2021 Data Science Institute Challenge Problem

Welcome to the 2021 Lawrence Livermore National Laboratory Data Challenge! We are all so excited to see your innovative solutions to such a cutting-edge problem. Now that you have set up your environment, this document will serve as a guide through this 3-week challenge!

Classifying Stars and Galaxies (Mini-Challenge):

With ever-increasing large-scale photometric surveys that collect data on upwards of billions of stars and galaxies, it is impossible for human experts to manually classify all of them, and therefore there is a large need for these images to classified automatically, and it is time to put your team's skills to the test to classifying stars vs. galaxies.

This data is taken from the HSC survey https://hsc.mtk.nao.ac.jp/ssp/survey/

In the "Working_with_stargalaxy_data.ipynb", jupyter notebook, you can see examples of how to access the csv file, and look at all of the different passbands and find the label of whether it is a star or a galaxy.

Within the stargal.csv file, there are 8 different "channels" within each image, 4 representing different bands of the image, (think it to be similar to the previously discussed "color channels"), and 4 channels representing the point spread function of each band. The first value of each image represents the classification of the data 0 (for stars) and 1 (for galaxies). First, choose a method to normalize the data, and then try your best to classify the images given the data using a CNN.

Here is your task:

1. Preprocess, Normalize Data and build a CNN classifier
2. Report overall accuracy, star accuracy, galaxy accuracy
3. Report how long it took to train and test
4. Show Confusion Matrix
5. Report Precision and Recall
6. Build ROC curve
7. Experiment with only using certain bands (does accuracy improve?)
8. (optional) Compare your results to our Astronomy Group's new paper https://arxiv.org/pdf/2105.01106.pdf

Asteroid Detector Challenge:

The Zwicky Transient Facility (ZTF) is a new time-domain survey that had first light at Palomar Observatory in 2017. Building on the highly successful legacy of the Palomar Transient Factory (PTF), ZTF uses a new camera with a 47 square degree field of view mounted on the Samuel Oschin 48-inch Schmidt telescope. ZTF's extremely wide field and fast readout electronics enable a survey more than an order of magnitude faster than that of PTF.

We have prepared difference images taken straight from ZTF need your help building a detection algorithm to find asteroids in the data to help us on our mission of planetary defense. This is an active research problem, and unlike star galaxy classification, you will be building your own training and testing set, running experiments, and seeing what conclusions you can make from the data.

A common approach to build a classifier in “data-starved” environments is to generate synthetic examples to which you can train, and use the classifier to find real examples. Our team created these synthetic asteroids and injected them into real science images from the ZTF telescope on which you can train. The data set contains 1000 images with 20 asteroids injected into each image. These difference images are roughly 3000x3000 pixels wide, whereas an injected asteroid only affects around 25 pixels, (talk about needle in a haystack!). We saved the entire difference image, as well as some header information. Checkout the “working with asteroids” jupyter notebook to see how to work with the data we saved for you all.

Here is your task:

1. Build Asteroid/No-Asteroid Classifier
    a. Build training set/testing set (normalization/preprocessing etc.)
    b. Determine how large you want each image to be
    c. Experiment with how classifier does if asteroid is in the center vs. edges/size of image/other interesting experiments and other tradeoffs you can think of
2. Find way to search through images intelligently
3. Report similar statistics as in the star galaxy (confusion matrix, precision recall, how long to search through images, roc curves, interesting experiments etc.)
4. (optional) Work with our team to run your search algorithm on images without synthetically injected asteroids