Kaggle 2018 Competition

**Useful information**

**Resources:**

Kaggle kernel Page: <https://www.kaggle.com/c/data-science-bowl-2018>

Video Introduction: <https://datasciencebowl.com/2018dsbtutorial/>

**TensorFlow:** we installed the TensorFlow 1.5.0, for more detail information <https://www.tensorflow.org/>

**Environment Configuration**

**[Windows 10]**

Creating a new python 3.5 environment in Anaconda. Then install:

TensorFlow (version 1.5.0).

OpenCV (version 3.3.1). Installed in Anaconda Navigator.

Sk-learn (version 0.19.1). Installed in Anaconda Navigator.

matplotlib (version 2.1.2). Installed in Anaconda Navigator.

pandas (version 0.22.0). Installed in Anaconda Navigator.

seaborn (version 0.8.1). This package is used for statistical data visualization. Installed in Anaconda Navigator.

**[IDE]**

Install Jupyter Notebook. Installed in Anaconda Navigator.

**EDA**

**Reference:**

[Some dataset exploration on Kaggle kernel]

1. Exploratory Analysis – *Jerry Thomas* <https://www.kaggle.com/jerrythomas/exploratory-analysis>
2. Exploratory Analysis - image stats – *Wesley Goi* <https://www.kaggle.com/etheleon/exploratory-analysis-image-stats>
3. Simple data science bowl 2018 exploration – *David Retana* <https://www.kaggle.com/davidretana/simple-data-science-bowl-2018-exploration>

Every one of the perspective is quite different. But they all look at the distribution of the nucleis. And they all notice the background problem.

**First look at the Dataset**

This file is to collect the training images and copy them into one single file. So we can actually see the data and have a better understanding of the pictures. Then we have a first look at the data set. We them display different types of the images with the following focuses:

* Color
* Number of nuclei
* Size and shape
* The different quality of the images

**Color**

Based on our observation, we sorted the images into four classes.

* Colored images with white background.
* Colored images with purple background.
* B&W images with black background.
* B&W images with white background.

First we convert these color images into gray-scale images. When converting, we use:

to normalize the image.

Then we look at the histograms of the four types of gray-scale images. We found that the histogram of B&W images with black background is different from the others. While others look very similar.

**Number of nuclei**

Number of the nuclei on every image varies. Some pictures only have few nuclei, while some have hundreds of nuclei.

**Size and shape**

The size and shape of the nuclei can be very different. Below shows some different sizes and shapes of the nuclei.

**The quality of images, possible noises**

According to the images, we identify four different types of the noises:

* blurring
* low resolution
* low contract
* background noise

The image below shows the four different types of the noises.