Data Engineer Take Home

Note

- **Please explain your answers and write based on your understanding**
- **Explain your design choices**

Instructions

- You have ***one day*** to finish the assignment.
- We will evaluate your code for accuracy, correctness, code quality, and comments.
- Create a private Github repository called `RD-Data-Takehome-[Firstname]-[Lastname]`
- Create a `.yaml` file with all environment dependencies and a README on how to run your code. You may need the following libraries:
  - Pandas, NumPy, SciPy, Matplotlib, PIL
  - PyTorch or Tensorflow
- For each question, add a folder `question_01`, etc.
- Finally, share your repository with our GitHub Username: `realitydefendercoder`

1. Design & build a small dataset (about 100+ images) to differentiate between real and fake face images. Please explain:

a. Considerations that went into deciding what data to collect.

b. How you went about collecting the data.

c. Besides fake/real labels, what other labels would you consider? Explain a simple method to sample a uniform dataset in the i.i.d sense, given the labels.

d. What API (e.g Pandas, etc.) you used to store and organize meta information about the dataset.

e. Please share your mini-dataset as a `zip` file.

2. You are given a classifier that reports high accuracy on the validation set:

   a. Would you be happy with these results or would you like to do more analysis.

   b. If so, what type of analysis would you perform?

3. Write a simple (supervised) deep classifier to train and test using the dataset collected in Q1.

   a. How will you divide your dataset into training and test sets.

   b. What data-augmentation techniques will to use for out-of-distribution (unseen) images?

   c. Please test accuracy on the attached, `rd_test_dataset` zipped face images, and save the output to a `.csv` file.

      https://drive.google.com/file/d/1jcdByJPkAGq9JsgsdLqeyLwI4Yl6plOf/view?usp=sharing

4. Now, consider the case where you had to manage a dataset with millions images rather than a few hundred. How will you change your dataset building and storing methods for:

   a. Faster access, given that data data lives on the cloud infrastructure like S3

   b. Faster data re-sampling, to create custom datasets

   c. Faster data-loader access for faster training