Geo-Landmark Recognition and Detection

Nishika Manira, Swelia Monteiro, Tashya Alberto, Tracy Niasso, Supriya Patil

Abstract*: The widespread use of smartphones and mobile data in the present-day society has exponentially led to the interaction with the physical world. The increase in the amount of image data in web and mobile applications makes image search slow and inaccurate. Landmark recognition, an image retrieval task, faces its challenges due to the uncommon structure it possesses, such as, buildings, cathedrals, castles or museums. These are shot from various angles which are often different from each other, for instance, the exterior and interior of a landmark. This paper makes use of a Convolutional Neural Networks (CNN) based efficient recognition system that serves in navigation, to organize photo collections, identify fake reports and unlabeled landmarks from historical data. It identifies landmarks correctly from a variety of images taken at different viewpoints as well as distances. An appropriate CNN architecture helps to provide the best solution for the currently selected dataset.

Keywords: Convolutional Neural Networks (CNN), Faster Region Based CNN (Faster RCNN), Histogram of Oriented Gradients (HOG), Rectified Linear Unit (ReLU), Region of Interest (RoI), Region Proposal Network (RPN), Residual Networks (ResNet), Visual Geometry Group (VGG).

I. INTRODUCTION

Objective to build a neural network model that correctly classifies and detects landmarks directly from images, in order to serve as an application in navigation, a tourist and shopping guide and help people organize their photo collections. Andrew Crudge, et al. [1] used Support Vector Machine Classifier where the training data consisted of a vector that contained all the labels, and a matrix whose rows represented the examples and columns represented the features. For feature extraction from images, Histogram of Oriented Gradients (HOG) was used. These methods did not guarantee efficiency as the HOG tool was not invariant to changes in orientation and the cell size of HOG affected the accuracy of the algorithm.

Manuscript received on May 17, 2021.
Revised Manuscript received on May 22, 2021.
Manuscript published on May 30, 2021.
* Correspondence Author
Nishika Manira*, Department of Electronics & Telecommunication Engineering at Padre Conceicao College of Engineering, Verna, Goa, India.
Swelia Monteiro, Department of Electronics & Telecommunication Engineering at Padre Conceicao College of Engineering, Verna, Goa, India.
Tashya Alberto, Department of Electronics & Telecommunication Engineering at Padre Conceicao College of Engineering, Verna, Goa, India.
Tracy Niasso, Department of Electronics & Telecommunication Engineering at Padre Conceicao College of Engineering, Verna, Goa, India.
Dr. Supriya Patil, Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa, India.

This method was complex and time consuming. Maxence Dutreix, et al. [2] elaborated on recognition and retrieval. For recognition, a Deep Neural Network architecture called ResNet was used to classify with a score for each prediction. For Retrieval, Deep Local *Feature* architecture was implemented which handled high-dimensional nearest-neighbor search. This method suffered crashes and memory leakages due to the large dataset, and the limited availability of processing power was a constraint.

In this paper, a CNN model is used for the classification and object detection of landmarks directly from images. Feature extraction is performed by convolution and pooling operations. The fully connected layers predict the class for the object in the image which serves as a classifier. The classification of landmarks is carried out by using various CNN architectures such as AlexNet, VGG-19, ResNet18, ResNet50 and ResNet152. Object detection is implemented using Faster Region Based CNN (Faster RCNN) algorithm. The paper is arranged as follows: data set details are described in part II, followed by pre-processing techniques in part III, feature extraction using CNN in part IV, various pre-trained CNN architectures in part V, followed by results and conclusion.

II. DATASET

This paper utilizes a custom dataset which contains a total of 1000 images consisting of 10 classes of landmarks; including churches, temples, forts and museums. Instagram and Google photos are the main sources to collect raw data. The images from Instagram are obtained by using an insta-scraper called Instaloader, a utility in the Python library. The data set is divided into the ratio of 3:1:1 (60:20:20) for training, validation and testing respectively.

III. DATA PRE-PROCESSING

For efficient training, the images are resized and renamed. So as to improve the model performance, different data augmentations are applied to the images by using the inbuilt functions from Fastai library such as random flipping, cropping and brightness transforms. For object detection, the LabelImg tool is used for annotating the images by creating bounding boxes around the landmark of interest. These annotations are then saved in the Pascal VOC format.

IV. FEATURE EXTRACTION

The deep learning [7] class of machine learning algorithms uses multiple layers to extract higher-level features from the raw input and learns progressively more about the input as it goes through each neural network layer.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Transfer learning can train deep neural networks with little data. In transfer learning, a machine exploits the knowledge gained from a previous task to improve generalization about another i.e., it reuses a pre-trained model on a new problem.

Convolutional neural networks are very effective in reducing the number of parameters without losing on the quality of models in spite of high dimensionality where each pixel is considered as a feature.

CNN has two main components:

Feature extraction part: The feature extraction is performed by a sequence of convolution and pooling operations. Convolution operation is carried out by passing a filter over the image that views a few pixels at a time, where the results are summed up into one value that defines all the pixels the filter observed. Pooling operation reduces the size of the image matrix and allows to train the network faster, that focuses on the most important feature of the image. The Rectified Linear Unit (ReLu) activation function is used to introduce non-linearity in the output.

Classification part: To serve as a classifier on top of these extracted features, the fully connected layers are used. Its output contains a list of class labels attached to the image with its respective probabilities. The classification decision is made based on the label that obtains the highest probability.

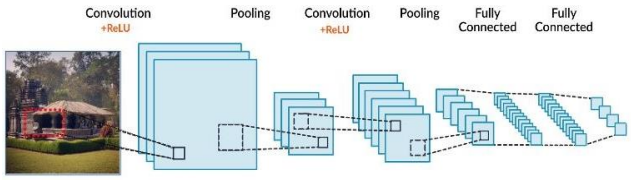

Fig1 -CNN

V. TRAINING MODELS

Fastai [11], an open-source library written in Python which is used to implement the deep learning models. It is built on top of PyTorch and is one of the leading flexible deep learning frameworks. This library can train fast and provide good accuracy. Fastai structures its training process around the Learner class, whose object binds together a PyTorch model, a dataset, an optimizer, and a loss function; the entire learner object further allows to launch training.

A. AlexNet Architecture

The general architecture of AlexNet is shown in the Fig2 [12]. The AlexNet architecture takes an RGB image as an input of size 227×227. AlexNet consists a total of 8 layers: 5 convolutional layers and 3 fully connected layers. The first two convolutional layers are succeeded by the overlapping max pooling layers. The remaining convolutional layers are connected directly. The last i.e., the fifth convolutional layer is followed by an overlapping max pooling layer, the output of which is an input to the fully connected layers. The width and height of the tensors are down sampled by the Overlapping Max Pool layers and the depth is kept unaltered. AlexNet makes use of ReLU to increase nonlinearity in the images. This important feature of AlexNet showed that deep CNNs could be trained much faster than using the saturating activation functions like tanh or sigmoid.

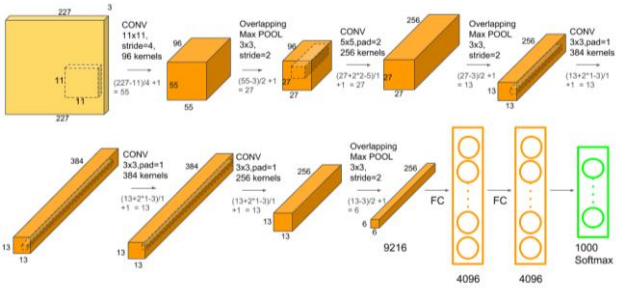

Fig 2 –AlexNet Architecture

B. VGG-19 Architecture

VGG-19 is a variant of VGG (Visual Geometry Group) model which comprises of 19 layers that are 16 convolution layers, 3 fully connected layer, 5 MaxPool layers and 1 SoftMax layer as shown in the Fig3 [13]. The VGG architecture is based on convNet and accepts a 224×224 RGB image as an input. The first two Convolution layers uses 64 filters of a constant size 3×3 with stride 1 that results in 224×224×64 volume. Next, the height and the width are reduced by the pooling layer resulting into a volume of 112×112×64. This pooling operation is achieved by using the max-pool of size 2×2 and stride 2. An additional two convolution layers with 128 filters are added that results in a dimension of 112×112×128. The further pooling layers reduces the volume to 56×56×256. Two more convolution layers are added with 256 filters each followed by down sampling layer that reduces the size to 28×28×512. Two more stack each with 3 convolution layers is separated by a max-pool layer. After the final pooling layer, 7×7×512 volume is flattened into Fully Connected layer with 4096 channels and SoftMax output of 1000 classes.

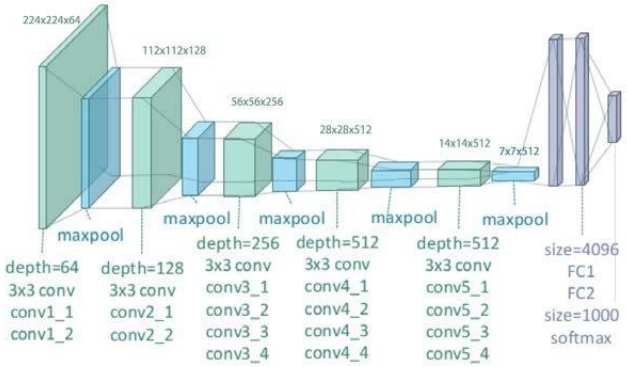

Fig 3 –VGG-19 Architecture

C. ResNet Architecture

ResNet [14], short for Residual Network is a type of neural network that was introduced in 2015 to solve the problem of the vanishing gradient.

Retrieval Number: 100.1/ijitee.G89830510721
DOI: 10.35940/ijitee.G8983.0510721

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

The network skips few layers of training and directly connects to the output by using this skip connection technique. This way the network is allowed to fit the residual mapping. Equation (1) gives Equation (2), the network fit G(x) instead of F(x), initial mapping, as shown in Fig4. This means that the output of a layer is connected to the input of the previous layer.

$$F(x) = G(x) - x \quad (1)$$

$$G(x) = F(x) + x \quad (2)$$

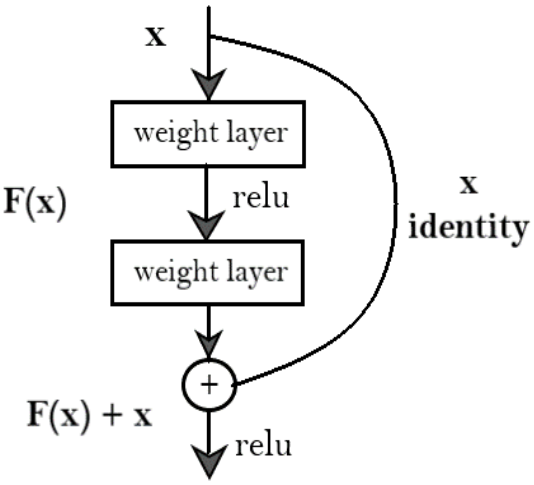

Fig 4 –ResNet Connections

This technique helps to skip by regularization if in any case the layer hurts the performance of ResNet. This aids in training a DNN, eliminating the vanishing gradient problems. The ResNet-18 architecture consists of 17 convolutional layers and 1 fully connected layer. Similarly, a ResNet-152 variant is made up of 151 convolutional layers and 1 fully connected layer. ResNet-50 architecture is shown in the Fig 5 [16]. It has four stages and is a variant of ResNet which comprises of 48 convolution layers, one max-pool and one average-pool layer. The input to the network is an image of size 224 x 224 x 3 which generally has its height and width as multiples of 32 with a channel width as 3. Initial convolution and max pooling are performed using a kernel size of 7×7 and 3×3 respectively. Stage 1 has three residual blocks containing three layers each. The convolution operation in all three layers of stage 1 is performed using a kernel size of 64, 64 and 128 respectively. The width of the channel is doubled and the input size is decreased to half, from one stage to the next. For ResNet152, bottleneck design is used. Three convolutional layers of size 1×1, 3×3, 1×1 are stacked, for each residual function. For the purpose of reducing and restoring the dimensions, the 1×1 convolution layers are used. With smaller dimensions, the 3×3 layer is left as a bottleneck. Lastly, the average pooling layer is followed by a fully connected layer.

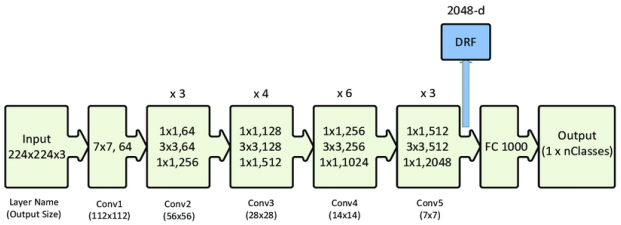

Fig 5 –ResNet Architecture

D. Faster RCNN Architecture

Estimation of an object in an image as well as its boundaries is object detection. In object detection, the object is identified and located by using a bounding box around it. Faster RCNN is a leading object detection architecture that uses CNN. As shown in Fig6 [17], Faster RCNN comprises of 3 parts:

Convolution layers

For the purpose of classification and detection, convolution layers, pooling layers and a fully connected layer are used that makes up the convolution networks. By sliding filters over the input image, convolution operation is performed, giving a 2D matrix called feature map.

Region Proposal Network (RPN)

In order to estimate whether further processing in a given region needs to be implemented, the RPN is used to scan every location quickly and efficiently. This is done by giving n bounding box proposals each with 2 scores that represents the probability of the existence of an object at each location. This is used to reduce the computational complexity. Region of Interest (RoI) pooling layer makes different sizes of region proposals generated from RPN into a fixed size.

Classes and Bounding Boxes prediction

The fully connected neural network is then used to take the regions proposed by the RPN as an input and predict the object class and bounding boxes.

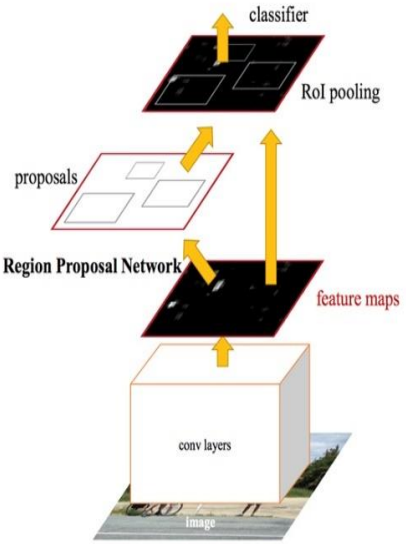

Fig 6 – Faster RCNN Architecture

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

VI. RESULTS

The comparison of classification accuracy obtained by implementation of various CNN architectures such as AlexNet, VGG, and ResNet is as shown in Table I.

Table- I: CNN Architectures and their Accuracy

Architecture	Accuracy
AlexNet	**86.53 %**
VGG-19	**91.12%**
ResNet-18	**84.98%**
ResNet-50	**96.89%**
ResNet-152	**92.23%**

VII. CONCLUSION

Landmark recognition along with detection plays an important role in guiding tourists to navigate through unknown locations, organize photo collections, identify fake reports and unlabeled landmarks from historical data. CNNs is used for the purpose of efficient and accurate classification. The model locates the landmark in the image by localizing it with a bounding box and displays the corresponding class name. The Resnet-50 CNN architecture provides an edge over AlexNet, Vgg-19, ResNet-18 and ResNet-152. It outperforms the method proposed by Andrew Crudge, et al. [1] and Maxence Dutreix, et al. [2] providing an accuracy of 96.89%. The Faster RCNN architecture improved the region proposal quality and thus the overall object detection.

REFERENCES

1. Andrew Crudge, Will Thomas and Kaiyuan Zhu, Article 'Landmark Recognition Using Machine Learning' 2015: http://cs229.stanford.edu/proj2014/Andrew%20Crudge,%20Will%20Thomas,%20Kaiyuan%20Zhu,%20Landmark%20Recognition%20Using%20Machine%20Learning.pdf
2. Maxence Dutreix, Nathan Hatch, Raghav Kuppan, Pranav Shenoy Kasargod Pattanashetty, Anirudha Sundaresan 'Google Landmark Recognition and Retrieval Challenges' Article. April 25, 2018 : https://nhatch.github.io/files/landmarks_report.pdf
3. 'Google Landmark Recognition using Transfer Learning' Article by *Catherine McNabb, Anuraag Mohile, Avani Sharma, Evan David, Anisha Garg :* https://towardsdatascience.com/google-landmark-recognition-using-transfer-learning-dde35cc760e1
4. Blog on end to end Image Classification in fastai.: https://rajaskakodkar.github.io/blog/deep%20learning/2020/10/07/image-classification.html
5. https://medium.com/@abhinaya08/google-landmark-recognition-274aab3c71ae
6. https://en.wikipedia.org/wiki/Deep_learning
7. Introduction to transfer learning: https://builtin.com/data-science/transfer-learning
8. Guide to CNN, by Daphne Cornelisse. April 24, 2018: https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/
9. Basics of Convolution Neural Networks, February 25, 2019: https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529
10. CNN working by Derrick Mwiti, May 8, 2018: heartbeat.fritz.ai/a-beginners-guide-to-convolutional-neural-networks-cnn-cf26c5ee17ed
11. Fast-ai: https://www.fast.ai/
12. 'Understanding AlexNet' by Sunita Nayak : https://learnopencv.com/understanding-alexnet/amp/
13. 'VGGNet Architecture' by Prabin Nepal, July 30, 2020: https://medium.com/analytics-vidhya/vggnet-architecture-explained-e5c7318aa5b6
14. Residual Networks (ResNet): https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/
15. 'Detailed guide to understand and implement ResNets' by Ankit Sachan: https://cv-tricks.com/keras/understand-implement-resnets/
16. https://www.researchgate.net/figure/ResNet-50-architecture-26-shown-with-the-residual-units-the-size-of-the-filters-and_fig1_338603223
17. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks' paper published on 6th January 2016: https://arxiv.org/pdf/1506.01497.pdf

AUTHORS PROFILE

Nishika Manira, student of Bachelor of Engineering in Electronics & Telecommunication Engineering from Padre Conceicao College of Engineering, Verna, Goa.

Swelia Monteiro, student of Bachelor of Engineering in Electronics & Telecommunication Engineering from Padre Conceicao College of Engineering, Verna, Goa.

Tashya Alberto, student of Bachelor of Engineering in Electronics & Telecommunication Engineering from Padre Conceicao College of Engineering, Verna, Goa.

Tracy Niasso, student of Bachelor of Engineering in Electronics & Telecommunication Engineering from Padre Conceicao College of Engineering, Verna, Goa.

Dr. Supriya Patil is an Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. She received B.E (Instrumentation), M.E. (Electronics) degree from Shivaji University and Ph. D (Electronics) from Goa University. She worked for her Ph. D in the area of microarray-based cancer classification. She has 25 years of teaching experience with specialization in Signal Processing and Artificial Neural Network. She has authored and co-authored over twenty conference/journal papers in field of Artificial Neural Network.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication