A VGG-16 Framework for an Efficient Indoor-Outdoor

Monika Dandotiya, Madhukar Dubey

MITS Gwalior, MP, India

Corresponding author: Madhukar Dubey, Email: md8120444411@gmail.com

Computer vision had reached a new level that allows robots from the limits of laboratories to explore the outside world. Even with progress in this area, robots are struggling to understand their location. The classification of the scene is an important step in understanding the scene. In many applications, a scene classification can be used such as a surveillance camera, self-driving, a household robot, and a database imaging system. Monitoring cameras are now everywhere installed. The accuracy of scene classification of indoor-outdoor techniques is weak. Using the Convolution Neural Net-work Model in VGG-16, this study attempts to improve accuracy. This research presents a new method for classifying images into classes using VGG-16. The algorithm's outputs are validated using the SUN397 indoor-outdoor dataset, and outcomesdemonstrates that the suggested methodology outperforms existing technologies for indoor-outdoor scene classification. In this paper, Very Deep Convolutional Networks for Large-Scale Image Recognition" is what we implement. In ImageNet, a dataset of over 14 million images belonging to 1000 classes, the model achieves 92.7 percent top-5 test accuracy. It outperforms Alex Net by sequentially replacing large kernel-sized filters (11 and 5 in the first and second convolutional layers, respectively) with multiple 33 kernel-sized filters. We attain Training loss is 10percent and Training Accuracy is 96 percent in our projected work.

Keywords: Outdoor-Indoor Classification, VGG-16, Classification, Deep Learning.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 321–330. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-93-91842-08-6-32

1 Introduction

One of the holy challenges of computer vision is the problem of scene classification. If we take an arbitrary photo we want to describe what kind of semantic scene it shows Very little work has currently been done in this field, probably due to the very difficult problem and the lack of an agreed scene description language Low-level image analysis that rarely attempts to bridge the gap to the description of the somaticized scene. The indoor and outdoor scene recognition methods are vividly implemented in handheld assistance to help visually challenged people in different environments of unknown public places like the library, temple, an airport terminal, cafeteria, etc. In the robotics field, scene classification algorithms help robots recognizing the type of environment in which they are working. Also, many pictures are being clicked by photographers at different places across the world every day [1]. Many aspects of image processing require the classification of image scenes. Indoor and Outdoor classification is a core component of scene processing as it is the initial step of many semantic scene evaluation methods [2].

Many novel methodologies have been applied to address this issue, but each technique relies on its image database, which reduces the likelihood of success. The problem of classifying the scene is a difficult one since the high-ranking entities are typically considered to be part of another type of scene. Similar objects like plants can be found in either class in indoor or outdoor classification, for example. Several methods were proposed for automatic classification with different success levels and rely largely on low color and texture characteristics. Study on the scene classification has grown considerably in the last decades. Rapid technological advancement in CBIR and growing demand for online storage spaces, improved organization & retrieval of the image database. The Scene classification has (SC) been known as the core field of CBIR study. Classification indoor/outdoor scene using fuzzy C means clustering where accuracy level is inadequate [3] future modern classification methods depended on image edge straightness analysis as indoor images include man-made objects with straight edges while tree-like outdoor, a mountain without straight edges [4].

For several years, the issue of scene classification has been examined from numerous angles in literature about all classification approaches to the Indoor-Outdoor scene can be summarized. There are usually 2 stages called training & classification phases. Extract image features are the first step for both training and classification phases. Researchers have developed different types of features to demonstrate a contrast between the indoor & outdoor images. Feature extraction is commonly considered to be critical for the classification of the indoor-outdoor scene [5].

Convolutional Neural Networks are a type of directed acyclic graphs. Such kinds of networks will have the ability to learn immensely high non-linear functions. A neuron is a primary unit in a CNN. Every layer in CNN is comprised of several neurons. The convolutional features have a more distinguish portrayal of scene images than those of features extracted by image processing methods. The convolutional features are learning-based features that contain rich semantic information, which is more potent and better applicable for scene classification. We should note that the low-level features that contain descriptive details cannot be ignored [6].

The following is the outline for the paper. We present the relevant work in Section 2. The third section provides an overview of the proposed methodology for scene radiation. Section 4 shows the outcome of the proposed approach's experiment. Sections 5 and 6 outline our future work and draw some conclusions.

2 Literature Survey

In recent times, there are many improvements and developments are made in image recognition which is used in scene classification that is to differentiate the different classes using neural networks. Scene recognition and classification or scene categorization has been broadly carried out in various environments. One tactic of this type is convolution Neural Networks (CNN's) with dense

net teaching proposals use image region proposals and another way is to discuss contextual Knowledge that is meant to distinguish the various images according to their groups and properties between different image segments.

A CorrFusion module was proposed which merged the components that are highly correlated with bi-temporal integrations. The deep depictions of the bi-temporal i/p are first extracted with deep convergence networks. Extracted features are then projected into a smaller space to extract most corresponding components & calculate the correlation between illustration and level. The details of the back-propagation gradients are also given for the proposed module. In addition, we presented a much superior scene detection dataset with more semanticized categories and carried out extensive experiments on the scene [7].

Report preliminary results using the two RS scene datasets UC Merced and optimal31.SC has become a hot matter in the field of remote sensing (RS). The distinctive answer entails manually labeling a large adequate set of RS scenes, with an expert opinion if necessary, and then training procedure on this set to learn how to correctly classify novel scenes. The best DL framework required a big labeled database for training. As a result, clever machine learning techniques that can learn to classify RS datasets with new unknown classes from a few tagged samples are critical. Few-shot machine learning is the name for this problem. In this paper, we propose a deep few-shot learning technique for RS scene classification. The planned method is built on prototypical deep neural networks combined with Squeeze Net pre-trained CNN for image embedding [8].

A model for understanding & distinguish a scene using depth data to let machines view real-time scenes as human beings. The proposed recognition method is a novel segmentation system, that usages multi-object statistical segmentation to learn & segregate objects in the scene. Unique features are then stripped for further recognition using linear SVM from these separated objects. Lastly, features & weight of scene recognition are given for multilayer perceptron. Our system has significantly built on state-of-the-art systems. Sport & security systems are proposed effectively in autonomous vision systems like robotic vision, GPS location Finder, etc. The intelligence capacity of computer is the day by day by technological advances. Researchers are committed to humanly equipping devices. The machines can currently sense and process sensor information. Though, the desire to think and understand real scenes remains enormous. Understanding of the scene is now a day for study [9].

Humans are extremely proficient to understand high-level systems of visually perceiving natural scenes. Scene understanding lately is a challenge and a major computer vision issue. Images are visual but visual data may have different characteristics similar to shape, edges, texture & color. Identifying objects in the image and where they are all located is the main aim behindhand object detection. The understanding of a scene includes meaningful information to extract semantic connections & patterns at multiple levels. Interaction of separate objects for human beings is most innate and natural. Finally, certain challenges in the detection of objects remain unresolved and may be used for further analysis [10].

Propose an indoor-outdoor classification ensemble based on cellular data from an LTE commercial network based on a traditional municipal field. The variables are derived by KPIs and radio propagation information of network core performance indicators. The DT grows & breaks into the Gini index of sampled characteristics, depending on these key variables. Through the development and introduction of mass-mobile devices for the next 5G era, wireless Big Data has attracted considerable interest. For personalized services in a smart world, the context information about these devices is essential. The ongoing change in scenarios, however, challenges the network operator [11].

O. Ye et al. proposes a mine video SC system, enhanced CNNs. algorithm increases the accuracy of feature extraction for complex background videos by the growing depth of the initial CNN network. This network communication structure contains 10 neuron layers. The problem of classification with a difficult background for video scenes is how to certify the accuracy of video processing & classification [12].

3 Proposed Work

In this segment, we investigated the diverse investigation.

3.1 Problem Statement

In this investigation, we rely on two publicly accessible datasets that comprise X-It is observed that the Previously proposed model is highly efficient while classifying outdoor classes and getting a little bit confused while classifying indoor scene categories. Airport terminal and Gymnasium are the two major classes, where the model is less efficient in classifying accurately. The major reason could be that the activations of these classes did not fetch the desired features for training. Recovering images or pictures from this large compilation of databases is a highly time-consuming and complex task. The deep learning approach in many computer vision tasks has recently achieved good performance. Thus, it also needs to be multidisciplinary scholars, in particular academics like neurobiology and machine learning, to focus on indoor-outdoor problems for scene classification.

3.2 Proposed Methodology

A VGG16 framework for indoor-outdoor scene classification from photos of diverse public scene contexts is proposed in this paper. For our proposed learning model, we used 'VGG16,' a well-known pre-trained network. Preprocessing is required in the first step. Preprocessing refers to all of the data transformations that occur before it is translated into the model, such as centering, normalization, rotation, shifting, shear, and so on.

3.2.1 Image DataGenerator

A DG can also describe the validation dataset and the evaluation dataset. An additional Image Data Generator instance is also used, using the same pixel scaling configuration (not documented) as the one used for Image Generator's training dataset. Because data augmentation is only used to artificially enlarge the training data set in order to improve model efficiency on an augment-ed dataset, this is the case. In the proposed technique, we used geometric transforms such scaling, zooming, horizontal flipping, image size, batch size, images, classes, color channel, test data image, and validation image.

3.2.2 Neural network model VGG16

In recent years, DL framework has been deployed. In this study, we employed VGG-16 in CNN to achieve this goal. CNN is a type of ANN that uses many perceptrons to evaluate image inputs and has learnable weights and bases for numerous sections of images that can be distinguished from one another. When various parameters are swapped, one advantage of utilizing CNN is that it exploits local spatial coherence of the input images to minimize their weight. The efficient usage of CNNs in image identification tasks has been expedited because to research in architectural design. CNN architectures follow a simple but effective design idea. Their architecture, known as VGG, was based on a layering pattern.

VGG16 is a CNN model presented in the publication "Very Deep Convolutional Networks of Large-Scale Image Recognition" by A. Zisserman and K. Simonyan from Oxford University. This model achieves 92.7 percent of the highest test accuracy in ImageNet, a dataset of over 14 million images belonging to 1000 classes. This was one of the most popular models during the ILSVRC-2014. It outperforms AlexNet by replacing broad kernel-sized filters (5, 11, and 11 on the first and second layers of convolution, respectively) with numerous 33 kernel-sized filters one by one.

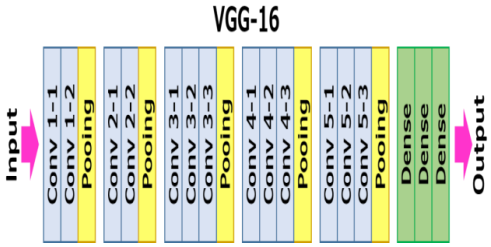

Fig.1. VGG16 Model

3.2.3Loss Function

Categorical Cross Entropy: For categorical classification, cross-entropy loss underwritten by training data point$_i$ (x_i, y_i), is simply the "negative log-likelihood (NLL)":

$$L_i = -\log(p_{y_i}) \qquad (1)$$

Because the chance of the correct label y i is one and zero for all other labels, the ground truth probability is 1for precise label y i and 0 for all other labels.

3.2.4 Optimizer (Adam)

Adaptive Moment Approximation is a gradient descent optimization algorithm. The process is effective if you work with a great number of data or parameters with a large problem. It needs less storage and is efficient. The algorithm 'gradient descent and momentum' is an intuitive combination of the algorithm 'RMSP.'

3.2.5 Mathematical Characteristic of Optimizer (Adam)

Taking the formulas used in the above two methods, we get:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\left[\frac{\delta L}{\delta \omega_t}\right] v_t = \beta_1 v_{t-1} + (1-\beta_1)\left[\frac{\delta L}{\delta \omega_t}\right] \ldots (2)$$

m_t and v_t are approximations of 1st moment (mean) & 2nd moment (uncentered variance) of gradients correspondingly, here after technique name. Because m t and v t are prepared as vectors of 0's, Adam investigators have discovered that they are biased towards zero, particularly during initial time steps and when decay rates are low (like _1& 2 are nearly 1).

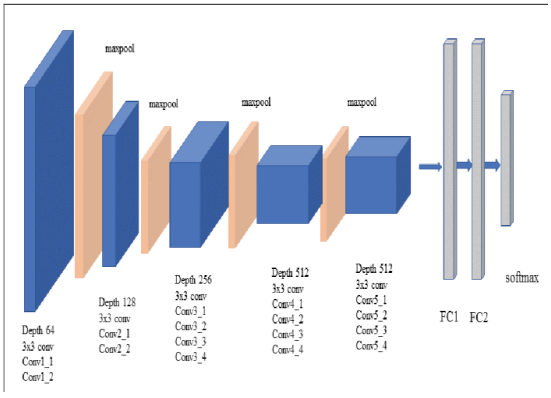

Fig.2. Proposed model Overview

3.2.6 Proposed Algorithm

Step 1: Gather the i/p images from Dataset (SUN397).
Step 2: Preprocess the collected images
Step 3: CNN model Training
Step 4: Modify proposed model (VGG 16) with different layers.
Step 5: Images Testing.
Step 6: Predicted Outputs.

The below figure shows the flowchart of pseudo code in which we will use the additional approach to decrease the loss and increase the accuracy percentage. The CNN architecture transfers the output of one layer to the next layer as an input. Layers of convolution are made with weights and biases used to filter a representation of the input. A collection of filtered images is the output of a convolution sheet. This, this, Activations of the layer are called performance. These activations are a 3-D series, where a channel is also considered the third dimension. In this work, DenseNet concatenates. Dense Connections, also known as Fully Connected Connections, are a type of deep neural network layer that uses a linear operation to connect every input to every output via a weight.

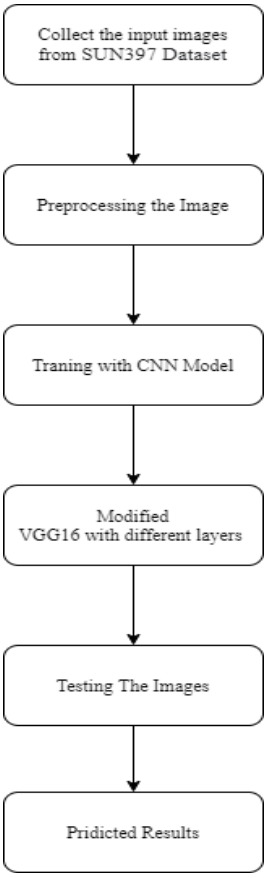

Fig.3. A Block chart of Proposed Approach

4 Experimental Results

To test the proposed approach, this study was carried out in Python programming. The precision of indoor classes is better than that of outdoor classes from given dataset after training the networks individually for indoor and outdoor classes. SUN397dataset is a publicly available dataset that was used for this study.

4.1 Dataset Description

SUN397 is a 397-category scene benchmark that includes indoor, man-made, and natural categories (at the least before places). The SUN397dataset is available to the general public. This dataset is difficult to process not only because of the vast number of categories, but also because of the short quantity of training data and the wide range of object and layout features. It is widely regarded as industry standard for scene classification. Seven scales are used in our tests, each of which is 227x227 by scale photographs.

Table 1. Model Summary

Layer(Type)	Output Shape	Parameter
Con2D	(None, 224,224,64)	1792
Con2D_1	(None, 224,224,64)	36928
Max_pooling2d(max pooling 2d)	(None, 112,112,64)	0
Batch Normalization	(None, 112,112,64)	256
Conv2d-2(conv2D)	(None, 112,112,128)	73856
Conv2d-3(conv2D)	(None, 112,112,128)	147584
Max_pooling2d-1(max pooling 2d)	(None, 56,56,128)	0
Batch Normalization-1	(None, 56,56,128)	512

Table 2. Parameter's information

Parameters	value
Dataset	sun397
Neural network model	VGG16
Batch size	64
Horizontal flip	True
Test data image	518
Validation image	515
Epoch	100
Image size	227*227

4.2 Results Analysis

The examination of the data acquired by the suggested model is presented in this subsection.

Table 3. Accuracy, Loss, Val loss, and Val Accuracy for the Base and Proposed Models are compared

Model	Loss	Accuracy	Val_loss	Val_Accuracy
Base model	0.2623	0.9113	1.7929	0.5615
Proposed model	0.1029	0.9637	1.2143	0.8913

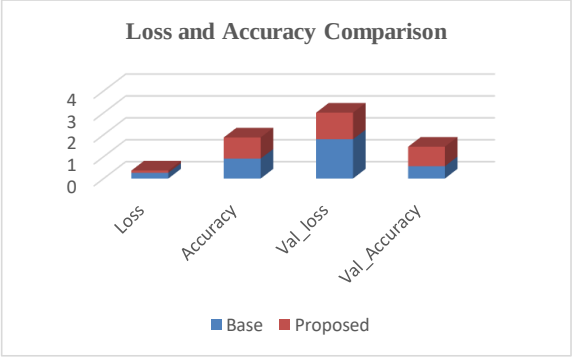

Fig. 4. Graph comparison of accuracy, Loss, Val loss & Val Accuracy for Base and Proposed Model

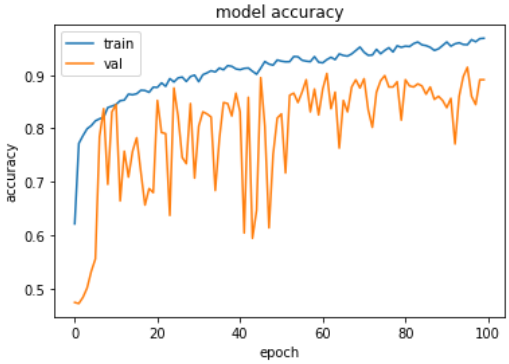

Fig. 5. Line Graph for Model Training Accuracy

The above graph for model training is shown in Figure 5. This cycle can last up to 100 epochs. It demonstrates the correctness of training and validation. It starts with a 62 percent training accuracy and subsequently increases to 96 percent accuracy after 100 epochs. It also demonstrates the precision of validation. It starts validation accuracy at around 23%, with a variable increment inaccuracy. It consistently improves its accuracy to around 89 percent.

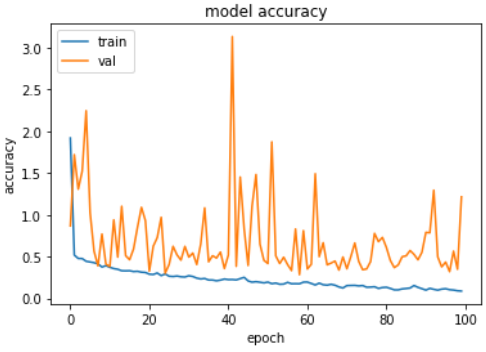

Fig.6. Line Graph for Model Loss Accuracy

The above figure represents a line graph for model loss accuracy. This cycle can last up to 100 epochs. It demonstrates the correctness of training and validation. It commences with 62 percent loss accuracy and subsequently increases to 96 percent accuracy after 100 epochs.

5 Conclusion

Scene categorization is an example of an application where the machine plays a key part in scene analysis. Scene categorization is the process of capturing a scene using machine vision and then attempting to categories it using the computer's past knowledge. It is possible to classify an indoor scene, such as a bakery, airport, garage, or bedroom, as well as an outdoor scene, such as a mountain, roadway, beach, or desert. Inside scene categorization is more challenging than outside scene classification due to the variety of indoor situations. Various indoor scene categorization algorithms have been developed over the years, each with its own set of problems, the most important of which is accuracy. In this study, we propose very deep convolutional networks for large-scale image recognition." The model achieves 92.7 percent top-5 test accuracy in ImageNet, a dataset of over 16 million pictures belonging to classes. It outperforms Alex Net by substituting big kernel-size filters with numerous 33 kernel-size filters in the first and second convolutional layers, respectively. We accomplish a Training Loss of 10% and a Training Accuracy of 96 percent in our suggested task.

6 Future Roadmap

Recently, in many computers vision tasks the deep learning approach perform well. But how to classify indoor-outdoor scene flawlessly is still difficult to explain. Therefore, multidisciplinary scholars need to focus on classification problems in indoor-outdoor situations, particularly academics for example, cross-referencing neurobiology and machine learning to obtain more breakthroughs. In the future, we'll create an indoor-outdoor dataset based on existing data sets. We'll also try out a comprehensive learning approach and compare it to what's already in place.

References

[1] Payne, A. and Singh, S. (2005). Indoor vs. outdoor scene classification in digital photographs. *Pattern Recognition*, 38(10):1533-1545.

[2] Deng, D. and Zhang, J. (2005). Combining multiple precision boosted classifiers for Indoor-outdoor scene classification. In *ICITA*, 720-725.

[3] Szummer, M. and Picard, R. W. (1998). Indoor-Outdoor Image Classification. *IEEE Intl Workshop on Content based Access of Image and Video Databases*.

[4] Tao, L., Kim, Y. H. and Kim, Y. T. (2010). An efficient neural network-based indoor-outdoor scene classification algorithm. *Digest of Technical Papers International Conference on Consumer Electronics,* 317-318.

[5] Vogel, J. and Schiele, B. (2006). Semantic Modeling of Natural Scenes for Content-Based Image Retrieal. *International Journal of Computer Vision,* 72:133-157.

[6] Noh, H., Hong, S. and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *International Conference on Computer Vision*, arXiv:1505.04366.

[7] Ru, L., Du, B. and Wu, C. (2021). Multi-Temporal Scene Classification and Scene Change Detection with Correlation Based Fusion. In *IEEE Transactions on Image Processing*, 30:1382-1394.

[8] Alajaji, D. et al. (2020). Few-Shot Learning for Remote Sensing Scene Classification. *Mediterranean and Middle-East Geoscience and Remote Sensing Symposium,* 81-84.

[9] Rafique, A. A., Jalal, A. and Kim, K. (2020). Statistical Multi-Objects Segmentation for Indoor/Outdoor Scene Detection and Classification via Depth Images. In *17th International Bhurban Conference on Applied Sciences and Technology*, 271–276.

[10] Pawar, P. G. and Devendran, V. (2019). Scene Understanding: A Survey to See the World at a Single Glance. In *2nd International Conference on Intelligent Communication and Computational Techniques,* 182–186.

[11] Zhang, L. et al. (2019). An Ensemble Learning Scheme for Indoor-Outdoor Classification Based on KPIs of LTE Network. *IEEE Access*, 7:63057-63065.

[12] Ye, O. et al. (2018). Video scene classification with complex background algorithm based on improved CNNs. In *IEEE International Conference on Signal Processing, Communications and Computing*, 1-5.