Google Landmark recognition using Image analytics

Contributors: Vikas, Nitesh Sharma, Ekta Agnihotri, Chetan Kathooria, Siddhartha Khushu

Contents

1. Summary of Problem Statement:

1.1. Problem Statement

In this Project we need to predict the name labels of landmarks given in the Google Landmark recognition picture dataset using Image analytics. The challenge at our hand is to build models that classify the images provided in such a way that it matches the correct landmark with each unique image. This image analytics technology can predict landmark labels directly from image pixels, to help people better understand and organize their photo collections.

Similar projects take place every year such as the ILSVRC (ImageNet Large Scale Visual Recognition Challenge), but the biggest difference in this project is the larger number of classes.

In this Project as part of the solutioning, we plan to use image analytics technology for image recognition by building various model architectures including Deep Learning models using Computer Vision with CNN (Convolutional Neural Networks) that recognize the correct landmark (if any) in a dataset of challenging test images.

The project will encompass the state-of-the-art approaches in instance-level recognition across three domains: artworks, landmarks, and products.

1.2. Data

The dataset https://www.kaggle.com/c/landmark-recognition-2020/data on which we have worked is derived from the Google Landmark Recognition Challenge that took place on Kaggle a few months back.

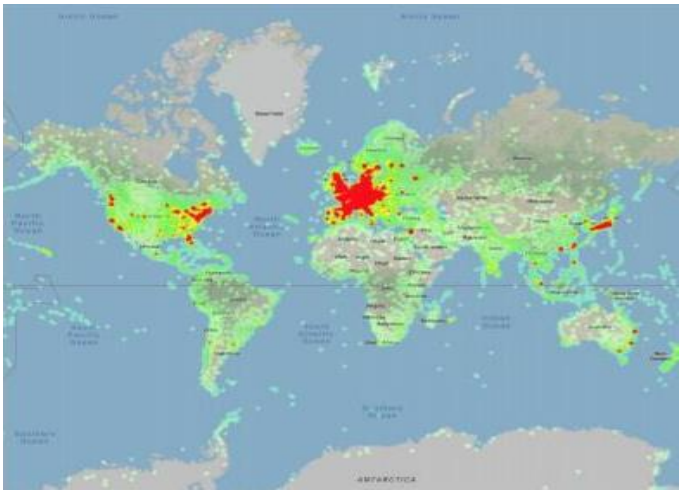


Figure-1: Heatmap of the places in the Google landmark datasets V2

The above dataset also poses several tough challenges. The landmark recognition training data originally contained over million images with around thousands of classes. The distribution of these classes is exceptionally long tailed, thereby creating an extreme class imbalance. The test set has a large proportion of out-of-domain images, underlining the need for low False-Positive (FP) recognition rates. The intra-class variability is particularly high, since images of the same class can include indoor and outdoor views, as well as images of indirect relevance to a class, such as paintings in a museum. To put things simply, this means that we would require a lot of computing power, coupled with a lot of time and patience. Like what was stated earlier, because there are so many classes, there are likely a lot of classes that contain only one image. This is a problem because with so few instances per class, our model's ability to accurately recognize images will be greatly hindered. To overcome this challenge, we will try using Image Data Augmentation techniques from Computer Vision. Also, downloading all the train and test images will take us about 48 hours (on a good bandwidth internet connection). Bottomline is that we need almost a week's time to just download and clean the images.

Analyzing these images on a system without GPU will also be taking a long time. Hence, we have decided to use Google Cloud Platform [5] for all our image classification and image retrieval models.

A similar list of landmark places is available country-wise in the Google landmark datasets V2 and can be viewed at the website - *https://storage.googleapis.com/gld-v2/web/index.html*

1.3. Findings

We started solving the problem through KNN and SVM algorithms. But, KNN was not able to work with available data set. After KNN algorithm, we tried to solve the problem with SVM, but the model accuracy obtained was dismally low (0.24%).

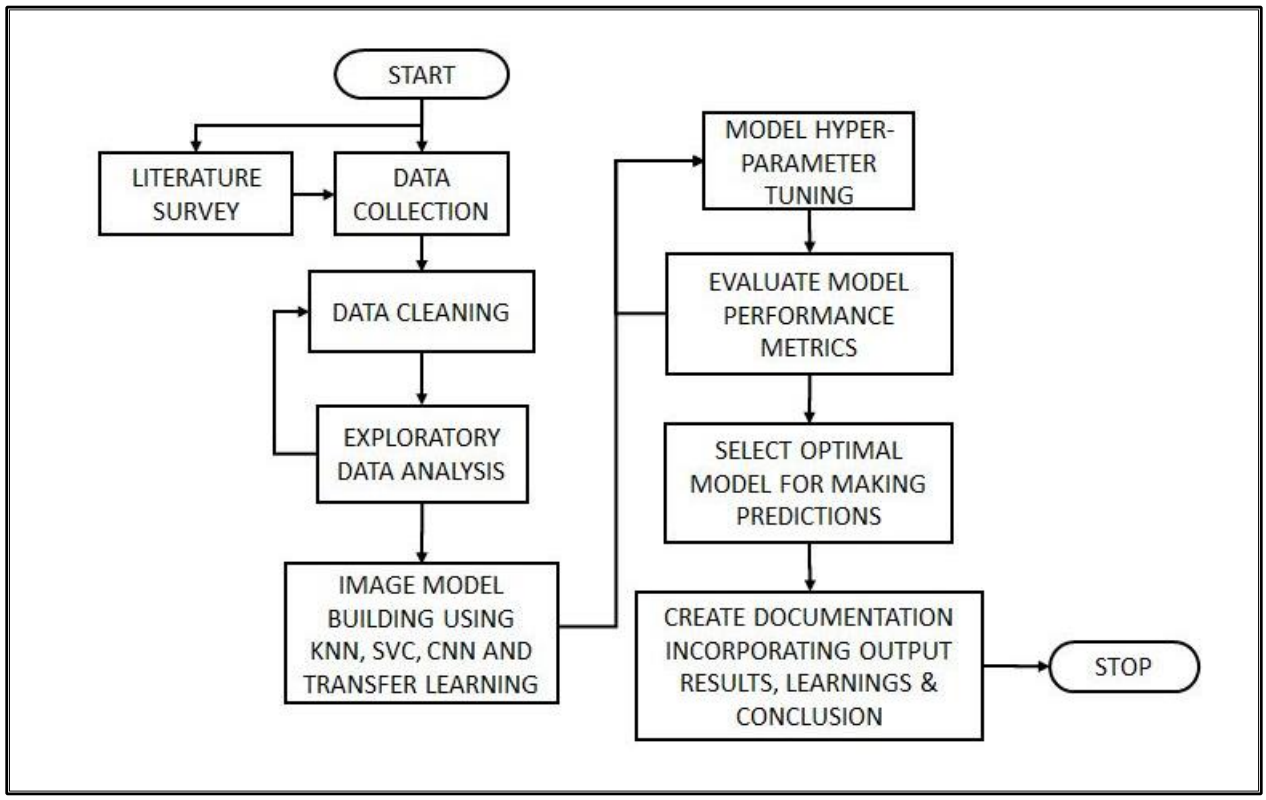
Later, we build CNN with 20 layers of convolution but again model accuracy scores obtained were still exceptionally low (around 3%).

So, finally we used CNN with Transfer Learning, using pre-trained models like ResNet-50 & EfficientNet-B2. It was found that the Model accuracies improved by a good margin when we started using Transfer Learning (with data augmentation) to predict and recognize the landmarks. When 200 classes were taken for the prediction study, the ResNet-50 model enabled getting a Training accuracy of 38% and Test Accuracy of 43%, whereas the EfficientNet-B2 model enabled getting a Training accuracy of 88% and Test Accuracy of 74%.

2. Overview of Final Process

2.1. Project execution workflow

In this study we are planning to undertake the following steps for executing the Project for prediction of name labels of landmarks given in the Google Landmark recognition picture dataset through Image analytics-based algorithms using Supervised Learning methods:



2.2. Exploratory Data Analysis (EDA)

There are 15,80,470 Train images & 10,345 Test images in the complete GLDV2 dataset.

- Total Number of Train, Test and Total Number of Images

```
The number of train images is : (1580470, 2)
The number of test images is : (10345, 2)
The total number of images is : 1590815
```

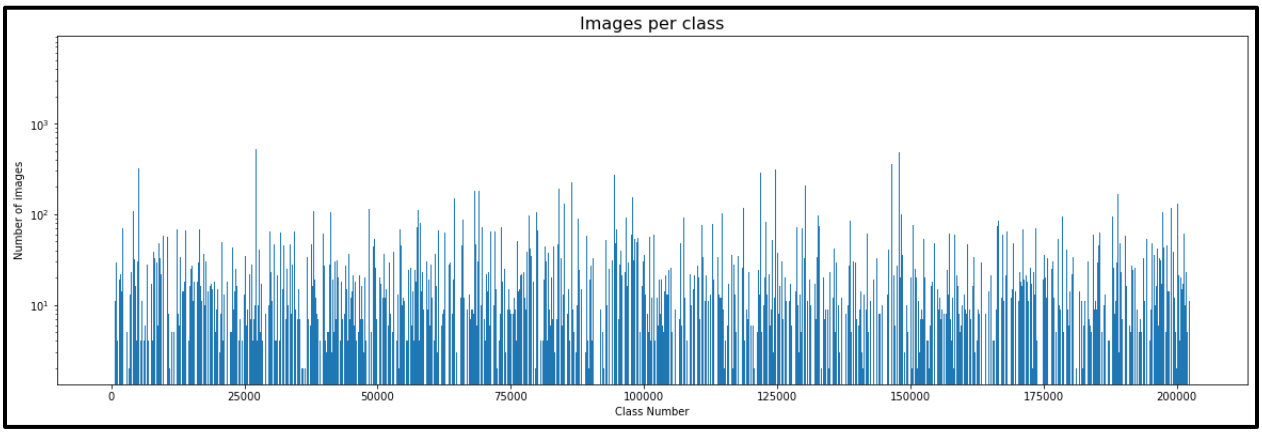
- Number of Images per landmark ID

```
landmark_id
183721       2
106110       2
193443       2
27475        2
116406       2
            ...
113209    1135
83144     1741
20409     1758
126637    2231
138982    6272
Name: id, Length: 81313, dtype: int64
```

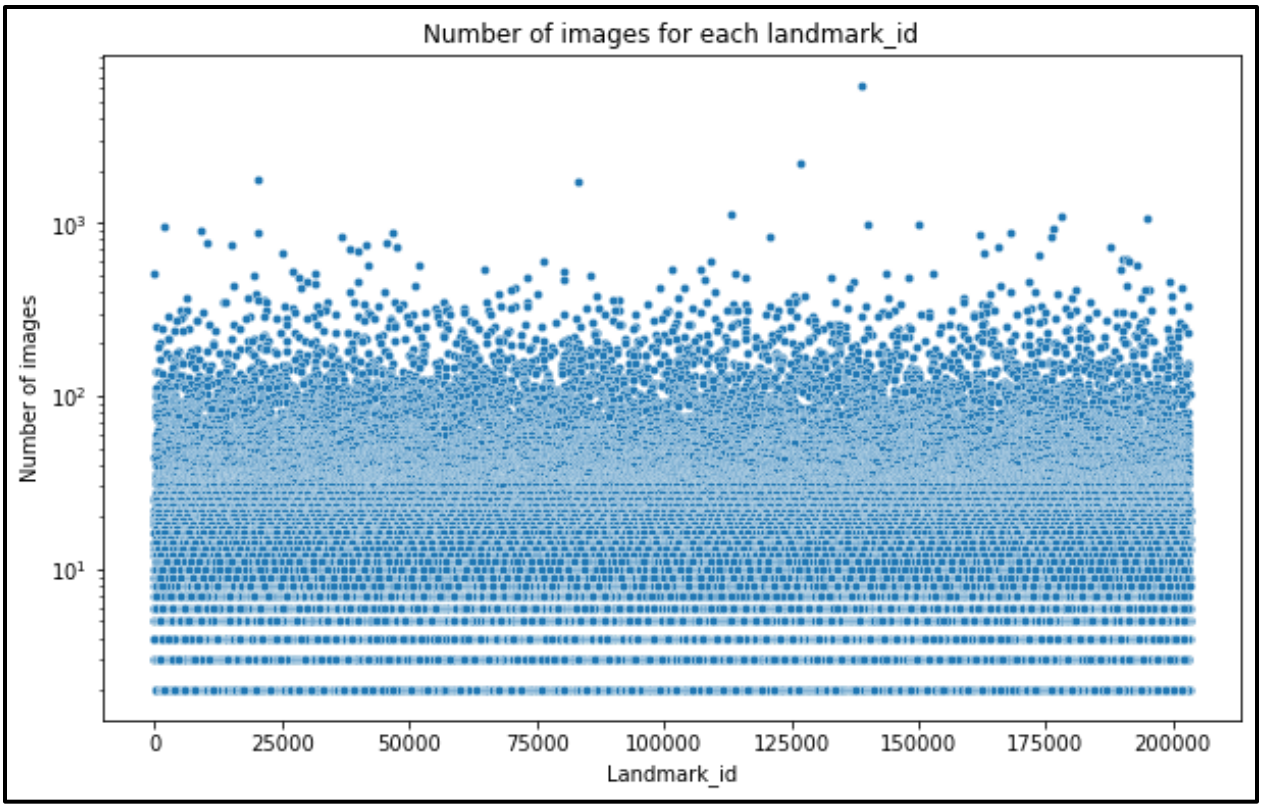
- There are 81,313 classes corresponding to 15,80,470 Train images in the complete GLDV2 dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 81313 entries, 0 to 81312
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   landmark_id  81313 non-null  int64
 1   frequency    81313 non-null  int64
dtypes: int64(2)
memory usage: 1.2 MB
```

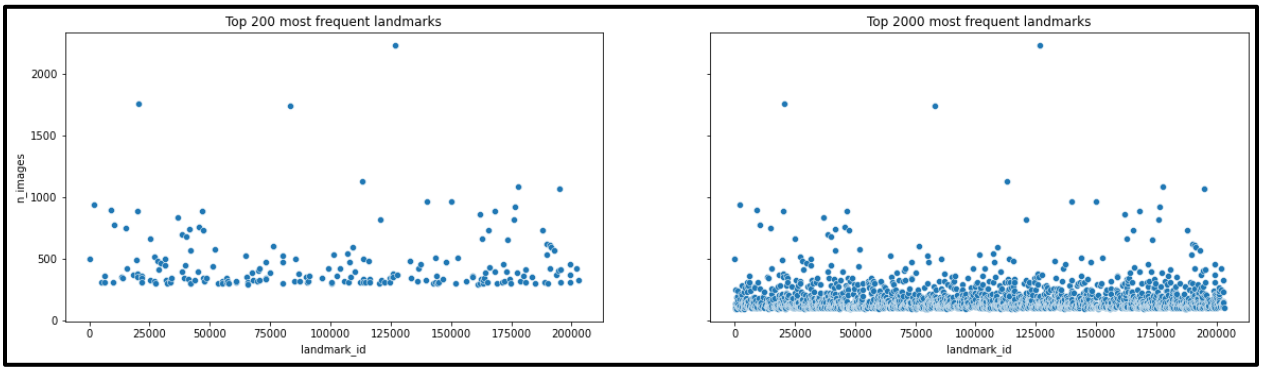
- Images Per Class



No. of images per Class ID - 81,313 classes that correspond to more than 2 Lac. Class ID's Numbers

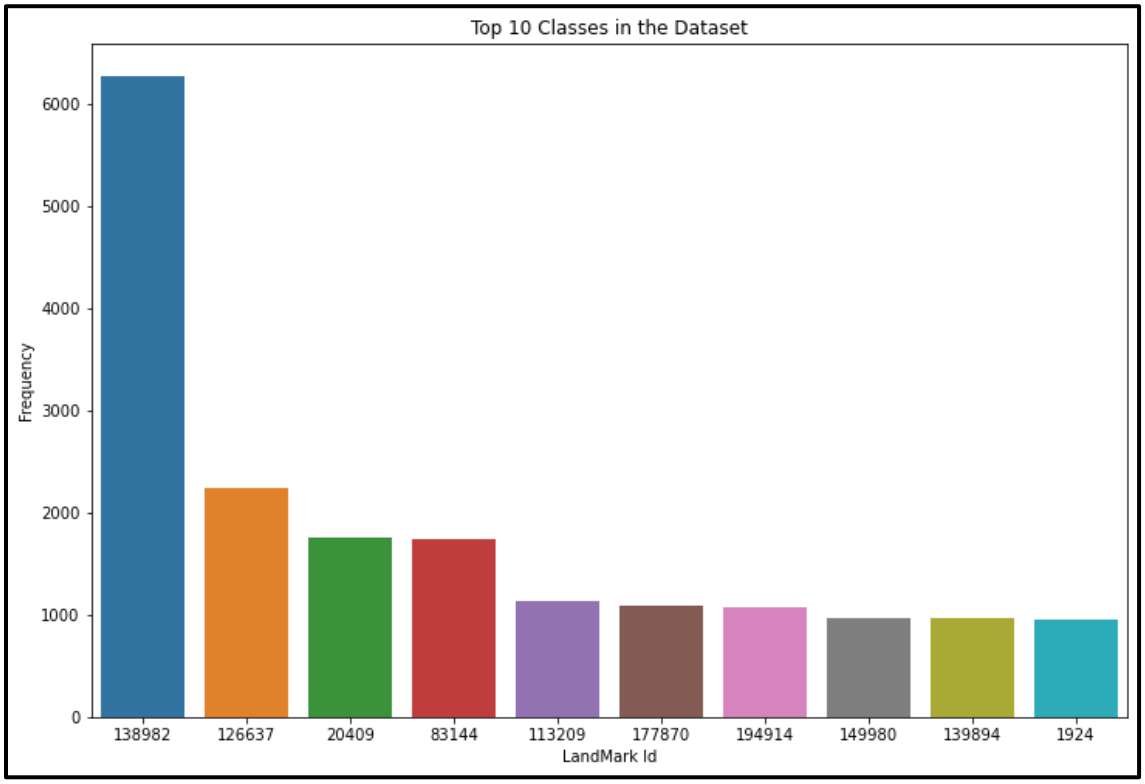


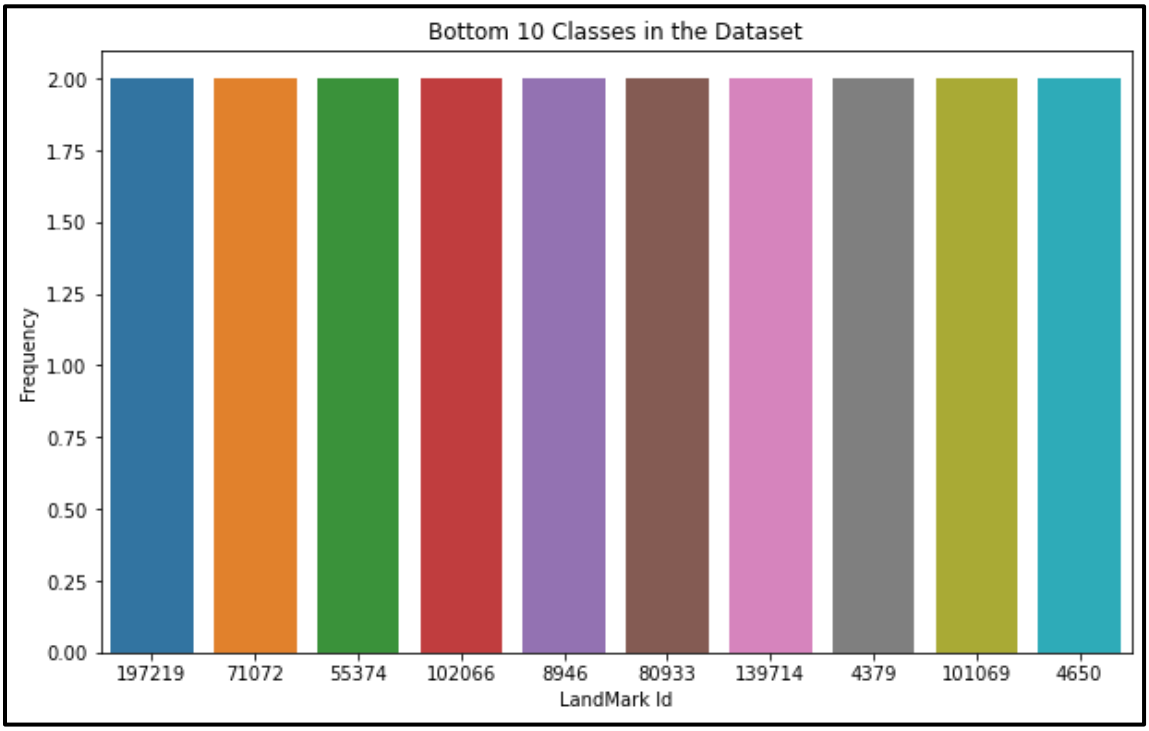
No. of images per Landmark ID in the complete GLDV2 dataset



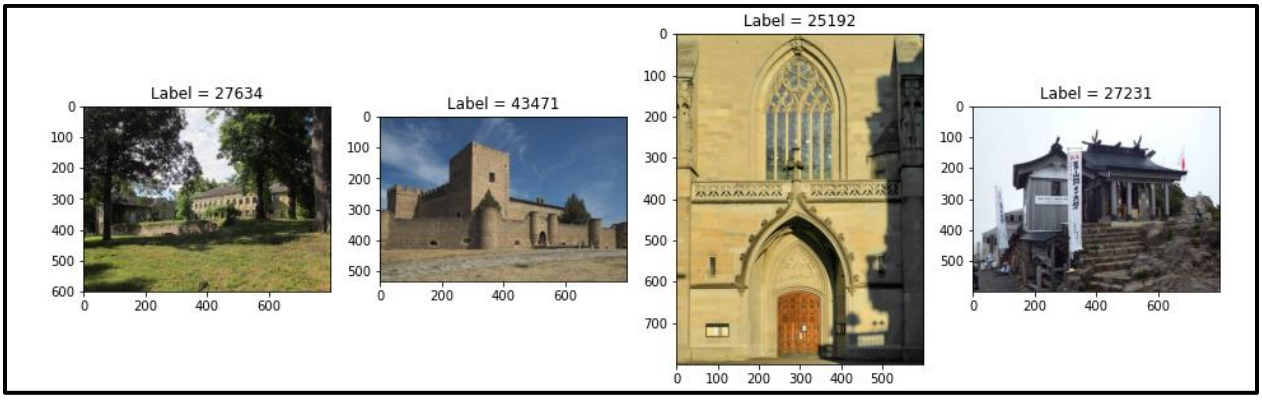
No. of images per Landmark ID in Top 200 most frequent & Top 2000 most frequent landmarks

It was found that the Maximum No. of images in a Landmark-ID (for a class) is 6,272 and the minimum no. of images corresponding to many Landmarks ID's (across various classes) is just 2 (as shown in the Frequency Histograms given below), hence exhibiting a class imbalance in the GLDV2 dataset.





Landmark Images with corresponding Labels retrieved randomly from Train Image directory using CV2



Landmark ID density distribution for Top 1000 Classes

	id	landmark_id	filename	label
0	0257e00b3b740c9e	49712	0/2/5/0257e00b3b740c9e.jpg	49712
1	59738410e2446d85	42016	5/9/7/59738410e2446d85.jpg	42016
2	e4a0bb227ac909ed	7666	e/4/a/e4a0bb227ac909ed.jpg	7666
3	960ad080644b1137	137203	9/6/0/960ad080644b1137.jpg	137203
4	4e7319e99bde4079	168106	4/e/7/4e7319e99bde4079.jpg	168106
...	...	...	...	...
251743	01935231c9f0f80f	176899	0/1/9/01935231c9f0f80f.jpg	176899
251744	d97c18d69c651b86	20409	d/9/7/d97c18d69c651b86.jpg	20409
251745	e9c09cf7de22bd16	169413	e/9/c/e9c09cf7de22bd16.jpg	169413
251746	f96ae7e492f744a1	164103	f/9/6/f96ae7e492f744a1.jpg	164103
251747	d33308ee99501d46	154243	d/3/3/d33308ee99501d46.jpg	154243

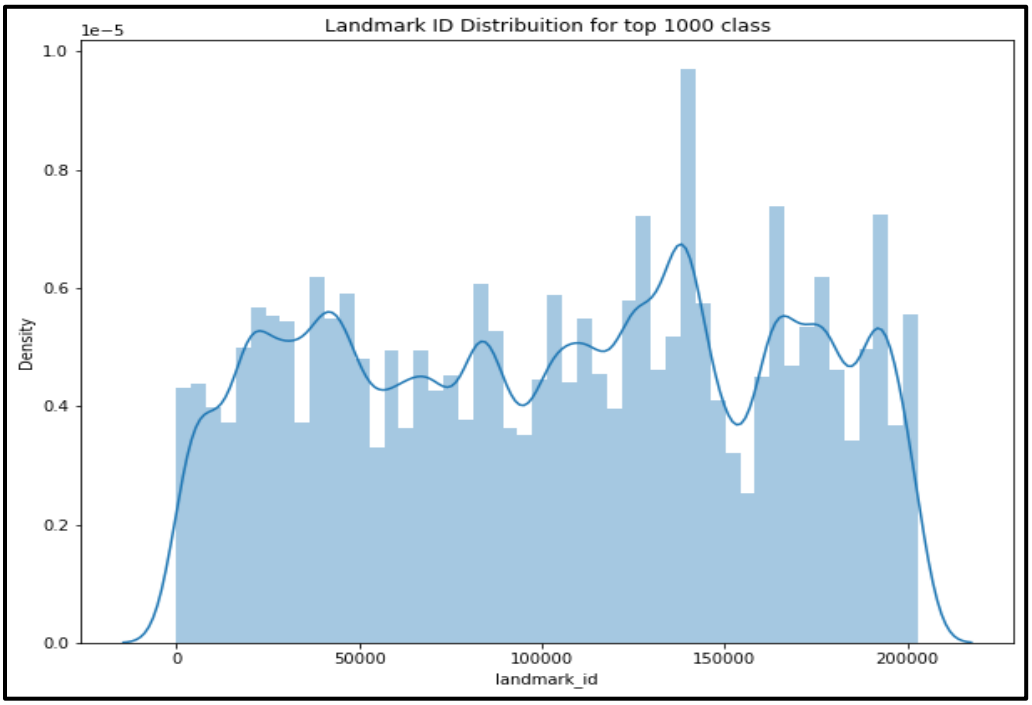
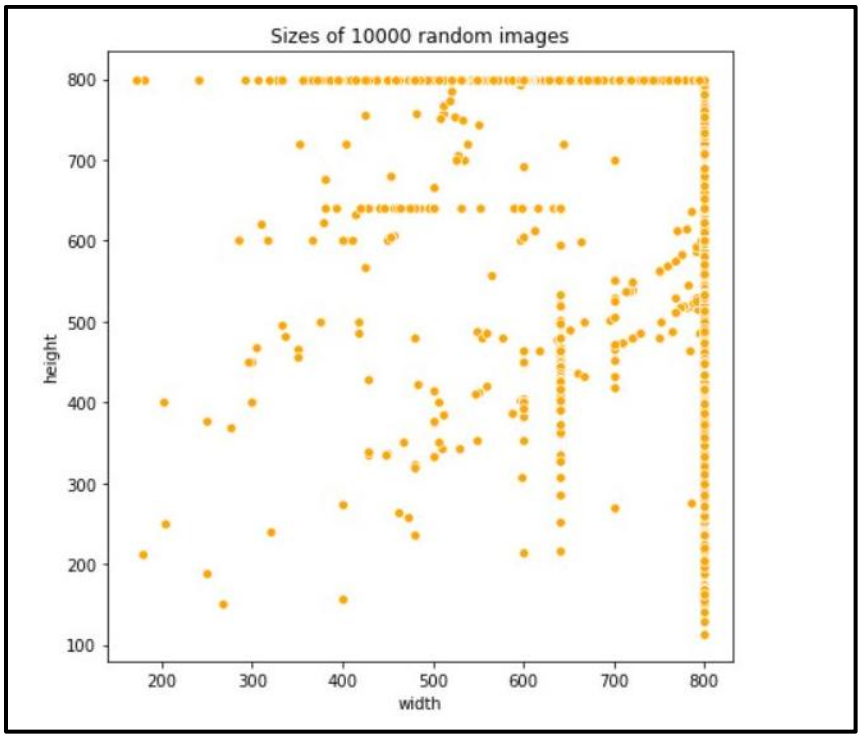
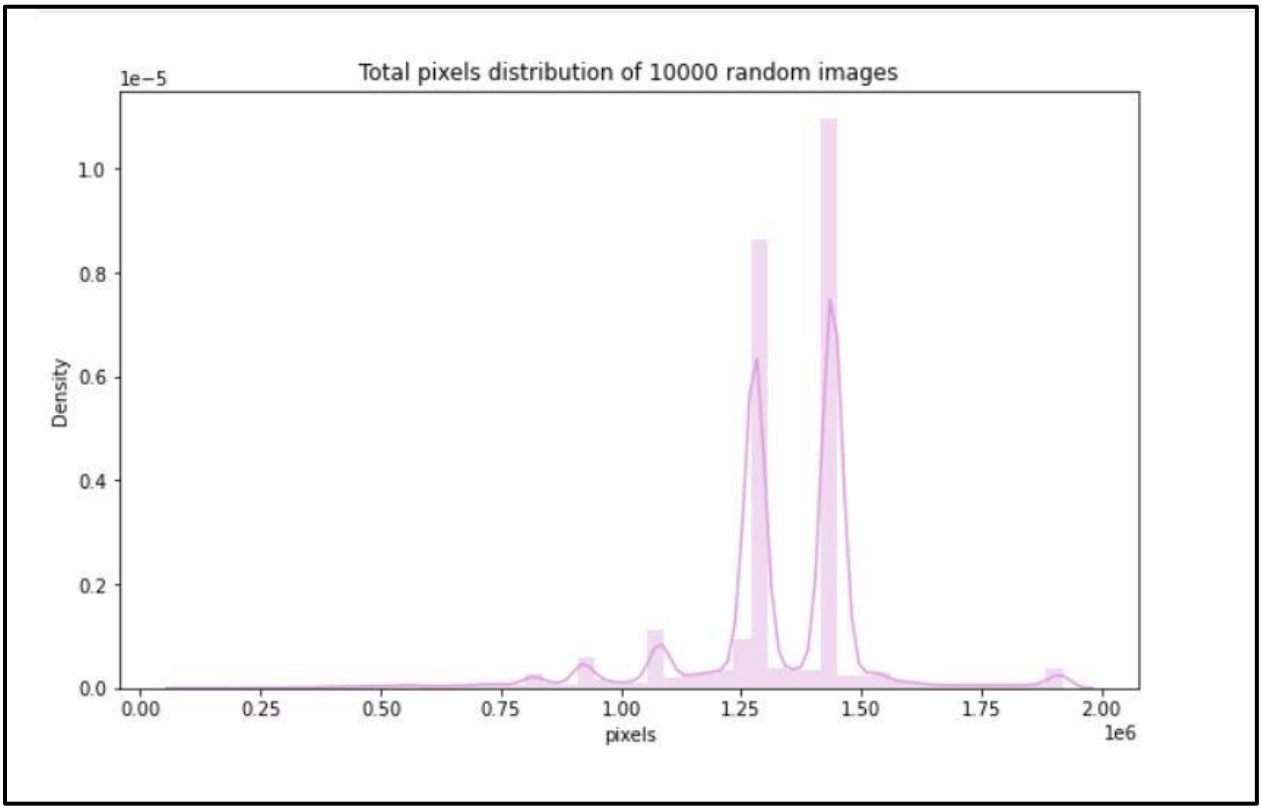


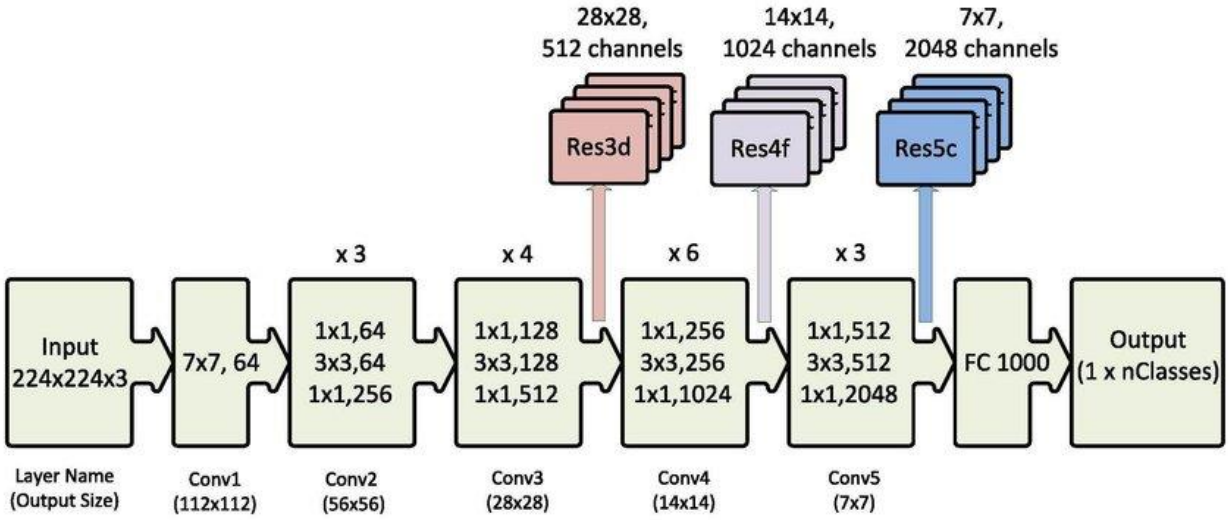
Image Characteristics for 10000 random images



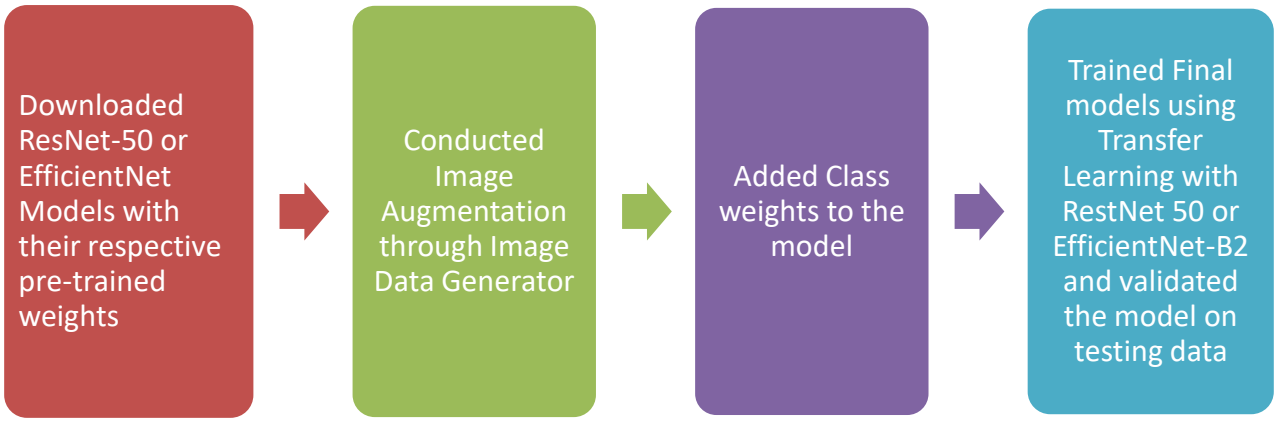


2.3. Base model & Architecture:

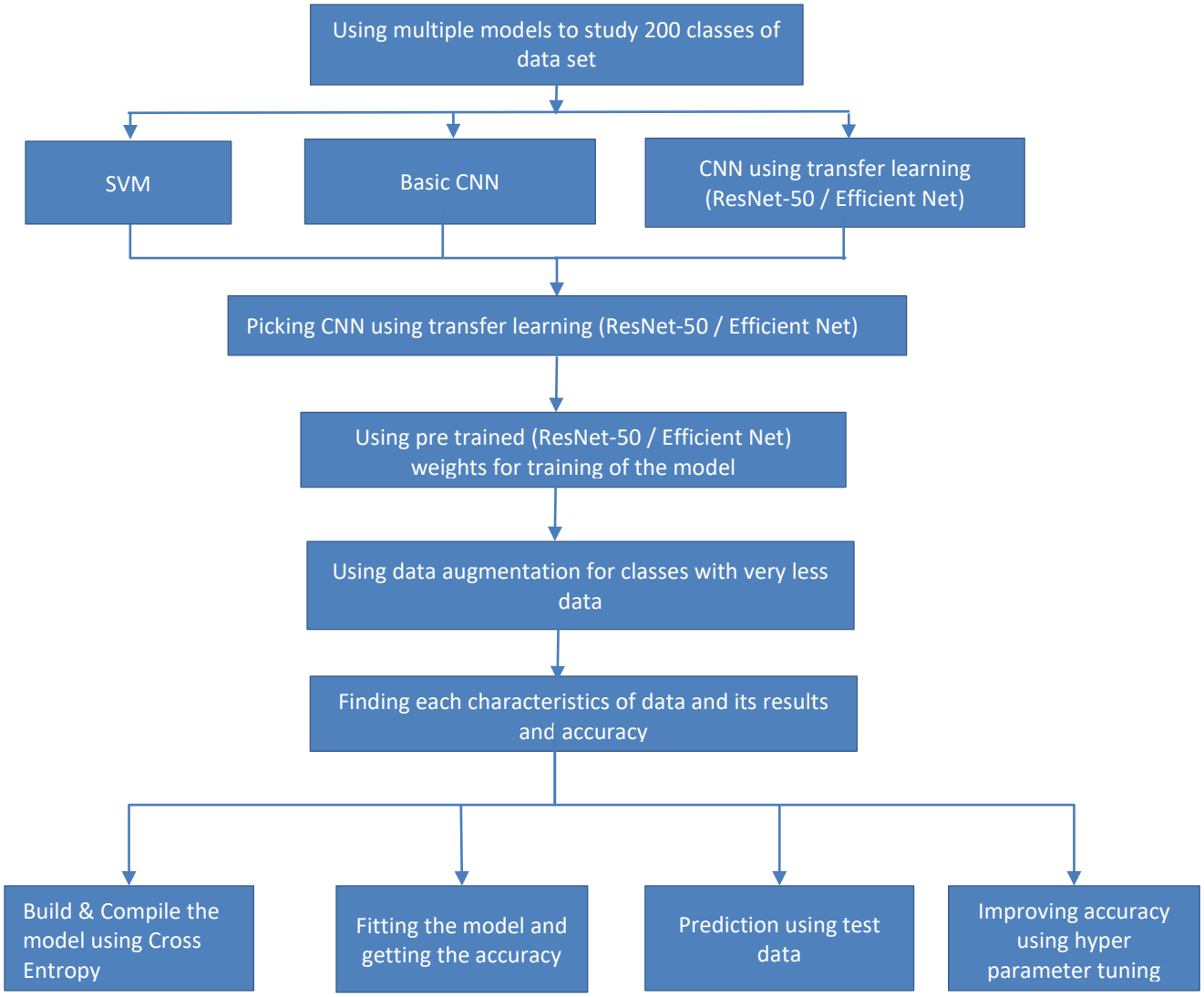
ResNet-50 is a convolutional neural network that is 50 layers deep. We can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224



2.4. Combined Techniques Flowchart



3. Walkthrough of the solution



4. Model evaluation

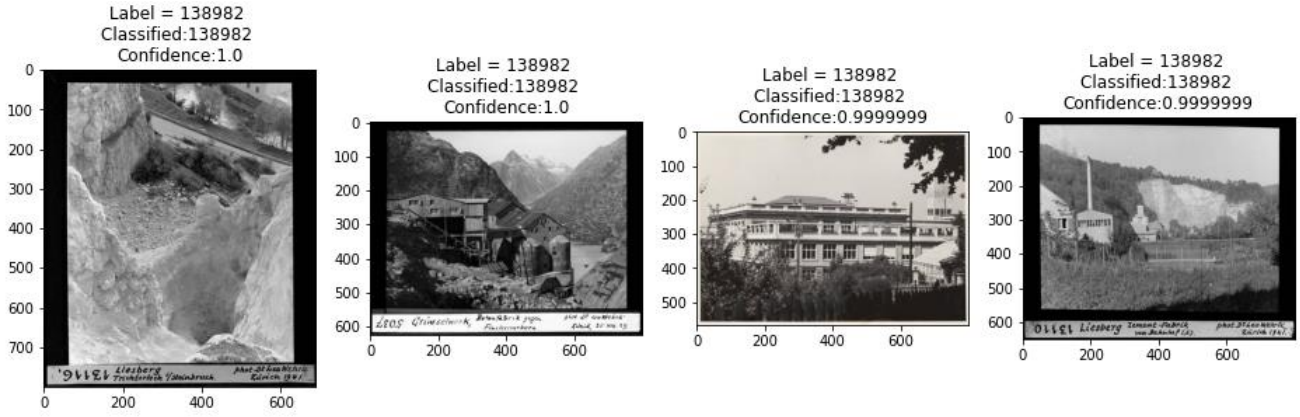
- The main objective of model is to recognize landmarks in an image in real-time.
- The CNN model is built using transfer learning of pre trained ResNet-50 & EfficientNet-B2 model.
- The pre-trained model was trained on 98,440 landmark images and the new weights were obtained.
- The model accuracy obtained improved upon decreasing the number of classes chosen for model evaluation.
- The pre-trained Models performed much better than the conventional CNN Models or Machine Learning models like KNN & SVM.

Results of various Models used for prediction (taking different number of classes)

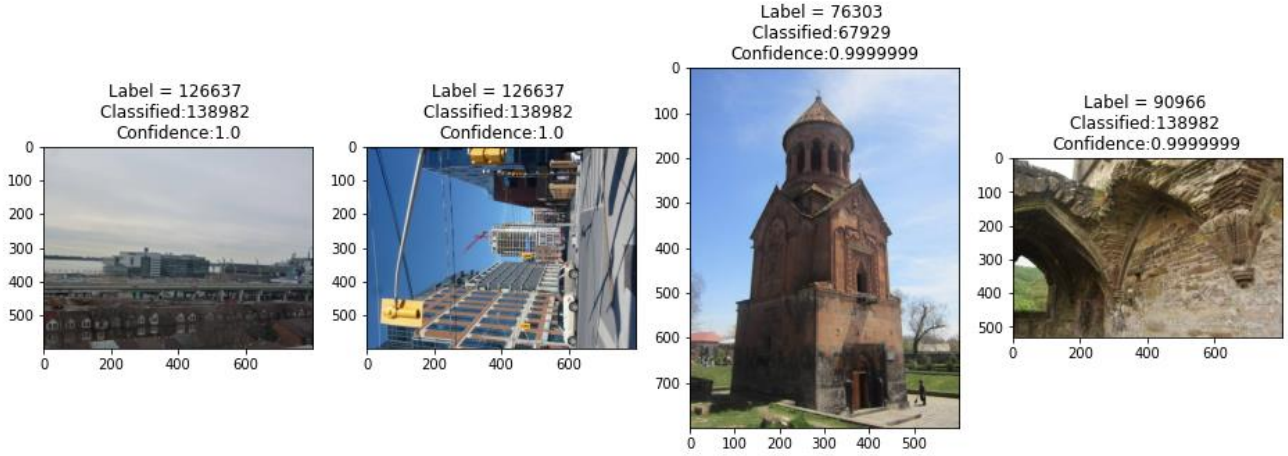
No of Classes Taken	Total No of Images	Algorithm Used	No of Epochs	Batch Size	Training Accuracy (%)	Testing Accuracy (%)
1000	2,51,748	Basic CNN	5	32	3.1	2.9
1000	2,51,748	CNN with ResNet-50	32	16	15.5	16.5
500	1,69,499	CNN with ResNet-50	40	16	22.6	23.6
200	98,440	CNN with ResNet-50	60	64	38.2	43.2
100	66,081	CNN with ResNet-50	100	32	47.0	53.2
50	45,077	CNN with ResNet-50	60	32	50	58
200	98,440	CNN with EfficientNet-B2	40	32	88	74.5
50	45,077	CNN with EfficientNet-B2	40	32	90	77.6

Predictions using ResNet-50 for 200 classes:

Good Predictions:

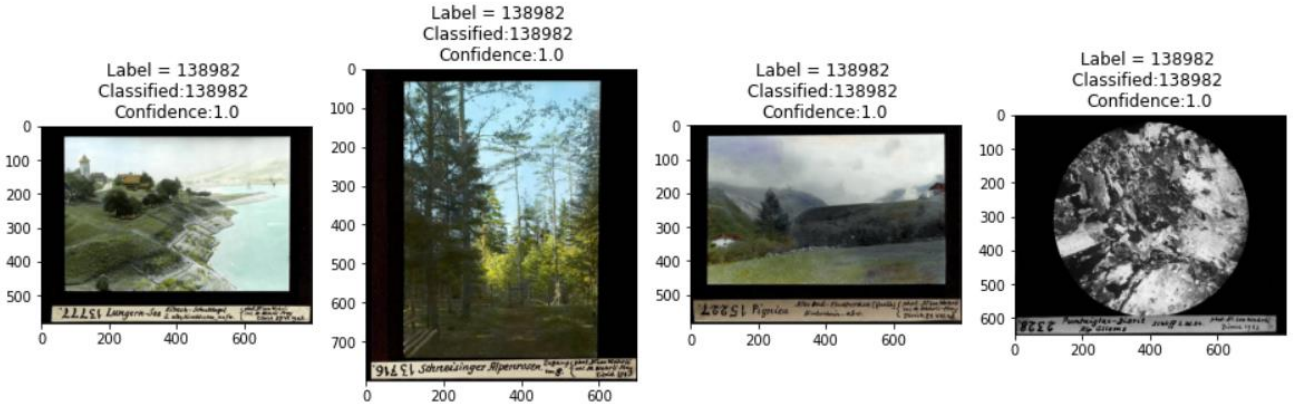


Bad Predictions:

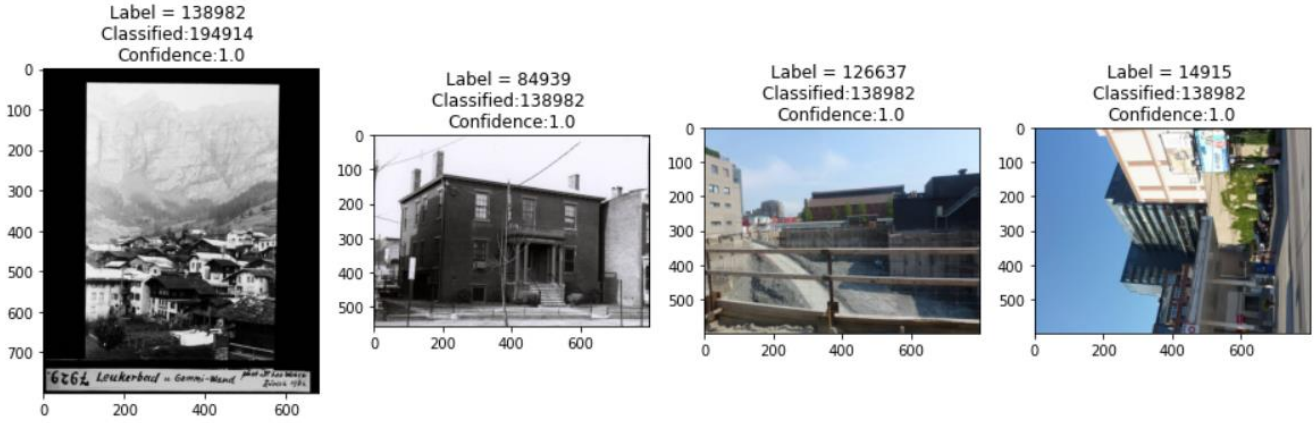


Predictions using EfficientNet-B2 for 200 classes:

Good Predictions:



Bad Predictions:

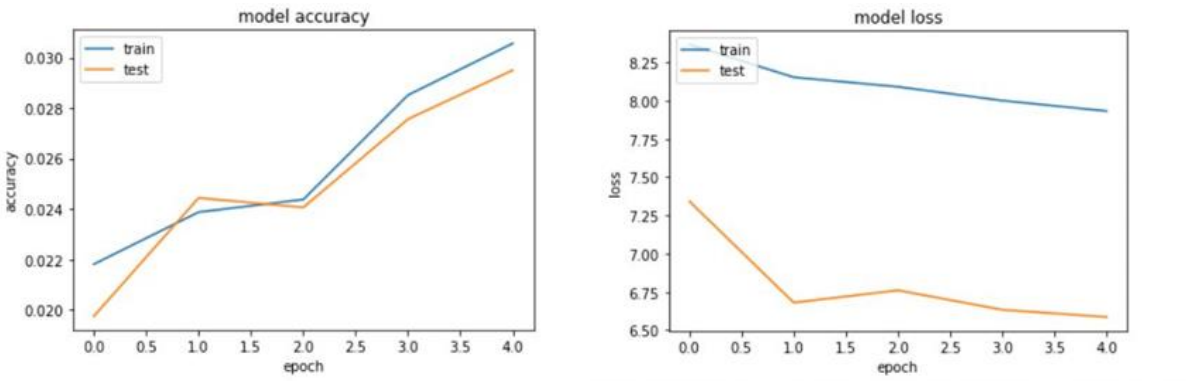


5. Comparison to benchmark

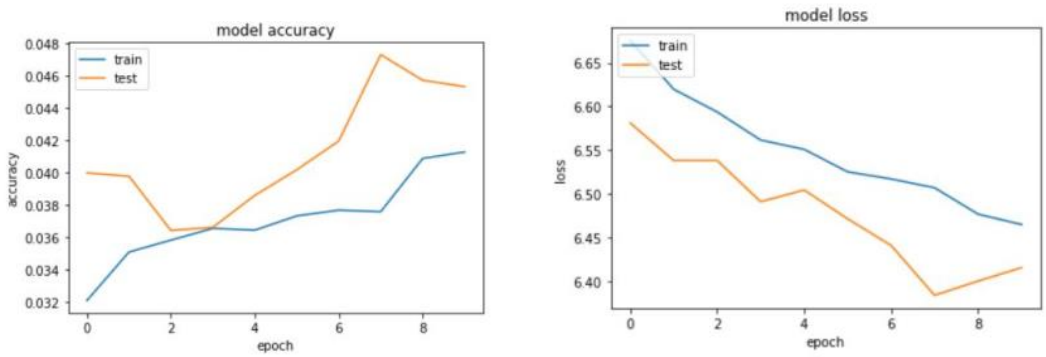
- We had chosen 80% model accuracy for 200 classes as our benchmark for the project.
- The final test accuracy obtained by us for 200 classes using pre-trained model ResNet-50 is 43% and using pre-trained model EfficientNet-B2 is 74.5%.
- We are closer to our benchmark using pre-trained model EfficientNet-B2.
- We were not able to meet the benchmark because of the limitations that have been explained thoroughly in limitations section below in the document.
- With respect to our chosen benchmark, we can also perform fairly good landmark detection and recognition.

6. Visualization(s)

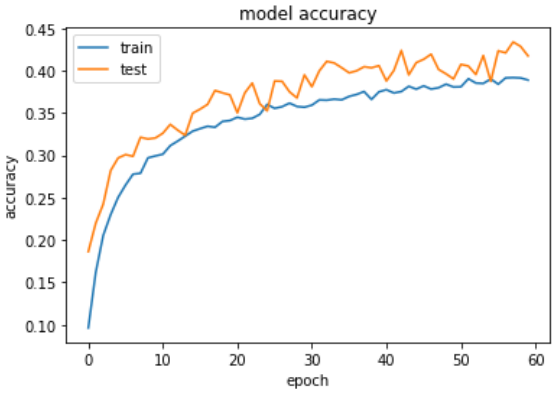
Model Accuracy and Loss for 1000 classes of Google Landmark Dataset using Basic CNN

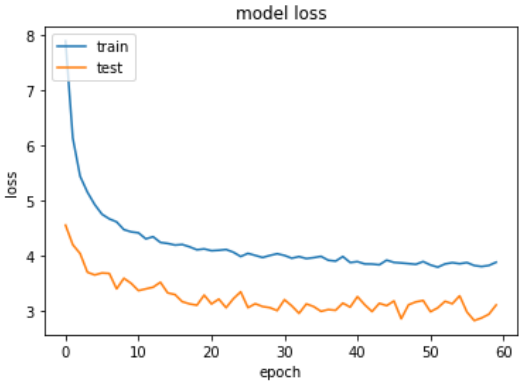


Model Accuracy and Loss for 1000 classes of Google Landmark Dataset using ResNet-50

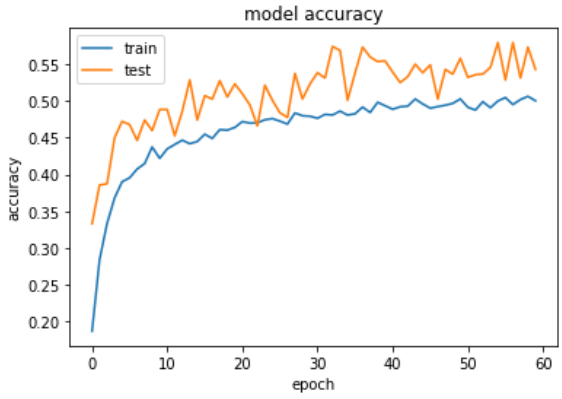


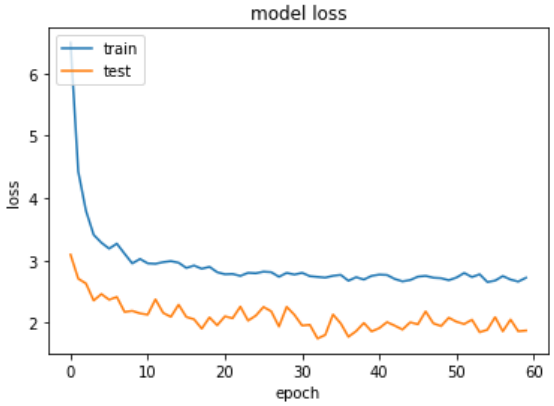
Model Accuracy and Loss for 200 classes of Google Landmark Dataset using ResNet-50



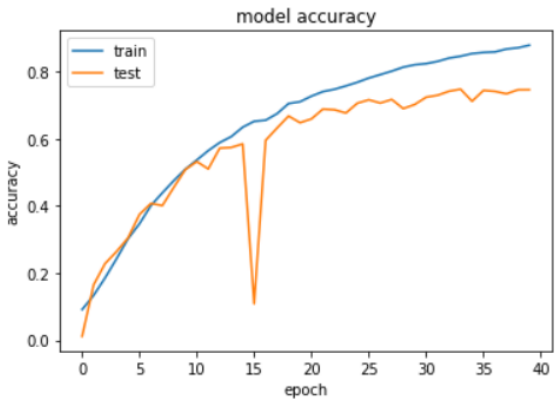


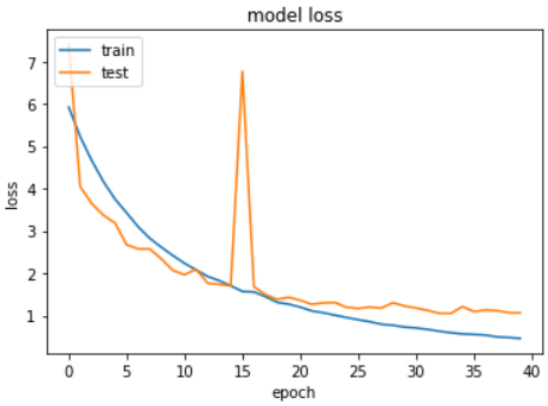
Model Accuracy and Loss for 50 classes of Google Landmark Dataset using ResNet-50



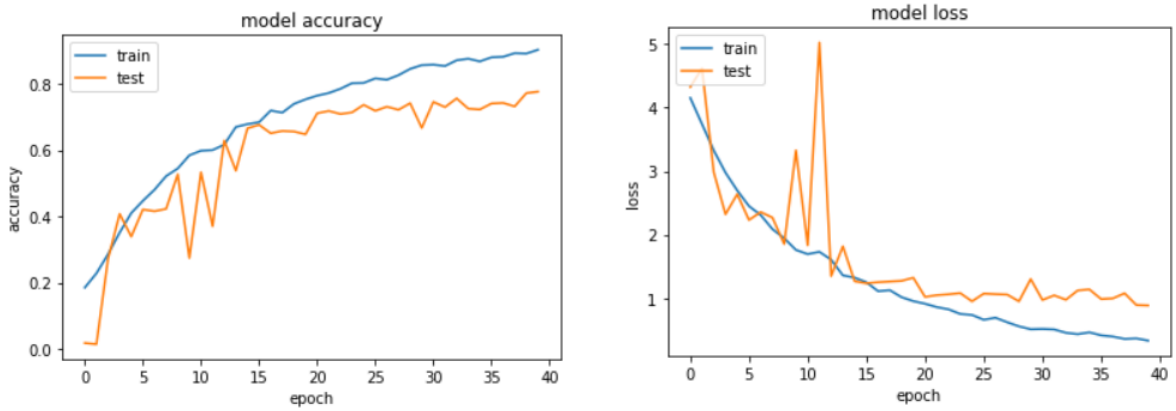


Model Accuracy and Loss for 200 classes of Google Landmark Dataset using EfficientNet-B2





Model Accuracy and Loss for 50 classes of Google Landmark Dataset using EfficientNet-B2



7. Implications

- Landmark Recognition solution can be implemented in image applications such as google photos. This technology can predict landmark labels directly from image pixels, to help people better understand and organize their photo collections.
- It can also be used for educational purposes. Teachers can show the images with landmark recognized to students that will enable their learning faster.
- It can be used in standalone cameras or inbuilt cameras on mobile phones to recognize the landmarks for which the pictures have been taken by the photographer.

8. Limitations

- The model accuracy of our solution is low because we are using 200 classes out of 81,313 classes in the data set.
- The distribution of spread of number of images per class is varying to a good extent thereby reducing the model accuracy.
- We need a huge IT infrastructure for processing this large dataset comprising of a total 15,90,815 images with 81,313 classes.
- We are running the solution on Kaggle directly whereby there is a time limitation of 9 hours for running the solution notebook using GPU processing power.

9. Closing Reflections

- We learnt the nature of Computer Vision problems and the solutioning approach to address the same problem.
- We have learnt how to tackle an imbalanced data set in computer vision problems.
- We can improve the model accuracy using Advanced Computer Vision with YOLO v3 pre trained model.