Data 603: Project Guidelines

A major goal of this class is to become comfortable processing a large amount of data. For the project you will be evaluating the Google Open Image Dataset in a manner of your choosing.

## About the Dataset

The [Google Open Image Dataset V5](https://storage.googleapis.com/openimages/web/index.html) contains ~9 Million images and is around 18TB of data. The dataset is split into a training set (9,011,219 images), a validation set (41,620 images), and a test set (125,436 images). The images are annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships.

For this project, we will be working with a *subset* of the images that have bounding boxes, object segmentations and visual relationships defined. This subset is split into 1,743,042 training images and the full validation (41,620 images) and test (125,436 images) subsets. The images are rescaled to have at most 1024 pixels on their longest side, while preserving their original aspect-ratio. The total size is 561GB plus a few GB of metadata files.

## Project Requirements

All work is to be done on the UMBC Big Data Cluster. The cluster contains the image subset and all required metadata. Your project must process each image at least once, and will most likely process a subset of the images several times.

## Part 1: Exploratory Data Analysis

### Metadata Exploratory Data Analysis

Is the dataset consistent? (e.g are there labels for each image?)(bounding boxes per image?)

Histograms to show Distributions of metadata values.

Are there any differences between Train/Test/Validation subsets?

Teach us something about the data we didn’t know.

Visualize the data somehow.

What techniques are you using? What libraries?

What subsets/facets are available?

### Feature Extraction

Run algorithms across the entire data set to extract a small piece of information from each image.

Normalize the Feature values.

Plot distributions/histograms and Cumulative Distribution Functions.

Attempt to fit distributions to the features. Does the data match the fit?

Compare your features to the EDA analysis. Do your features have the same trends, distributions, etc.

## Part 2: Model Evaluation

This section will evaluate a pre-trained model of your choosing. With instructor approval, a novel idea of your choosing may replace running a pre-trained model.

### Research

Find a pre-trained model of interest to evaluate over the open image dataset.

Keras: <https://keras.io/applications/> (e.g. ResNet50, MobileNet, etc)

Read the relevant papers! Read the papers the model authors cite. Why did they create this model? (i.e. MobileNet isn’t as good but way faster) Try to find if anyone has done this comparison before. What are the differences between this data set and the training data set.

### Data Subset Identification and Extraction

Identify a subset of the data for further processing. For example, if you’re doing a face detection model, find images with faces, and extract the relevant pieces of the image. Depending on the model chosen, images or pieces of images will have to be downsampled to an appropriate image

### Running a Model

Run the model on the subset of images extracted above.

How did it perform compared to the original creators?

Did it perform on a particular subset of the extracted images?

Iterate on the data extraction technique to see if you can improve accuracy.

Basically, do EDA on the model run outputs. Create a confusion matrix.

## Part 3: Custom Work

* Retrain the last stage of the model using the OID labels. Re-evaluate the performance of the model.
* Another option is to devise something novel after approval from the instructor. Perhaps re-do ETL, evaluate performance of different layouts, file-formats, etc.
* Something of your own design. Photo collage, photo
* Update the data on the cluster
* Join some other data! (a lot of other data)
* Explore another big data system
* I’ll work with you! The right answer isn’t the point, showing you can process data is.