Project Proposal

Dataset Selection

Our Dataset is called the Places365 it consists of 365 different classes and 5000 images for each class. After going through all the classes we selected 5 as per project use. They are - ***airfield, bus stand, canyon, market,*** *and* ***temple*** and each of these classes contains 5000 images, we may reduce the size if this leads to an unreasonable amount of training time. The size of our finalized dataset is 1.52 GB, it can be downloaded from here and the full place365 dataset can be downloaded from here.

Possible Methodology

1. **Data Collection**: Searched the internet and found a dataset with 5 places. Each place has its own characteristics and features. The model will be trained, validated, and tested with the dataset.

2. **Data Division:** The dataset will be split into three classes. One class is for training, named the training dataset, which contains about 60% of the entire dataset. The other is for validation, named the validating dataset, which contains 20% of the entire dataset. The last one is the testing dataset, which contains 20% of the entire dataset. To balance the data in the three datasets, it is ensured that each dataset contains the same portions of features.

3. **Model Building:** Tentatively, two models will be built for this project. One is the decision tree (DT), which is classified as a conventional machine learning method. The other is a convolutional neural network (CNN), which is a kind of deep learning method. It is a typical application of CNN in image classification. For DT, it serves as a baseline to evaluate the performance of CNN.

4. **Training and Testing:** The model will be trained with the training dataset. After training, the trained model will go through validation to see if the trained model can perform satisfactorily. Finally, the models will be tested using the testing dataset. For each step, the results will be evaluated in accordance with the evaluation matrix as described below.

5. **Model Evaluation:** After training and testing, the results will be evaluated based on the evaluation metrics for classification, including accuracy, precision, recall, F1 score, and area under the curve (AUC). The evaluation will be conducted for DT and CNN for comparison.

Bibliography

- Dataset - http://places2.csail.mit.edu/download-private.html

- Methodology - https://www.datascience-pm.com/crisp-dm-2/