**Folders:**

1. ‘train’ is the original training set, this folder is required for ‘scrap image from google.ipynb’ python Jupyter notebook to work as the notebook reads the labels of the subolders to use as search queries. (refer to notebook for details).
2. ‘test’ is the original test set, used for prediction for the submission file. All test images are copied into a subfolder ‘test2’ in this folder, keep it this way for the code to work as the code requires the test folder to have images in subfolders.
3. ‘train\_expanded\_character’ includes both the original training set (with original file names) augmented with the images from webscraping obtained via the ‘scrap image from google.ipynb’ code.

The ‘final code.ipynb’ runs on this train dataset to train the model. Keep the folder this way, the code uses the subfolder names as the image labels.

Refer to the disclaimer in the report (also copied below) and the webscraping section of report for more information, and also the comments in the ‘scrap image from google.ipynb’ code as to how the images were obtained.

|  |
| --- |
| *Disclaimer:*  *As the training images that were obtained via webscraping were saved directly onto Google Colab’s Linux runtime session to save time, the* ***original additional images to train the data are automatically deleted*** *after 12 hours. Great care has been taken to reproduce these additional images for training by running the webscraping code to download the images again and zipping them up for CAX to use.*  ***These final training images are in the zipped folder named ‘train\_expanded\_character’ within the uploaded code folder.*** |

1. ‘expand\_train\_set\_character’ includes only the images from webscraping, for reference only.
2. ‘valid’ folder allows user to randomly put images from the training set into it for validation. Refer to last part of the report for the reason that there is only 1 dummy image in it now.

*Before the final submission, only 1 dummy image is placed in the validation directory (as this code requires validation directory with at least 1 image to run, feel free to randomly copy and paste any amount of images from the training set to this directory to use as validation images when running it). This is to ensure that the maximum number of images are used for training.*

**Codes (Python Jupyter Notebook):**

All codes have to be run via jupyter notebook.

1. ‘scrap image from google.ipynb’ provides code to scrap images to add to training set. Refer to webscraping section of report for more information. This code has full comments as to how the images were obtained and how to use it.
2. ‘final code.ipynb’ contains the full code for training and predicting. This code also has very detailed comments as to how everything works. Just run this code with all the folders in place until the end to output the submission file.