Blood Cells Classification with CNN

Jiaqi Hu, Fan Liang, Tuo Wu

Georgetown University, Master of Science in Analytics

Abstract

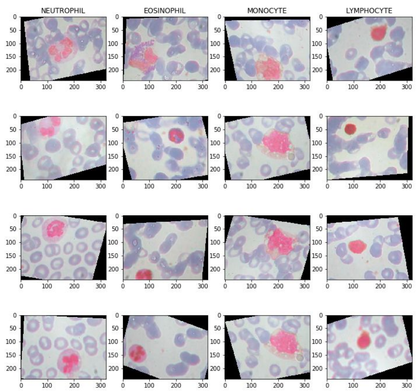
Convolutional neural network is commonly applied to analyzing visual imagery. In CNN, the Input is a tensor with shape (number of images) x (image width) x (image height) x (image depth), and convolutional kernels whose width and height are hyper-parameters, and whose depth must be equal to that of the image. In this project, we use CNN to research on image classification.

Introduction and Problem Statement

The diagnosis of blood-based diseases often involves identifying and characterizing patient blood samples. Blood Cell classification is important for these diseases' diagnosis and treatment.

In this project, we use CNN to recognize the colored cell in each image and classify it in 4 types of blood cell: Eosinophil, Lymphocyte, Monocyte, and Neutrophil. And we will compare the loss and accuracy rate with different dropout rates.

Sample images of 4 types of blood cells:



Related Work

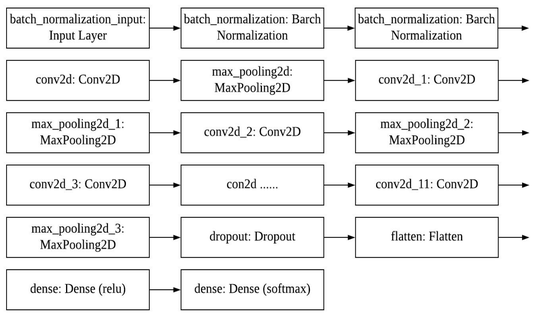
The diagnosis of blood-based diseases often involves identifying and characterizing patient blood samples.

From Yale's Peripheral Blood Lab, the identification of blood cells is based primarily on observations of the presence or absence of a nucleus and cytoplasmic granules. And previous kaggle researchers also used CNN to classify the blood cells. The previous models have over 400 thousand parameters. Our project focuses more on the effects of different dropout rates and try to simplify the CNN model to have similar accuracy rates.

Dataset and Features

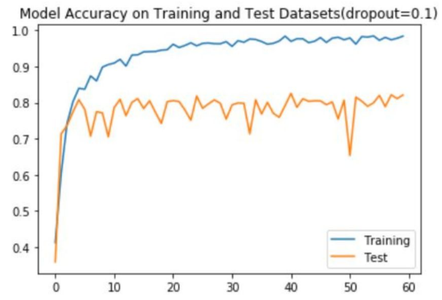
We use dataset called Blood Cell Images from Kaggle. This dataset has 12,500 augmented images of blood cell, and these images are labeled by cell type in csv file. There are four different type of cells: Eosinophil, Lymphocyte, Monocyte, and Neutrophil. There are approximately 3000 images for each of 4 cell types splitted in training and test datasets.

Description of Model



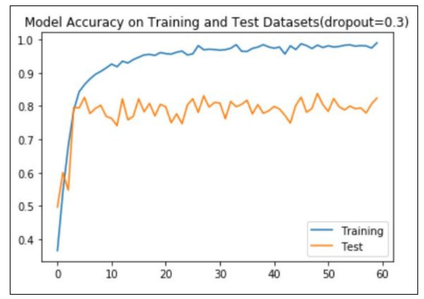
Results

We have built many different models to test accuracy. The main differences are dropout rate and number of dropout layer.

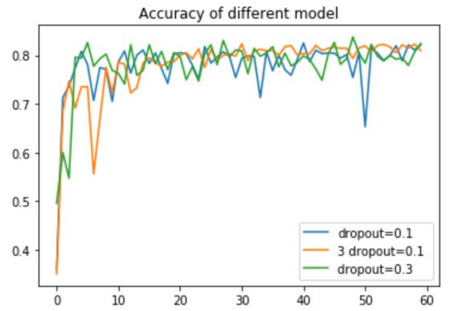


In this model, we set dropout rate as 0.1 which means every node has 10% possibility to be ignored. In this graph, we found this model has much better performance in training data. When iteration went through, accuracy of training dataset is increasing. However, the accuracy of validation dataset has large fluctuation when iteration went through.

In this model, we set dropout as 0.3. We found accuracy of validation set increase a little bit. Also, accuracy curve is more stable than first model.



In this model, we add three dropout layers and dropout rate is 0.1. We did not find dramatic increase in accuracy of validation set.



When we compare these three models, we find model whose dropout rate is 0.3 has the highest accuracy, model which has three layers of dropout has the most smooth accuracy curve as iteration went thought.

Bibliography

1. Mooney, P. (2018, April 21). Blood Cell Images. Retrieved September 20, 2019, from https://www.kaggle.com/paultimothymooney/blood-cells.
2. Systems Cell Biology@Yale. (n.d.). Retrieved December 1, 2019, from http://medcell.med.yale.edu/systems_cell_biology/blood_lab.php.