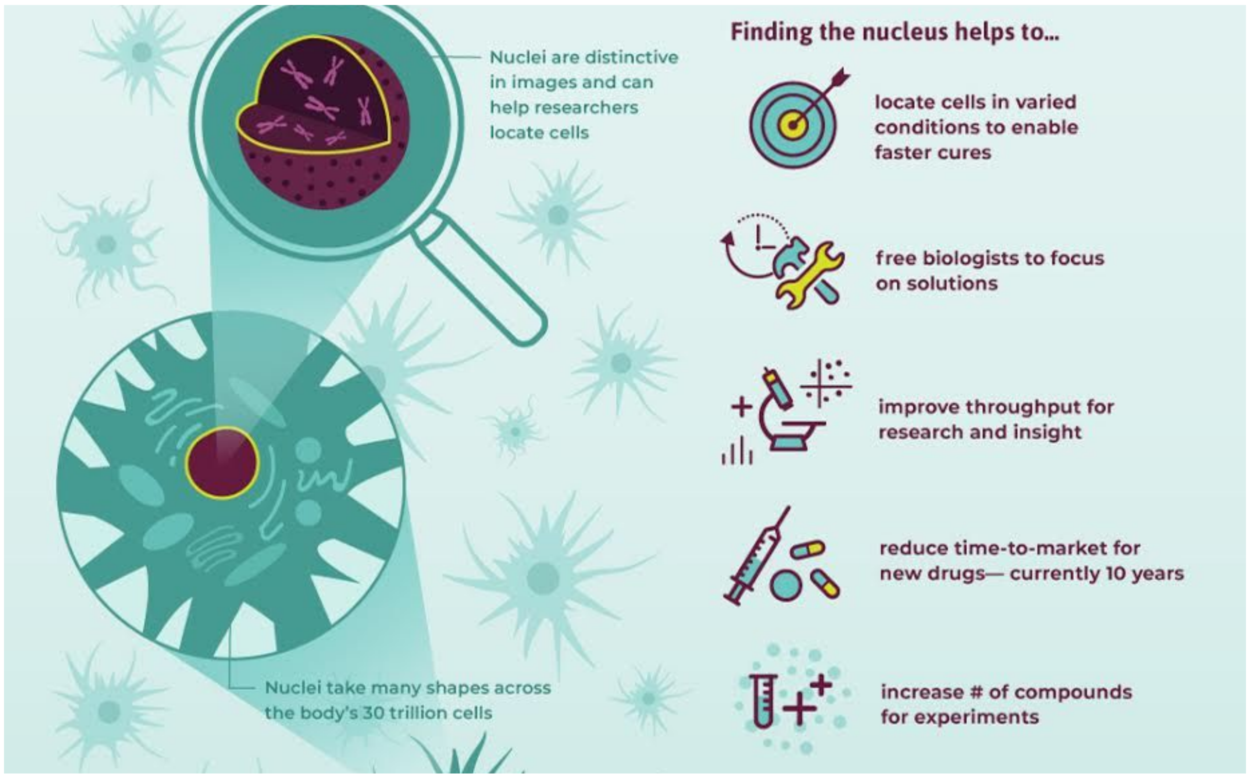
Nuclei are distinctive in images and can help researchers locate cells
Nuclei take many shapes across the body's 30 trillion cells
Finding the nucleus helps to...
locate cells in varied conditions to enable faster cures
free biologists to focus on solutions
improve throughput for research and insight
reduce time-to-market for new drugs— currently 10 years
increase # of compounds for experiments

Find the nuclei in divergent images to advance medical discovery

Capstone Project Proposal

15.05.2018

—

Jimit Jaishwal

Domain Background

Imagine speeding up research for almost every disease, from lung cancer and heart disease to rare disorders. We've all seen people suffer from diseases like cancer, heart disease, chronic obstructive pulmonary disease, Alzheimer's, and diabetes. Many have seen their loved ones pass away. Think how many lives would be transformed if cures came faster.By automating nucleus detection, you could help unlock cures faster—from rare disorders to the common cold. Please saw this video you can better understand problem, this is a part of Data Science Bowl 2018 in kaggle - View This Video

Problem Statement

Identifying the cells' nuclei is the starting point for most analysis because human body's 30 trillion cells contain a nucleus full of DNA, the genetic code that programs each cell. Identifying nuclei allows researchers to determine each cell in a sample, and by measuring how cells react to various treatments, the researcher can understand the underlying biological processes at work.

I've used this dataset - Find the nuclei in divergent images to advance medical discovery we will created a computer model that can identify a range of nuclei across varied conditions.

Datasets and Inputs

Our dataset contains a large number nuclei images.there are different different type cell available inside our images.

We have already training and testing data available, Each image is represented by an associated ImageId. Files belonging to an image are contained in a folder with this ImageId. Within this folder are two subfolders, Images and Mask,

- `Images` this folder contains the image file. And every image name is same main folder name
- `masks` this folder contains the segmented masks of each nucleus. This folder is only included in the training set. Each mask contains one nucleus.

File descriptions

- `/train/*` - training set images (images and annotated masks) Download
- `/test/*` - stage 1 test set images (images only, you are predicting the masks) Download

The most important file is `train` training set images and annotated masks, we have 670 training examples, there are different sizes images in our dataset but we resize those images to fit in our machine learning model, our image resized size is 256 * 256 pixels and depth is 3

Solution Statement

I am using deep learning algorithm and make with Tensorflow/Keras and trained with training data. In this project I use Convolutional neural network will be implemented Tensorflow/Keras library and optimized to minimize multi-class logarithmic loss as defined in Evolution and Metrics section, In many CNN architecture I use U-net Convolutional neural network. The U-net is convolutional network architecture for fast and precise segmentation of images. Prediction will be made on the test dataset and will be evaluated.

Benchmark Model

The model with the Private Leaderboard score(Intersection Over Union) of 0.614 will be used as a benchmark model. Attempt will be made so that score(Intersection Over Union) obtained will be among the top 50% of the Private Leaderboard submissions.

Evaluation Metrics

Submissions are evaluated using the Multiclass logarithmic loss.Multiclass logarithmic loss is used to assess predictive models in high-cardinality classification problems. Each image has been labeled with one mask image file. For each image, you must submit a set of predicted mask image (one for every image). The formula is then,

$$\text{log-loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(p_{ij}\right)$$

where N is the number of images in the test set, M is the number of image class labels as different type Image Mask file, log is the natural logarithm, y is the true mask image and p is predicted mask image.

Intersection Over Union Metric(IoU)

This project is evaluated on the mean average precision at different intersection over union (IoU) thresholds. The IoU of a proposed set of object pixels and a set of true object pixels is calculated as:

$$IoU(A, B) = \frac{A \cap B}{A \cup B}.$$

The metric sweeps over a range of IoU thresholds, at each point calculating an average precision value. The threshold values range from 0.5 to 0.95 with a step size of 0.05: (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). In other words, at a threshold of 0.5, a predicted object is considered a "hit" if its intersection over union with a ground truth object is greater than 0.5.

At each threshold value tt, a precision value is calculated based on the number of true positives (TP), false negatives (FN), and false positives (FP) resulting from comparing the predicted object to all ground truth objects:

$$\frac{TP(t)}{TP(t) + FP(t) + FN(t)}.$$

A true positive is counted when a single predicted object matches a ground truth object with an IoU above the threshold. A false positive indicates a predicted object had no associated ground truth object. A false negative indicates a ground truth object had no associated predicted object. The average precision of a single image is then calculated as the mean of the above precision values at each IoU threshold:

$$\frac{1}{|thresholds|} \sum_t \frac{TP(t)}{TP(t) + FP(t) + FN(t)}.$$

Lastly, the score returned by the project metric is the mean taken over the individual average precisions of each image in the dataset.

Project Design

The general sequence of steps are as follows:

- **Data Visualization** - Visual various type images and mask file of nuclei, we also use visualization to show predicted mask file of give microscopic images.
- **Data Preprocessing** - we scale and normalize our all features.The values should be in the range of 0 to 1
- **Model Selection** - In our project dataset is images so i choose deep learning U-net CNN model, because that was good image segmentation.

We have already training and testing dataset. When we train our model then we use our test dataset and predict mask files and save results.

References

- https://www.kaggle.com/c/data-science-bowl-2018
- https://www.kaggle.com/c/data-science-bowl-2018/data
- https://www.kaggle.com/c/data-science-bowl-2018#description
- https://www.kaggle.com/c/data-science-bowl-2018#evaluation
- https://arxiv.org/pdf/1505.04597.pdf
- https://www.datasciencecentral.com/profiles/blogs/polymorphic-malware-detection-using-sequence-classification
- https://storage.googleapis.com/kaggle-media/competitions/dsb-2018/dsb.jpg
- https://www.youtube.com/watch?v=eHwkfhmJexs&feature=youtu.be