# **DATS-6103: Data Mining Final Project** **Group Proposal**

# **Popular Attraction/Landmark Recognition Using Google Landmark Dataset**

With advert increase in use of smartphones and other social apps, Image recognition, Image classification, Image processing are latest interest for data engineers in computer vision tasks. Major challenge with image classification is lack of annotated large dataset to train better and robust models.

**Problem Statement:**

Recognizing and training model to identify landmark is a challenging task as appearance of the landmark varies with geometry, illumination, different aspect ratio of the image presented. To overcome this issue, collection of images is used to capture typical appearance of the location. This project focus to build a model which recognizes given popular attraction/landmark using google landmark dataset. This landmark recognition model will be handy to identify name of the landmark in image. This will also helpful for photo organization in smartphones and fields like Aviation, maps, etc.

**Dataset:**

In order to capture typical appearance of image by collection of images, we need large annotated landmark dataset. Google has released latest landmark v2 dataset (sept 2019) which makes ideal choice for landmark recognition and retrieval purposes. This dataset includes over 5 million images with more than 2 lakhs diverse landmark classes. Google has also published this dataset in 3 sets – train, index and test, in which train and test files are used for landmark recognition and index file is used for retrieval purpose. Train dataset consists of image details of different landmarks, while test dataset consists of images that include no landmark, one landmark or more landmark. The major challenge to use this dataset is highly imbalanced training dataset, as number of categories is large, there many classes with single digit training data which makes its difficult to classify and train model for such classes.

Train dataset – 4132,914 location data with 203,094 unique classes

Test dataset -- 117,577 data

As dataset is highly imbalanced, certain data preprocessing need to be considered before training the model and dataset need to be cleaned to look for any broken url (analyze the image). Dataset is created by crowdsource the landmark available in internet. Each image might have different pixel size; hence these images need to resize to one uniform pixel size for analysis and training.

**Data Mining Algorithm & Network:**

HOG (Histogram Oriented Gradient) feature descriptor is used to extract feature from images and feature vectors are created. Further, we aim to use SVM (Support Vector Machine) classifier for this project. The dataset is labeled, high dimensional feature spaces are needed to handle, SVM usually produces best performances at these scenarios. With help of customized machine learning packages, we planned to code and train this dataset in python programming language, hence latest python version and any IDE is required to run code.

**Reference Materials:**

Announcing Google-Landmarks-v2: An Improved Dataset for Landmark Recognition & Retrieval (2019, September), Retrieved from: <https://ai.googleblog.com/2019/05/announcing-google-landmarks-v2-improved.html>

The Common Visual Data Foundation (2019, September), Google Landmarks Dataset v2, Retrieved from: https://www.kaggle.com/c/landmark-recognition-2019

Y. Li, D. J. Crandal and D. P. Huttenlocher, Landmark Classification in Large-scale Image Collections, Retrieved from: <http://www.cs.cornell.edu/~yuli/papers/landmark.pdf>

A. Crudge, W. Thomas and K. Zhu, Landmark Recognition Using Machine Learning, Retrieved from: http://cs229.stanford.edu/proj2014/Andrew%20Crudge,%20Will%20Thomas,%20Kaiyuan%20Zhu,%20Landmark%20Recognition%20Using%20Machine%20Learning.pdf

Y. Takeuchi, P. Gros, M. Hebert and K. Ikeuchi, Visual Learning for Landmark Recognition, Retrieved from: http://www.cs.cmu.edu/~takeuchi/iuw97/iuw97.html

**Performance Metrics:**

We aim to use most commonly used performance metrics used in image classification like accuracy score, confusion matrix, precision, sensitivity (Recall), specificity, AUC curve, F1 score, etc.

**Rough Schedule:**

We planned to complete this project in 5 weeks span (Dec 1, 2019). Below are rough split of events and time associated of it.

* Dataset Finalization – 1 week
* Understanding of dataset, Proposal Draft, Git set up, Action plan – 3 Days
* Data Preprocessing – 1 Week
* Data Training & Prediction using SVM – 1 Week
* Performance analysis – 5 Days
* Documentation (Group & individual Report, PPT) – 1 Week

Sharmin Version:

**Objective:** For our group project, we have taken up the concept of Landmark Recognition. Through this project, we aim to address this hot topic and possibly find out its usability in a more efficient prediction. This concept is quite popular with today’s social media. Hence, its impact is of a great magnitude.

To our advantage, we will be able to use the latest version of this dataset (September 2019).

**Dataset:** Google-Landmarks-v2

**Link:** <https://ai.googleblog.com/2019/05/announcing-google-landmarks-v2-improved.html>

**Possible Algorithm:** We aim to use the Support Vector Machine algorithm for our analysis and prediction. However, future work may also include comparing multiple algorithms and defining the most accurate one for prediction.

# talk about software and network

**Reference Materials:** During our initial discussion, we came across several reference papers and websites related to our topic. Hence, we will be learning concepts from them.

**Performance Metrics:** We will make use of the common performance metrics such as the Confusion Matrix, Precision Matrix, etc. # Can add specific ones as well

**Rough Schedule:**

* Deciding and discussing the dataset
* Plan of action (procedure, creating Git repository, etc.)
* Pre - processing and cleaning the dataset
* Training the model and performing prediction
* Calculating the accuracy and analyzing the performance
* Documentation and updating the GitHub Repository