog, Dog, Cat => Multiple Object

Explanation of UNET: input image file to output segmentation map.

Is a neural network architecture, flow of information in a U-shape.

*In the U shape.*

*First arrow, convolution layer. Take a matrix, 3x3, iteratively, take the dot product. That’s your new image. Reduces the size of image.*

Copy and crop.

Max pool

Single depth slice

*Occur every step down in the U.*

*Up convolutions, take the small matrices, and puts to bigger space. Performs dot product operation plus padding. It will make the image bigger.*

*UNET. Image encoding and decoding. And also produces image mask.*

*Main benefit of UNET compared to mask or CNN. It involves 2 neural networks, and outputs 2 target boxes. Height, width.*

UNET – Overlapping Tile Strategy

*A way of creating an image mask, is creation of a very small box. N-squared notation, back-propagation.*

UNET – Accuracy and Method

*Penalized edges, forces edges to be created.*

*Instead of allowing blends to happen, it will severely punish.*

First Implementation:

Pre Process: Data Augmentation Elastic Deformations

*If you perform slight deformation, ….*

*Elastic deformations – won’t change the structural property.*

Data Augmentation:

My Solution:

Winning Solutions:

Clache, Sharpen, Embass

Gaussion Noise

Color to Gray

Inverting

Remaping grayscale images to random color images

Blur, Median Blur, Motion Blur

Contract and brightness

Random scale, rotates, and flips

Heavy geometric

Post Processing: Watershed Algorithm

*Watershed algorithm allows you to check for separation and separates them.*

*PyTorch*

Data Variation of Stage 2 (competition)

* Using unseen data

Q&A

*How did you tune your hyper parameters?  
learning-ray(?) Or initializations.*

*Learning ray is monotonically decreasing.*

*If learning ray is upward curve, learning curve is too high.*

*If spikes, undetermined.*

*If converges after 4 epochs(?), learning curve is too low.*

Keras vs PyTorch

*14 hours to training the model*

*Are there machine learning searching for better machine learning?*

*Grid searching.*

*You can create a learner, looks at every epoch, and checks the better learning model.*

*Given enough neurons, and layers, better learning.*

*The stuff the winners did was preprocessing.*

*In general, the machine learning tries to do, transforms the r2 space.*

*Two circles, two layers, no math for that for now.*

*Machine learning solves by continuous function. You can do regression.*

*Theorem states that they can learn in a continuous function with a big enough neuron. You can have discrete functions.*

*Google developed an algorithm producing algorithm.*

*You can’t throw random crap at a neural network, and produce good stuff.*

*Training two connected neural networks is pretty hard.*

*Traditionally what biologist have to do.*

*Kaggle Competition of Whale.*

*Finding individual whale, this Whale is Jim, this Whale is Sam.*

*They got their HR department to label them.*

*Aerial footage of rainforest*

*UNET is so cheap*

*Did you write separate code for GPU?*

*PyTorch, Keras, …*

*Hardest problem is installing the GPU.*

*Google Collab = Jupyter Notebook.*