A Hybrid Approach for Landmark Recognition using Deep Local Features and Residual Network-50

Nishant Nimbare, Parth Shah, Shail Shah, Ramchandra Mangrulkar*

Computer Engineering Department, Dwarkadas J Sanghvi College of Engineering, Mumbai, India.

Abstract. As smartphones and mobile data become universal in modern society, the opportunities to interact with the real world would grow tremendously. Latest Technologies such as Oculus Rift and Google Glass attempt to bridge the gap between the virtual and the material. With advancements in computing speed and image recognition, the idea of augmented reality (AR) becomes more tangible. However, the sheer complexity of image processing and feature recognition is an area of concern for AR. A successful AR system must distinguish among many landmarks and identify or classify the existence of new landmarks. AR algorithms naturally lend themselves to using deep learning because of the adaptability required to various factors. This paper aims to develop and refine a deep learning algorithm that can distinguish landmarks from images using a google landmark database of known landmarks. Instance-level recognition is universally used in areas of Landmark recognition and is also the upcoming research area. Instance-level recognition is the brain behind Landmark recognition. As in Landmarks, the goal is to seek an instance of a common group instead of a group, requiring new deep learning techniques. In this paper, three different VGG16, Inceptionv3, and ResNet50 models are trained using the transfer learning technique and a Pure Convolutional Neural Network (CNN) model is also trained from scratch. This paper proposes a modified version of the ResNet50 model to increase the accuracy and performance of the models used. The revised version of Resnet50 contains an additional Deep Local Features (DeLF) processing layer before generating the final output.

Keywords. Landmark Recognition, DeLF, ResNet50, Transfer learning, CNN, Inceptionv3, VGG16, Instance-Level Recognition

1 Introduction

Deep learning approaches have outperformed previous advanced machine learning strategies in many areas, including video, audio, imaging, medical, social, and sensor. Object recognition has gained tremendous interest by engineers and scientists in artificial intelligence and computer vision. Deep learning enables models with multiple computing layers to learn and represent data at multiple levels of abstraction, simulating how the brain perceives and understands multimodal information and indirectly capturing complex constructs of large-scale data. Among various methods developed in object recognition, Convolutional Neural Network (CNN) is of interest for this project. CNNs (Convolutional Neural Networks) are influenced by the structure of the human visual system and hence outperform conventional machine learning approaches in computer vision and pattern recognition. CNNs are used in a variety of applications, including self-driving cars, segmentation, object identification, facial recognition, and many more. [1].

This paper presents a deep learning model based on CNN. A technology that can accurately predict landmark labels directly from image pixels can broadly benefit applications in various areas such as photo management, maps, aviation, and satellite images.

Image processing and classification is a fascinating and interesting problem. This branch of Computer Science is developing exponentially, given the industry's current rise of computational power. As Image Processing has complex matrix-based calculations, the time to train the model depends on processing power. The availability of storage to masses has made image processing fast and inexpensive [2].

Anything from mobility pattern to stop signs in a self-driving car to cancers in a brain cell can be predicted using image recognition. Convolutional Neural Nets (CNNs) are one of the most widely used image recognition models. This paper presents one such Image Processing application of the real-world, Landmark Recognition.People often fly to foreign countries or abstract places, take several photographs, and then discover a month later that they have no idea what they were standing in front of. People would be able to remember exactly what it was that made the picture so memorable if they are able to train models that can identify landmarks. Analysing popular landmarks is also appealing because it is one of the basic building blocks for making AI smarter. Recognizing landmarks in sequences of images is a challenging problem for several

* Corresponding author: shail.shah822@gmail.com

reasons. To begin with, the appearance of any particular landmark varies significantly from one observation to the next. In addition to variation due to various aspects, illumination change, external clutter, hardware of the device capturing and changing geometry of the imaging devices are other factors affecting the variability of the observed landmarks. Due to given problems, normal object detection is not applied for Landmark Recognition, instance-level learning is used for such types of problems [3].

Instance-level recognition (ILR) is a method in image recognition or computer vision, where the goal is to find a specific instance of an object. For example, instead of labeling an image as "water-fall" it identifies the image as Niagara waterfall. ILR distinguishes the instance of the object class from the object class. There are many problem statements that require to find a specific instance from a collection, like artwork, landmarks, products and can also be useful in visual searching, shopping, copyrighting etc. Instance-level recognition will unravel the true potential of deep learning technologies for semantic image classification/retrieval for eCommerce, travel, media & entertainment, agriculture, etc. [4].

2 Literature Review

Authors Asia Kausar, Mohsin Sharif, Jinhyuck Park, Dong Ryeol Shin, published a paper with title "Pure-CNN: A Framework for Fruit Images Classification" have implemented a system to classify fruits based on PCNN. Unlike a traditional CNN where the last few layers are fully connected layers, they used a Global Average Pooling (GAP) after Convolutional and Pooling layers, thus using a Pure-CNN (PCNN). GAP is used to reduce the number of parameters and prevent the model from overfitting. GAP behaves like a Max pooling layer but performs more reduction compared to Max pooling. Finally, the GAP layer was attached to the last SoftMax layer for prediction. ReLU was used as an activation function, and the whole architecture was 7-layers deep. They were able to achieve 98.8% accuracy. Authors make a case that fully connected (FC) dense layers at the end of a usual CNN model are computationally expensive and tend to overfit quickly. Also, layers of CNN behave as object detectors, but this ability vanishes on using FC layers for classification. There are two suggested solutions for this. First is the use of dropout layers, wherein dropout layers are introduced. In the Dropout concept, some connections between the layers are randomly dropped off during training. The second way to reduced overfitting due to FC layers is to replace them with a Global average pooling (GAP) layer. Paper compares a traditional CNN model, with and without dropout layers, to a pure-CNN (PCNN) model, consisting only of convolution layers and a GAP layer on top. The pure-CNN model was found to better at classifying fruit images. Dataset used was Kaggle's fruit-360 dataset with 81 classes of fruits. The model was trained over 40,000 iterations with a batch of 50 images. Images were RGB 100 x 100. Global Average Pooling (GAP) was used to minimize the number of parameters and protect the model from overfitting. It reduces filter size by merely averaging the entire function map. GAP, although similar to the max-pooling layer, performs a more drastic method of dimension reduction. Here dimension [h, w, d] is reduced to dimension size [1, 1, d].

Authors Andrei Boiarov, Eduard Tyantov, in published paper titled "Large Scale Landmark Recognition via Deep Metric Learning", have stated metric-based learning of the model to train on the dataset of the Landmarks. Since there isn't a lot of data about a certain landmark, the model ought to have a high accuracy and a low false-positive score. In this paper, the authors have used a deep neural network to train the model. This paper proposes an algorithm to be used while training the model. It also mentions the cleaning of the dataset required in order to apply the metric learning. The paper has also provided overall architecture, learning process, and feature extraction by measured approach. The authors used a Convolutional Neural Network approach to train their landmark recognition model. They divided the CNN into three parts – the main network, the embedding layers, and the classification layer. They trained their model to extract as many characteristics as possible using the fine-tuning method, which is widely used in computer vision. As Landmarks and Scenes are similar with respect to the number of features thus, the authors first trained their model on the scene's dataset. The authors also removed the last fully connected layer and instead added a fully connected layer along with batch normalization. It is also a fast method and has been tested with other methods. The method is like the state-of-the-art method but is scalable and is faster in deployment. The concept has been applied and scaled up for use in image recognition on user images.

Authors Ilja Kuzborskij; Fabio Maria Carlucci; Barbara Caputo, in published paper titled "When Naïve Bayes Nearest Neighbors Meet Convolutional Neural Networks", mentions that since it cannot use CNN activations as input, NBNN has lost traction since introduction of CNN, and it cannot be used as the final layer of CNN architecture. NBNN cannot handle huge datasets as it is not scalable. In paper [7], the authors introduce a structure that solves these issues while also reintroducing NBNN into the scene. Firstly, CNN activation is extracted from local patches at multiple scaling levels. Then, "a scalable version of Naïve Bayes Non-Linear Learning (NBNL)" [7] is addressed, which exploits the learning power of local SVM. This paper provides a method to use CNN activation features and NBNN-based classifiers together. The key elements are:

1. Obtaining CNN activations from local patches at different scales

2. a scalable NBNN based algorithm that makes use of the learning power of locally linear SVMs [7].

Authors: Mahbub Hussain, Jordan Bird, and Diego Faria, in their published paper titled "A Study on CNN Transfer Learning for Image Classification" have used Convolutional Neural Network for image classification. Image classification has been a core problem in the computer vision field with a significant variety of applications. Machine Learning has gained a lot of attention when dealing with image classification. This paper has proposed a CNN architecture model, InceptionV3, to find out its accuracy and efficiency when dealing with new datasets via Transfer Learning. This model is trained on human-face, cars, and animals. Higher accuracy was achieved when InceptionV3 was used on the above-classified images.

Authors: Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, Bohyung Han, in their published paper titled "Large-Scale Image Retrieval with Attentive Deep Local Features", have suggested a local attribute descriptor appropriate for large-scale image retrieval. It is based on Convolutional Neural Network (CNN). This system generates dependable confidence scores in order to exclude false positives.

The four main blocks used are: -

I. dense localized feature extraction,

II. selection of keypoint,

III. dimensionality reduction,

IV. indexing and retrieval.[9]

The dense features are extracted by giving the image to Fully Convolutional Network (FCN). Then, the features are localized based on their respective fields, which can be computed considering the configuration of pooling and convolutional layers of FCN. Hence, Local descriptors indirectly learn representations that are more applicable to a topic. It can be deduced from the results that although DELF shows great empirical results improving the accuracy nontrivially when combined with Deep Image Retrieval (DIR), it does not show optimum results on its own. Author mentions that "This indicates that DELF has capability to encode complementary information that is not available in global feature descriptors" [9].

3 Methodology

3.1 Transfer Learning

Transfer learning is a machine learning technique where a model trained on one task is re-purposed for a second related task. It is an optimization that improves performance and allows rapid progress when modeling the second task [10].

Transfer learning is the most used deep learning alternative since deep learning models need significant resources to train as well as huge and difficult datasets in which deep learning models are trained.

Transfer learning is effective when features learned by the model in the first task are generic. This deep learning transfer learning technique is known as an inductive transfer. The selection of possible models for the task, is done on the basis of its previous learning so that the model can fit the new task by making small changes.

Usual Transfer Learning Workflow: -

1. Selecting Base Model: A pre-trained base model is selected from a range of models. Many research institutions publish models on the internet that have been conditioned on vast and difficult datasets and are part of a collection of candidate models from which to select.
2. Reuse Model: The pre-trained model will then be used to construct a model for the next task of concern. Depending on the modeling process, this can include any portions of the model.
3. Tune Model: The base model may not be used for the new task in its raw or original state and thus the model needs to be tuned by adding some layers according to the problem statement.

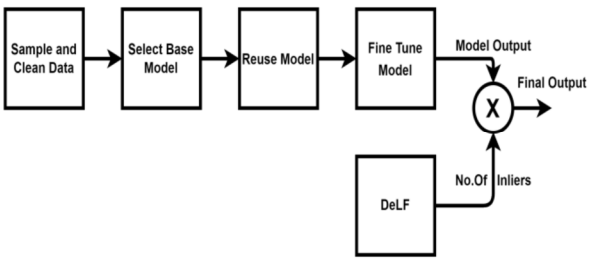

Fig. 1. Workflow

Fig. 1 shows the basic workflow of the system used for the training of the model.

3.2 Optimizing the tuned transfer learning model

After generating the fined tuned model, by transfer learning, of various CNN models (InceptionV3, VGG16, and RESNET50), a maximum accuracy of 90.4 on our "Test" dataset was achieved on the RESNET50 model.

On analyzing the Test dataset, it was found that some structural similarities, color luminance, and shortage of training images were causing errors in the prediction of the model. Thus, to tackle the problems mentioned earlier, DeLF, a Convolutional Neural Network that identifies semantically local features, was used.

The pre-trained DeLF module can be used for image retrieval as a drop-in substitute for other keypoint detectors and descriptors. It describes each noteworthy point in each image with 40-dimensional vectors known as the feature descriptor.

So, to optimize the desired result from the CNN model, a new layer was added in front of our Tuned model, which took the output from both the predicted RESNET values and multiplied that by the top 3 prediction matches with the input image using DeLF.

The maximum multiplied result determined the prediction of the model.

This inclusion of DeLF increased our model's accuracy from 90.4 to 95.6.

3.3 Dataset

This paper uses the Google landmark recognition challenge dataset. The dataset contains around 15k classes and 1.2 M images. Due to limitations of computing resources, a sample of 100 classes containing between 10 and 50 images was taken [11].

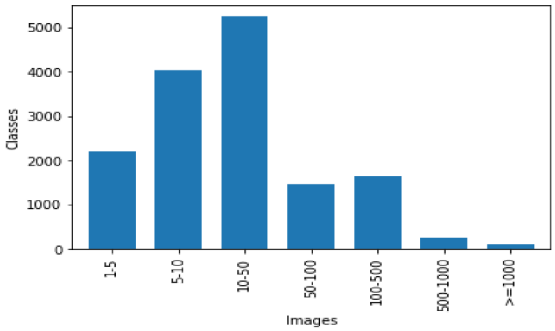

Fig. 2. Distribution of images per class

Fig. 2 shows the distribution of number of images on different class in the dataset.

Total Images (over 100 classes) = 4100
Training split = 75% [3128]
Validation split = 20% [764]
Test split = 5% [208]

3.4 Selecting the base model

The base model was selected from the following three models: VGG 16, InceptionsV3, Resnet 50.

The ImageNet dataset was used to train all the models mentioned above. The ImageNet dataset is a massive archive of images that have been annotated by humans. It was created by academics for computer vision science. The dataset contains slightly more than 14 million images, slightly more than 21 thousand categories or classes (synsets), and slightly more than 1 million images with bounding box annotations. (e.g., boxes around identified objects in the images).

3.5 Architectural Design

For all three of the above models, the following architecture was prepared.

Input Images are 224 x 244 with RGB channels.

Input shape 224 x 224 x 3

Fig. 3 shows the basic architecture of the models used for the paper. The base model is fine-tuned and trained with the dataset to produce accurate results.

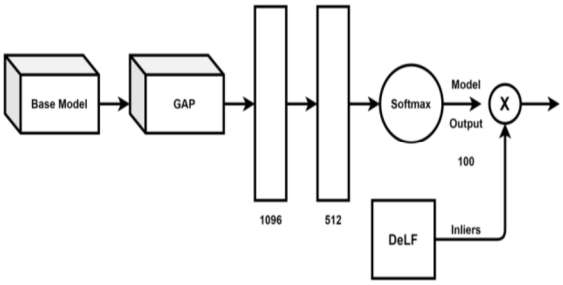

Fig. 3. Base architecture

1. VGG 16:

VGG 16 was proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University in 2014. VGG-16 was one of the best performing architecture in the ILSVRC challenge 2014.

VGG 16 has a 16-layer architecture: 13 convolution layers, two fully connected dense layers, one softmax layer, as shown in Fig. 4.

Number of parameters: ~ 138 million
Output shape: 7 x 7 x 512

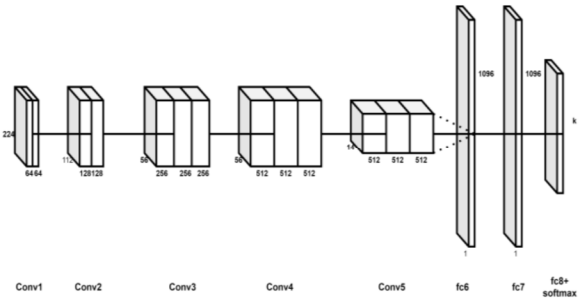

Fig. 4. VGG16 architectures

2. Inception V3:

Inception v3 is a popular image recognition model that has been shown to achieve greater than 78.1 percent accuracy on the ImageNet dataset. The model is the result of several concepts generated over time by various researchers.

Inception-v3 is a 48 layers deep convolutional neural network, as shown in Fig. 5.

Number of parameters: ~ 24 million
Output shape: 5 x 5 x 2048

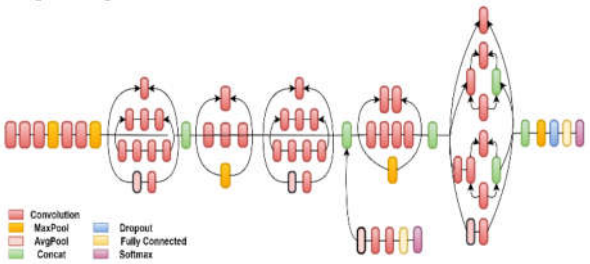

Fig. 5. InceptionV3 architecture

3. ResNet50:

ResNet-50 is a smaller version of ResNet 152. Residual Network is a classic neural network used as a backbone for many computer-vision tasks. In 2015, this model was the winner of the ImageNet challenge.

Number of parameters: ~ 22 million
Output shape: 7 x 7 x 2048

Fig. 6 shows the ResNet50 architecture used to train the model.

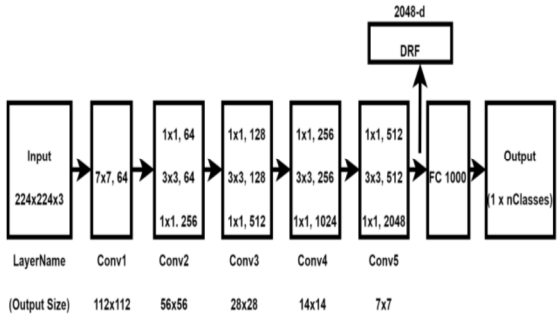

Fig. 6. ResNet50 architecture

4. Pure CNN :

Unlike in traditional Transfer Learning apporaches in Convolution Neural Networks, Pure CNN, as the name suggests, does not contain any Fully Connected layer on the top. On Top of the base model (ResNet 50), the model consisted of another conv2D layer and GAP layer and finally a softmax layer, as shown in Fig. 7.

Total no. of parameters: ~ 23.7 million
Trainable parameters: 105,000

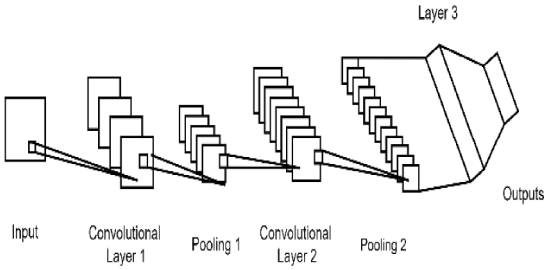

Fig. 7. Pure CNN architecture

5. ResNet50 + DeLF :

This model is an extension of ResNet 50 model used, to increase the accuracy of the earlier trained model. A new DeLF module, produces number of inliers matched between test and an image of classes of three predictions from the ResNet model, as shown in Fig. 8. The number of inliers produced by the DeLF module multiplied by the probability of the class produced by the ResNet model, gives the final output.

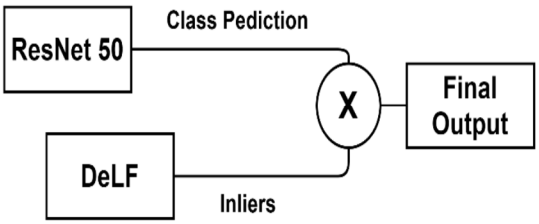

Fig. 8. ResNet50 + DeLF architecture

3.6 Results

Table 1 shows the comparison of various models used in this paper. It is seen from the table that the testing accuracy of most of the models is comparble, but that of ResNet 50 + DeLF is the greatest.

Table 1. Results

Model	Training Accuracy (%)	Testing Accuracy (%)
VGG 16	99.8	89.1
Inception V3	27.3	18.6
ResNet 50	100	90.4
Pure CNN	97.55	89.67
ResNet 50 + DeLF	100	95.67

Fig. 9 shows the comparison of all the models with their training and testing accuracy, on the dataset.

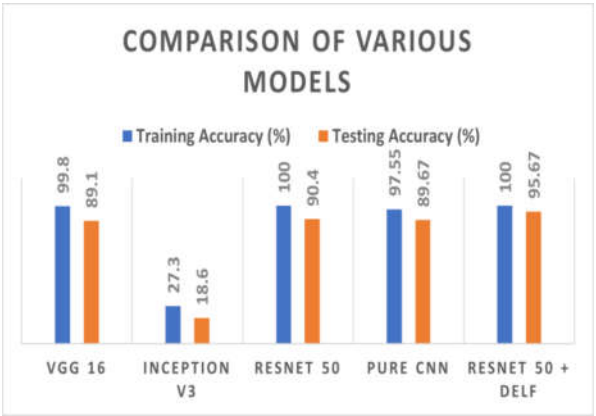

Fig. 9. Comparison of Various models implemented

4 Conclusion

A Landmark Recognition model was implemented, using transfer learning techniques with three different models, namely VGG16, InceptionV3, ResNet50. The accuracy of the models is not high and can be improved using further fine-tuning techniques.

VGG 16 was one of the two pure models which performed well, though slightly behind ResNet it should good empirical results, even with certain manual tests. Overall, it performed extremely well considering it had the shortest training time amongst others. InceptionV3 turned out to be the worst performing base model. It showed a clear case of under-fitting, which was shocking considering it is a complex model. ResNet 50 was by far the most accurate simple model, it showed the best empirical results. Considering this it was the chosen as the base model for the next two models (PCNN and DeLF).Pure CNN model was based on ResNet (base model). Though it initially looked promising, the PCNN model quickly started to show signs of overfitting. It was deduced that the lack of a Fully connected layer was the cause. Combination of ResNet and DeLF outperformed all other approaches discussed above. But addition of the DeLF stage highly increased the recognition processing time, which would be a hindering factor in testing large datasets.

The model was also hosted on a web app made using basic HTML, CSS, and Javascript webpage with python flask framework at server side.

References

1. Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, "Deep Learning for Computer Vision: A Brief Review", *Computational Intelligence and Neuroscience*, vol. **2018**
2. A Data Geek's Guide to Recognize Landmarks https://medium.com/@abhinaya08/google-landmark-recognition-274aab3c71ae
3. Visual Learning for Landmark Recognition http://www.cs.cmu.edu/~takeuchi/iuw97/iuw97.html
4. Instance-level Recognition https://towardsdatascience.com/instance-level-recognition-6afa229e2151
5. A. Kausar, M. Sharif, J. Park, and D. R. Shin, "Pure-CNN: A Framework for Fruit Images Classification," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018,**pp.404-408.**
6. Andrei Boiarov and Eduard Tyantov. 2019. Large Scale Landmark Recognition via Deep Metric Learning. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management Association for Computing Machinery, New York, NY, USA, 169–178.
7. I. Kuzborskij, F. M. Carlucci and B. Caputo, "When Naïve Bayes Nearest Neighbors Meet Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2100-2109
8. Mahbub Hussain, Jordan Bird, and Diego Faria, "A Study on CNN Transfer Learning for Image Classification" at School of Engineering Aston University,Birmingham.
9. Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han, "Large-Scale Image Retrieval with Attentive Deep Local Features"
10. A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning. https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a
11. Google-Landmarks Dataset https://www.kaggle.com/google/google-landmarks-dataset