NIH Research Festival

Machine Learning-based Production of Training Images
to Use in Image Analysis Workflow

Team Leads: Xinlian Liu and Yanling Liu

September 10–12, 2018

1 Introduction

This hackathon addresses current challenges with a community coding effort which will result in open-source solutions in applying machine learning to image analysis at NIH.

2 The Image Analysis Workflow

The image analysis workflow is descried as such: After a large image data set is acquired, some quality control methods to convert them into a suitable format. Working with domain experts using data analysis tools, subsets of the images are identified for further analysis. We use Generative Adversarial Network to generate synthesized images to mitigate data imbalance. At this state, the resulting data is ready for common deep learning tasks such as classification, detection, and segmentation.

2.1 Acquiring Data

The Data is acquired from CXR14 Chest X-Ray Images, published by Dr. Summer's Group of NIH.

2.2 Data Quality Control

In this step we remove images that are either wrongly included or not usable for any reason.

- size/resolution variation
- image recording format variation (extra alpha channel, color vs. grayscale, file permissions, etc.)
- wrong position, wrong orientation

- bad scan

Hackathon participants can contribute algorithms and implementations to:

- detect images with poor quality/value
- correct salvageable images by applying intelligent transformations

2.3 Data Exploration

In this step we learn more about characteristics of the data and make decisions on identifying a subgroup that can be used for final goals. For example, by analyzing distributions of available data and consulting with medical experts, we set the targeting group demographics as: 20-70 years old male with PA position.

Hackathon participants can help us set optimal range for over 800 phenotype combinations.

2.4 GAN with Convolutional Networks

We have prepared a GAN model with deep convolutional networks as both generator and discriminator.

Participants are welcome to explore different network architectures, or port the solution to Tensorflow for easier performance tuning or PyTorch as a flavor variation.

2.5 Containerization and Workflow

Participants are encouraged to create a Docker container and other workflow tool Binder of this project.

2.6 Performance Improvement

The project may utilize multi-GPUs or distributed multi-nodes acceleration.

2.7 GAN with VAE

Participants are welcome to explore combining Variational Auto-Encoder as the generator and use it in a GAN network.

2.8 Examples of Image Analysis

2.8.1 A classifier with adequate data

We will provide a classifier between images of healthy lungs and images of a phenotype with adequate training data such as 'infiltration'.

2.8.2 A classifier with inadequate data

We will provide a classifier between images of healthy lungs and images of a phenotype with inadquate training data such as 'mass'. We will also provide an enhanced data set with synthesized 'mass' data for improved performance.

Hackathon participants are welcome to work on alternative algorithms and implementations for the classifiers or expand the work to other domains such as segmentation.

3 Platform and Skills

We expect participants are familiar with Deep Learning with Python and job submissions on ~~BioWulf~~ AWS. Experiences with tools and libraries such as Pandas, Matplotlib, and PIL, etc. will be helpful.

4 Conclusion

This hackathon project will examine the readiness of the NIH community on applying machine learning in image processing workflow. Code contribution will also enhance the capacity of deep learning in image analysis at Frederick National Lab for Cancer Research.