Student: Elshan Naghizade

Project: Python Library To Transform Image Datasets in Query-able SQL Tables

Date: 11 July

CURRENT PROGRESS

The main modules have already been built:

- ✓ **Feature Extractor** – the refactored functions to retrieve dimensions, color channels, edges, color histograms, corners, etc. from a variety of image formats.
- ✓ **BLOB Image Storage** – 2 functions capable of transforming numpy arrays into BLOBs (since most features and images are stored as arrays) for PostreSQL and SQLite.
- ✓ **SQL Transformation:**
  - ***Lite version*** - Generation of SQL statements as texts and storing them as .sql files.
  - **Sophisticated version** – Using psycopg2 or sqlite3 to directly operate on database systems.

COMPROMISES AND CHALLENGES

1) **Memory optimization** – Gigabyte-tier datasets require batch-by-batch processing which begs the question of the optimal batch size to be processed at once.
2) **Folder traversal** – Traversing subdirectories of the dataset folder is a trivial task, however, keeping the separate subfolders in separate tables with somehow

still maintaining normality is a problem still to be faced. My current approach is just to treat only the highest level of subfolders as separate/normalized tables.

3) **Python Library Standards** – Conventions pertinent to Python libraries are to be upheld within the tradeoff of standardization and internal structure of the modules. At the moment, the plan is to build a separate command line interface to interact with the library where the external calls are to be handled inside a single INTERFACE() class.
4) **Feature Selection** – The research part of the project partially revolved around elucidating the list of the features to be included in the library. The current assortment contains edges, corners, histograms, and color filters.

ANALYSIS ASPECTS

- **Metrics – t**he main Key Performance Indicator, for now, is chosen to be memory usage as Python allows us to easily keep track of it with the 'memory-profiles' library.
- **Feature Survey –** A comprehensive web/blog/social-media survey is in progress to sieve out the image-processing features that are necessary for the library at hand. The main factors are the time/resource requirements to retrieve the feature examined.
- **Possible Feature Selector Module –** In the course of the search for the most appropriate features it has been discovered that there is an abundance of feature engineering techniques to determine the dataset-specific importance of them. Currently, Recursive Feature Elimination and Sequential Feature Selection are being examined.
- **Scope of the library: A**scertaining the distinction between the datasets that would rather be transformed into relational schemas rather than NoSQL ones.