

**Pune Institute of Computer Technology  
Dhankawadi, Pune**

**A SEMINAR REPORT  
ON**

**IMAGE CLASSIFICATION AND ANNOTATION IN SCENE  
IDENTIFICATION**

**SUBMITTED BY**

**Name : Aditya Vyawahare**

**Roll No. : 31468**

**Class: TE-4**

**Under the guidance of**

**Prof. R. A. Kulkarni**



**DEPARTMENT OF COMPUTER ENGINEERING  
Academic Year 2021-22**



DEPARTMENT OF COMPUTER ENGINEERING  
**Pune Institute of Computer Technology**  
**Dhankawadi, Pune-43**

**CERTIFICATE**

This is to certify that the Seminar report entitled

**“IMAGE CLASSIFICATION AND ANNOTATION IN  
SCENE IDENTIFICATION ”**

Submitted by

Aditya Vyawahare                      Roll No. : 31468

has satisfactorily completed a seminar report under the guidance of  
Prof. R. A. Kulkarni towards the partial fulfillment of third year  
Computer Engineering Semester II, Academic Year 2020-21 of  
Savitribai Phule Pune University.

Prof. R. A. Kulkarni  
Internal Guide

Dr. M.S.Takalikar  
Head  
Department of Computer Engineering

Place:Pune  
Date: 16/11/2021

## ACKNOWLEDGEMENT

It is my pleasure to present report on "Image classification and annotation in scene identification". First of all, I would like to thank our Seminar Coordinator Prof. Deepali Kadam, Head of Department Dr.M.S.Takalikar and Principal Dr. R.Shreemati for their encouragement and support.

I would also genuinely express my gratitude to my guide Prof. R. A. Kulkarni, Department of Computer Engineering for her constant guidance and help. She has constantly supported me and has played crucial role in completion of this report. Her motivation and encouragement from beginning till end to make this seminar a success.

Last but not the least I would thank all the faculty,my parents and friends who have helped me.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>2</b>	<b>MOTIVATION</b>	<b>3</b>
<b>3</b>	<b>LITERATURE SURVEY</b>	<b>4</b>
<b>4</b>	<b>A SURVEY ON PAPERS</b>	<b>5</b>
4.1	Object Detectors Emerge In DeepScene CNNs . . . . .	5
4.2	Learning Deep Features for Scene Recognition using Places Database	5
4.3	Scene recognition with CNNs: objects, scales and dataset bias . . .	5
4.4	Semantic-Aware Scene Recognition . . . . .	5
<b>5</b>	<b>PROBLEM DEFINITION AND SCOPE</b>	<b>7</b>
5.1	Problem Definition . . . . .	7
5.2	Scope . . . . .	7
<b>6</b>	<b>CHALLENGES FACED IN SCENE CLASSIFICATION</b>	<b>8</b>
6.1	Large intraclass variation . . . . .	8
6.2	Semantic ambiguity . . . . .	8
6.3	Computational efficiency . . . . .	8
<b>7</b>	<b>DIFFERENT ALGORITHMS USED FOR SCENE CLASSIFI- CATION</b>	<b>9</b>
7.1	Object-Centric CNNs . . . . .	9
7.2	Scene-Centric CNNs . . . . .	9
<b>8</b>	<b>METHODOLOGY</b>	<b>10</b>
8.1	Workflow . . . . .	10
8.2	Mathematical Model . . . . .	12
8.3	VGG Model Architecture . . . . .	13
<b>9</b>	<b>RESULTS</b>	<b>14</b>
9.1	Evaluation Of The Models . . . . .	14
9.2	Visualizing Predictions . . . . .	15
<b>10</b>	<b>CONCLUSION</b>	<b>16</b>
	<b>References</b>	<b>17</b>

## List of Tables

1	Literature survey . . . . .	4
---	-----------------------------	---

## List of Figures

1	Object-Centric Workflow(VGG16+SVM) . . . . .	10
2	Scene-Centric Workflow(VGG16) . . . . .	11
3	Architecture of VGG16 (Source: <a href="https://neurohive.io/en/popular-networks/vgg16/">https://neurohive.io/en/popular-networks/vgg16/</a> ) . . . . .	13
4	Model evaluation of Object-Centered (VGG16 + SVM) . . . . .	14
5	Model evaluation of Scene-Centered (VGG16) . . . . .	14
6	Correct predictions . . . . .	15
7	Incorrect predictions . . . . .	15

## Abstract

In computer vision and machine learning Image classification and annotation are one of the important problems. Image classification (or Image recognition) is a sub domain of computer vision in which an algorithm looks at an image and assigns it a tag from a collection of tags which are predefined or categories that it has been trained on whereas image annotation is the process of assigning labels to the images of a dataset to train a machine learning model. Intuitively, annotations provide meaningful insights for the class label, and the class label provides meaningful insights for annotations. The meaningful annotations for the image help us to identify the objects in the image and can also be used to classify the image into different scenes. It can be used for automatic tag generation on social media websites, movie genre prediction, video websites, etc. I evaluate two models, object centric (VGG16+SVM) and scene centric (VGG16), to classify images into it's scenes topics and provide conclusions based on the accuracy of the models.

## Keywords

Machine-learning, Image Annotation, Image Classification, Scene Classification, Deep-Learning, CNNs

# 1 INTRODUCTION

Automatic scene classification (sometimes referred to as scene recognition, or scene analysis) is a longstanding research problem in computer vision, which consists of assigning a label such as 'beach', 'bedroom', or simply 'indoor' or 'outdoor' to an image presented as input, based on the image's overall contents. Object classification focuses on classifying objects in the foreground, Scene Classification uses placement and layout of objects within the scene for classification.

Humans can recognize and classifying scenes in a tenth of a second or less, thanks to our ability to capture the gist of the scene, even though this usually means having missed many of its details. For example, we can tell an image of a bathroom from one of a bedrooms quickly, but would be dumbfounded if asked (after the image is no longer visible) about specifics of the scene (for example, how many nightstands / sinks did you see?). It is also not possible for a human to classify big datasets. Hence the need of automatic scene recognition increased.

Scene classification is image classification task where an image is to be classified into some defined set of topics. Another similar term is Image annotation, which is, the process of labeling images of a dataset to train a machine learning model. Therefore, image annotation is used to label the features you need your system to recognize. The annotation task usually involves manual work, sometimes with computer-assisted help. A Machine Learning engineer predetermines the labels, known as "classes", and provides the image-specific information to the computer vision model. After the model is trained and deployed, it will predict and recognize those predetermined features in new images that have not been annotated yet.

Image annotation as well as classification are both critical and challenging work in computer vision research. Due to the rapid increasing number of images and inevitable biased annotation or classification by the human curator, it is desired to have an automatic way. Recently, there are lots of methods proposed regarding image classification or image annotation. However, people usually treat the above two tasks independently and tackle them separately. There is a relationship between the image class label and image annotation terms. For example, image of class highway is more likely annotated with words "road", "car", "truck" and "signal" than words "fish". "boat" and "scuba" which are the descriptions of beach.

I will evaluate two methods of classifying scenes. In the both the methods. I will use the pre-trained VGGNet16 model for image processing that is trained on the ImageNet dataset provided in Keras. ImageNet is a standard dataset used for classification. It contains more than 14 million images in the dataset, with little more than 21 thousand groups or classes. In the first method I will classify the scenes directly using the pretrained VGGNET16 and evaluate the results. In the second model I will train the VGGNet16 model to find the objects in the image and that will be our annotation, and then I will use the annotations to classify it into the scenes. Finally, I will evaluate both the methods and find out the most suitable method to classify scenes.

The dataset I will use to evaluate our methods is the LabelMe dataset. It consists of eight category classes: coast, highway, inside-city, forest, open-country, mountain, street and tall-building. In order to keep the balance of the shape number of images for each class, followed I reduced the size of each image as  $256 \times 256 \times 3$ . Thus, the total number of images is 2688.



## 2 MOTIVATION

In recent times, for several decades scene classification has been an active area of research , and it has a wide range of applications like content-based image retrieval, intelligent video surveillance, robot navigation, automatic tag generation, augmented reality and disaster detection applications. As the core of scene classification, scene representation is the technique of transforming a scene image into its concise descriptors, and still attracts tremendous and increasing attention. However, scene classification has been a difficult task because of the high variability, ambiguity, and the huge range of illumination and conditions like scaling in the pictures.

The recent resurrection of interest in Artificial Neural Networks (ANNs), generally deep learning ,has revolutionized computer vision and been ubiquitously used in various tasks like semantic segmentation, object classification and detection and scene classification. Hence to accurately classify the images into their scenes using deep learning models has been a need in today's world.

### 3 LITERATURE SURVEY

The Following table shows the literature survey by comparing techniques proposed in various references:

Table 1: Literature survey

No.	Paper	Summary	dataset	limitations
1	Simultaneous image classification and annotation	Two models developed multi-class sLDA and multi-class sLDA with annotations	UIUC-Sport LabelMe	No Neural Networks used
2	Indoor/outdoor scene classification project. Pattern Recognition and Analysis.	Both a k nearest neighbour classifier, and a linear classifier were used. Fisher's Linear Discriminant was used for dimensionality reduction	1343 consumer photographs	Dataset not large enough to classify and test images. Restricted to indoor-outdoor images.
3	Learning Deep Features for Scene Recognition using Places Database	Compared 2 models. a. ImageNet-CNN+SVM b. Scene-centric CNN Proved scene-centric models perform better than object-centric	Places Dataset	Pre-trained CNN used and was not fine-tuned on the dataset
4	Object Detectors Emerge In Deep Scene CNNs	Trained a CNN model to extract object details and then classify into scenes	ImageNet Places Dataset	The performance of object-centric CNNs is affected by scaling any shift to describing scenes would suddenly affect their performance
5	Scene recognition with CNNs: objects, scales and dataset bias	address two problems: 1) scale induced dataset bias in multi-scale CNN 2) how to combine object-centric and scene-centric knowledge.	ImageNet Places Dataset	Fine Tuning model limits overall improvement of the model

## 4 A SURVEY ON PAPERS

### 4.1 Object Detectors Emerge In DeepScene CNNs

Showed that object detectors are created from training CNNs to perform scene classification. Because scenes are composed of objects, the CNN for scene classification automatically catches meaningful objects detectors, representing the learned scene categories. Hence, the same network can perform both object localization and scene recognition in a single pass, without ever having been taught explicitly the notion of objects.

Proved that object detectors are created as a result of learning to classify scene categories, showing that a single network could support recognition at various levels of abstraction (e.g., edges, textures, objects, and scenes) without needing more than one networks or outputs.

### 4.2 Learning Deep Features for Scene Recognition using Places Database

Introduced a new Places dataset which is a scene-centric database. Places with about 7 million pictures labeled of different scenes and it also proposes some new methods to compare the diversity and density of image datasets and show that Places is as rich in density than other scene dataset and has more diversity. Demonstrated that scene-centric and object-centric neural networks are different in their internal representations by introducing a simple visualization of the receptive fields of CNN units. Hence proved that classification performance of scene-centric Convolutional Neural Networks are better than object-centric Convolutional Neural Networks.

### 4.3 Scene recognition with CNNs: objects, scales and dataset bias

This paper addresses to main problems- Scale induced dataset bias in multi-scale CNNs, and how to combine scene-centric and object-centric knowledge in CNNs. Moreover it describes them as a couple of particular cases in a more general view of how multi-scale features can be combined for scene identification. They are not incompatible, and actually when combined properly to reduce the dataset bias the results can be excellent, even reaching human recognition performance simply with just two or three networks carefully chosen. The evaluation was done on two datasets Places and ImageNet and it proved that carefully chosen combinations on those datasets can increase the accuracy up to 66.26% and with deeper architectures it can reach up to even 70.17%.

### 4.4 Semantic-Aware Scene Recognition

Addresses the problem of scene classification with data augmentation to fine-tune CNNs. It concluded that fine-tuning a CNN model have certain “equal-

izing” effect between final accuracy and the input patch scale, that is, to some extent, with too minuscule patches as CNN inputs, the final classification accuracy is much worse.

Also examined the effect of fine-tuning CNNs on distinct scales that is with distinct scale patches as inputs. There was a average accuracy gain in the range of scale patches where the original CNNs perform badly. This paper demonstrated that Places-CNN has the best performance in the whole ranged patches of scale, in this case, fine-tuning on target dataset leads to slight performance improvement.

## 5 PROBLEM DEFINITION AND SCOPE

### 5.1 Problem Definition

To propose and evaluate two deep-learning models, object-centric (VGG16+SVM) and scene-centric (VGG16), that accurately identifies the scene labels from the image inputs.

### 5.2 Scope

The factors which are crucial and need to be considered for a successfully classify images into scene are variability, ambiguity, and the wide range of illumination and scale conditions of the images in the dataset. These factors have a major impact on the optimization of the results.

Hence the model needs to be evaluated on log-loss score and accuracy score on train, validation and test sets to accurately predict the scenes.

## 6 CHALLENGES FACED IN SCENE CLASSIFICATION

### 6.1 Large intraclass variation

Intraclass variation is originated mainly from intrinsic factors of the scene and conditions of imaging itself. In terms of the factors, each scene can have many different example of images, possibly differing with large variations among various different objects, human activities, or background. Conditions of Imaging like changes in viewpoint, illumination, scale and heavy occlusion, shading, clutter, motion, blur, etc. are major factors in contributing to large intraclass variations. Further 3 challenges may be added by filtering distortion, poor resolution, noise corruption, and digitization artifacts. For example, three hotel rooms can be shown with different viewing angle, lighting conditions, and objects.

### 6.2 Semantic ambiguity

Since images of different classes may share similar objects, background, textures, etc., they look very alike in visual looks, which causes ambiguity among them. There is a strong visual relation between three different indoor scenes, which are archive, library, hotel rooms and bookstore. The problem of semantic ambiguity would be more serious with the emergence of new scene categories. Adding to it, scene category annotation relies on the experience of the annotators, therefore a scene image could belong to multiple semantic categories.

### 6.3 Computational efficiency

The popularity of social media sites and mobile devices has led to increase in demand for various different and interesting computer vision tasks including scene recognition. However, what makes an efficient scene recognition a pressing requirement is that mobile devices have constrained on computing related resources.

## 7 DIFFERENT ALGORITHMS USED FOR SCENE CLASSIFICATION

### 7.1 Object-Centric CNNs

Object-centric CNNs extract features which are local to the image from different regions of the image scene. After detecting the objects in the image, machine learning algorithms are used to further classify the image into its specific classes. An important factor is the relational size of images in the source and target datasets, in the deployment of object-centric CNNs. Although CNNs are usually robust against scale and size, because such models are originally pre-trained on datasets to detect and/or recognize objects the performance of object-centric CNNs is influenced by scaling.

### 7.2 Scene-Centric CNNs

Scene-centric CNNs extract the representation scales in the whole range. A CNN is trained to classify the images directly without producing annotations. We allow the CNN to identify the features itself rather than explicitly training it to identify features and then classify it into scenes.

## 8 METHODOLOGY

### 8.1 Workflow

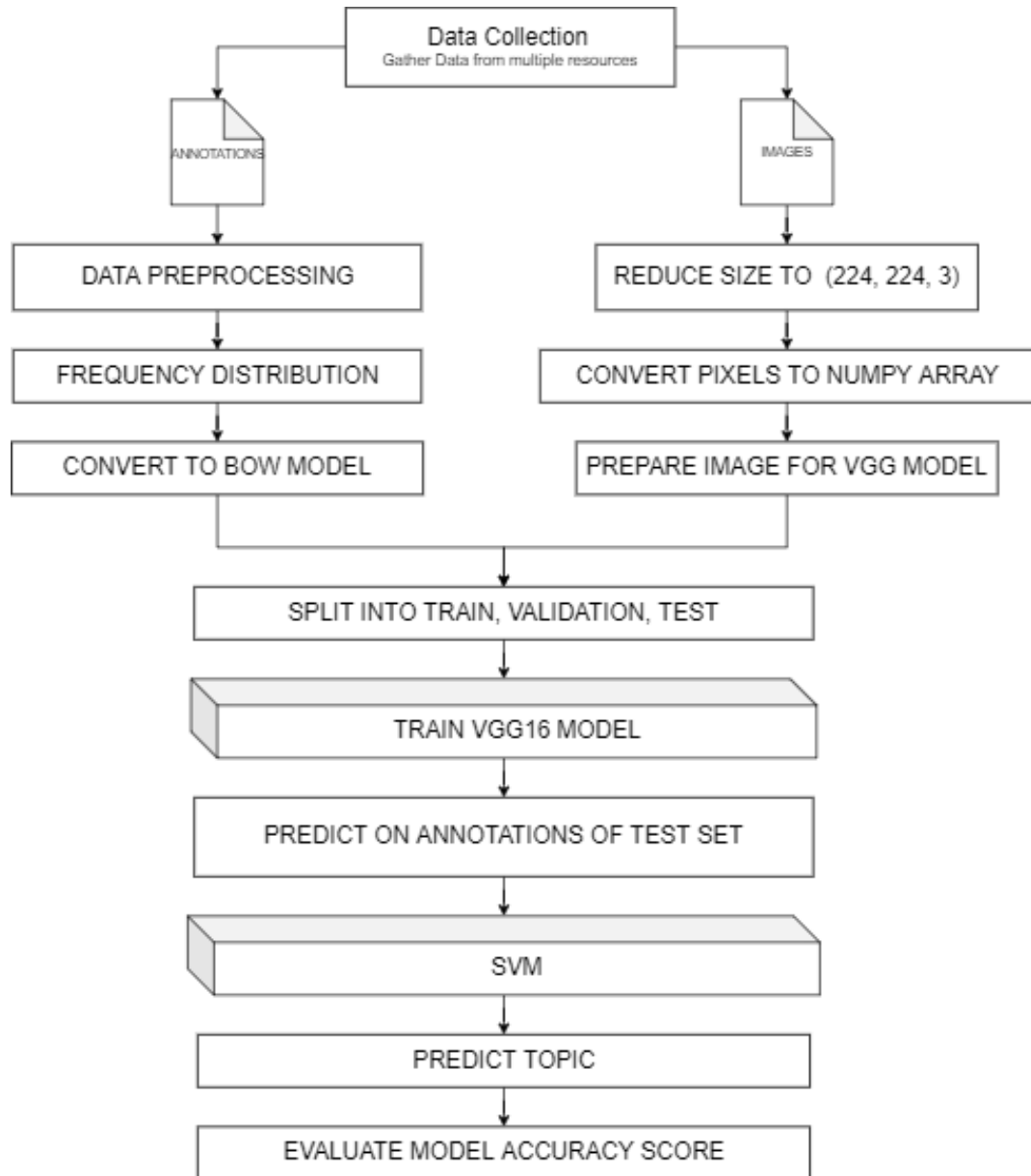


Figure 1: Object-Centric Workflow(VGG16+SVM)



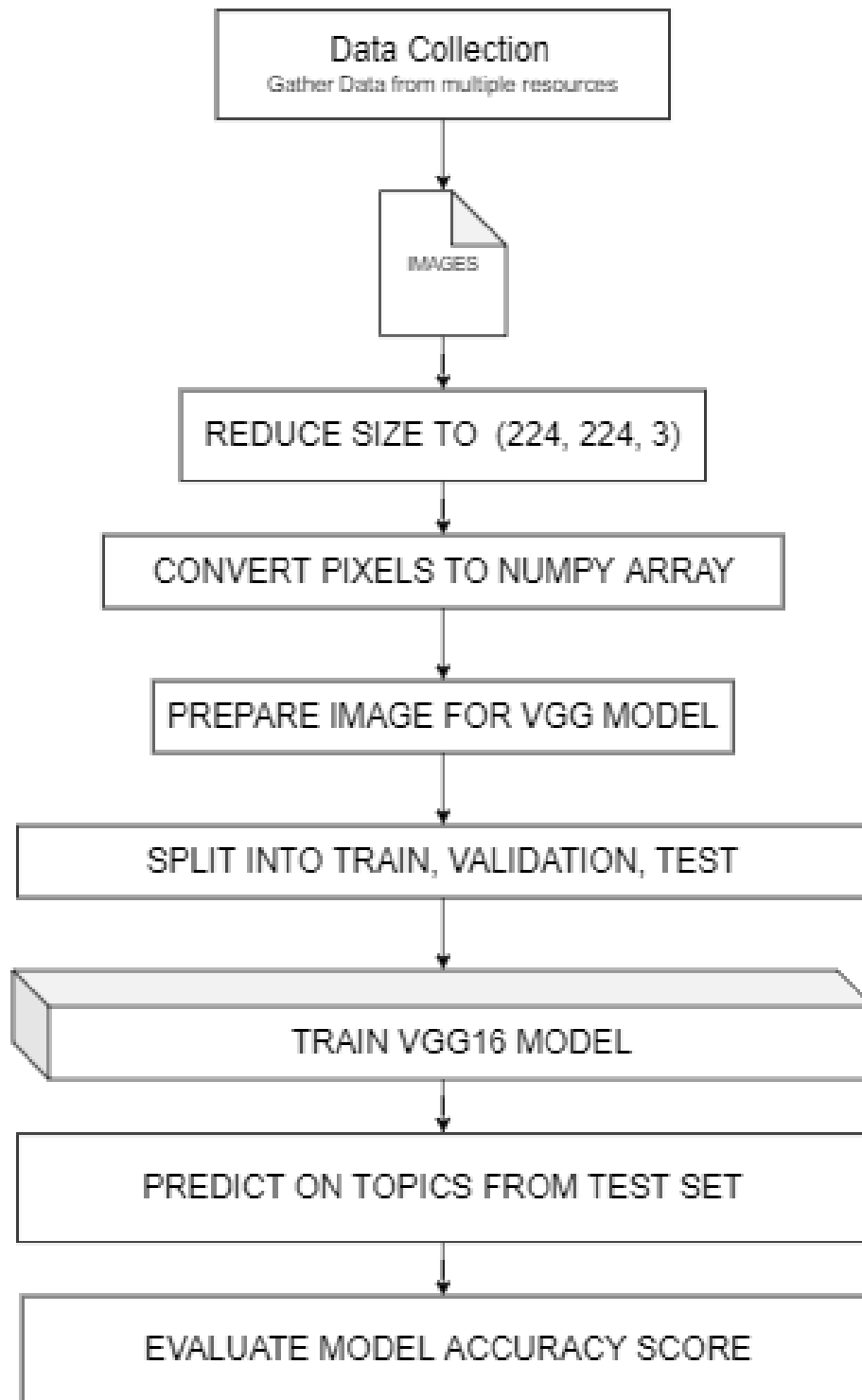


Figure 2: Scene-Centric Workflow(VGG16)

## 8.2 Mathematical Model

### Performance Metrics:

To perform real-world analysis of scene classification of varying images using deep learning models performance metrics like log-loss, accuracy and confusion matrix are used.

### Log Loss Score:

This is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of a logistic model that returns  $y_{\text{pred}}$  probabilities for its training data  $y_{\text{true}}$ . The more the predicted probability diverges from the actual value, the higher is the log-loss value.

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$$

### Precision:

The precision is the ratio true positives divided by the total number of positive predictions(i.e, the number of true positives plus the number of false positives). Precision tells us about how accurate the scene is classified from the images and out of them how many of them are actual positive.

$$Precision = \frac{TP}{TP + FP}$$

### Recall:

Recall is a metric that quantifies the number of correct positive predictions made out of all the positive predictions that could have been made. Increase in recall represents the less false negative and vi ca versa

$$Precision = \frac{TP}{TP + FP}$$

### F1 Score:

F1-Score is the harmonic mean of precision and recall. F1-measure can have best value as 1 and worst value as 0.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

### Accuracy Score:

Accuracy is calculated as the ratio of correctly predicted outcome to total outcome.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

### 8.3 VGG Model Architecture

VGG16 is a convolution neural network (CNN ) architecture which won the ILSVR(Imagenet) competition in 2014. The 16 in VGG16 means it has 16 layers that have weights. Till date it is one of the most powerful vision model architecture. The most important thing about VGG16 is that instead of having a huge number of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and used always with the same padding and maxpool layer of the 2x2 filter of the stride 2. Throughout the whole architecture this arrangement of convolution and max pool layers is followed . At the end it has 2 fully connected layers followed by a softmax for output. It has around approximately 138 million parameters which makes the network huge.

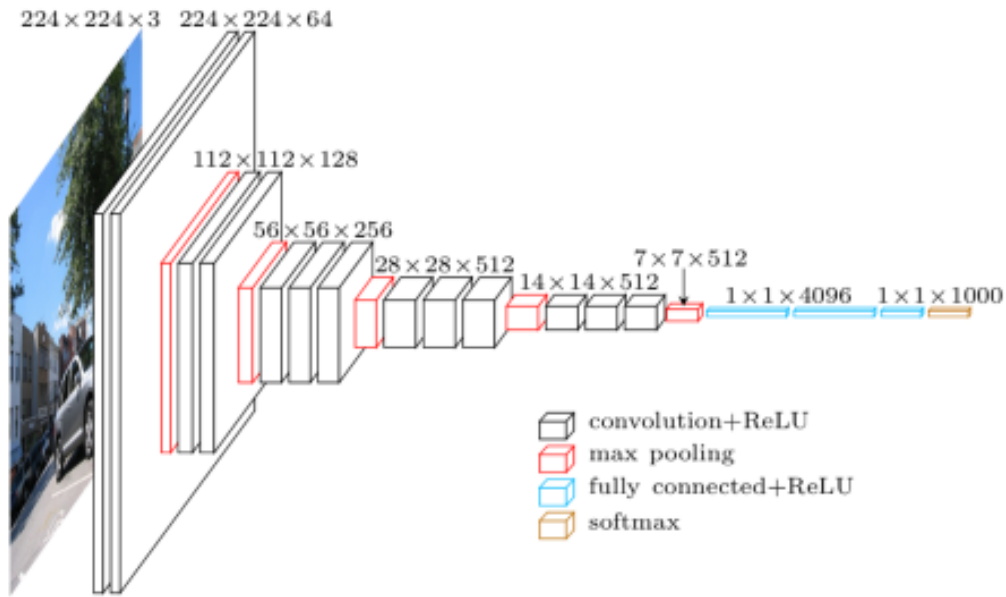


Figure 3: Architecture of VGG16 (Source:<https://neurohive.io/en/popular-networks/vgg16/>)

I will use the pre-trained VGGNet16 model for image processing that is trained on the Imagenet dataset provided in Keras. For classification ImageNet is used. Greater than 14 million images is contained in the dataset, with as upto 21 thousand groups or classes. We can modify the VGGNet16 model to fit our needs. We can remove the softmax layer and attach the below layers.

- Dense layer with 2056 units and ‘tanh’ activation
- Dropout layers with 0.5 percentage
- Dense layer with 1024 units with ‘tanh’ activation
- Dropout layers with 0.5 percentage
- Softmax layer with units of an optimal number of topics- 20 for Object-centric approach and 8 for Scene-centric Approach

## 9 RESULTS

### 9.1 Evaluation Of The Models

The Object-Centric Model(VGG16 + SVM) performed as follows:

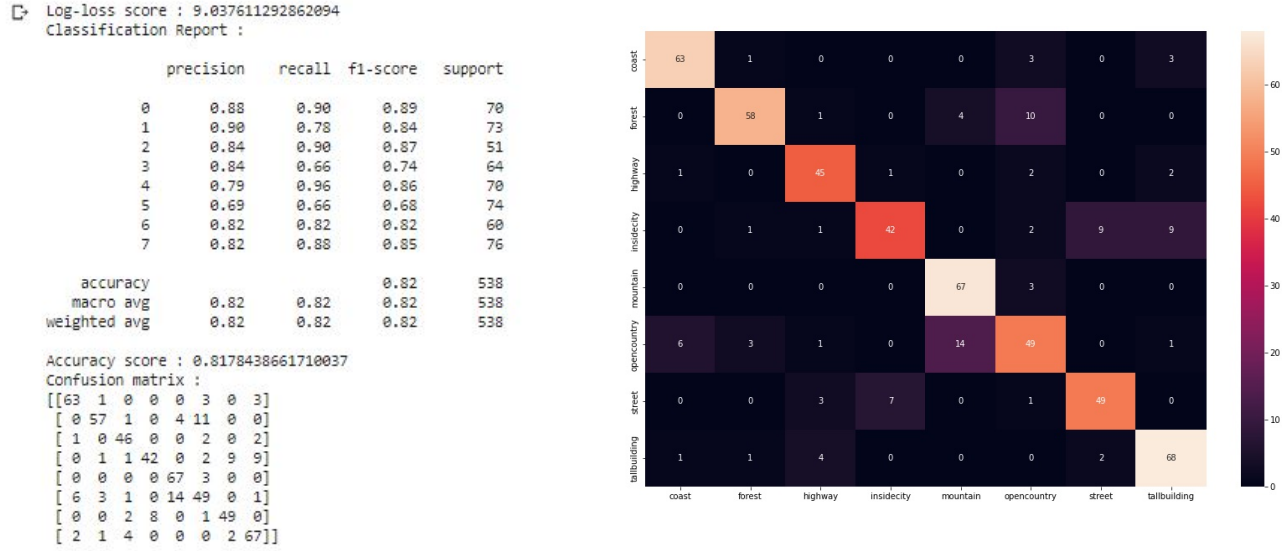


Figure 4: Model evaluation of Object-Centered (VGG16 + SVM)

The Scene-Centric Model(VGG16) performed as follows:

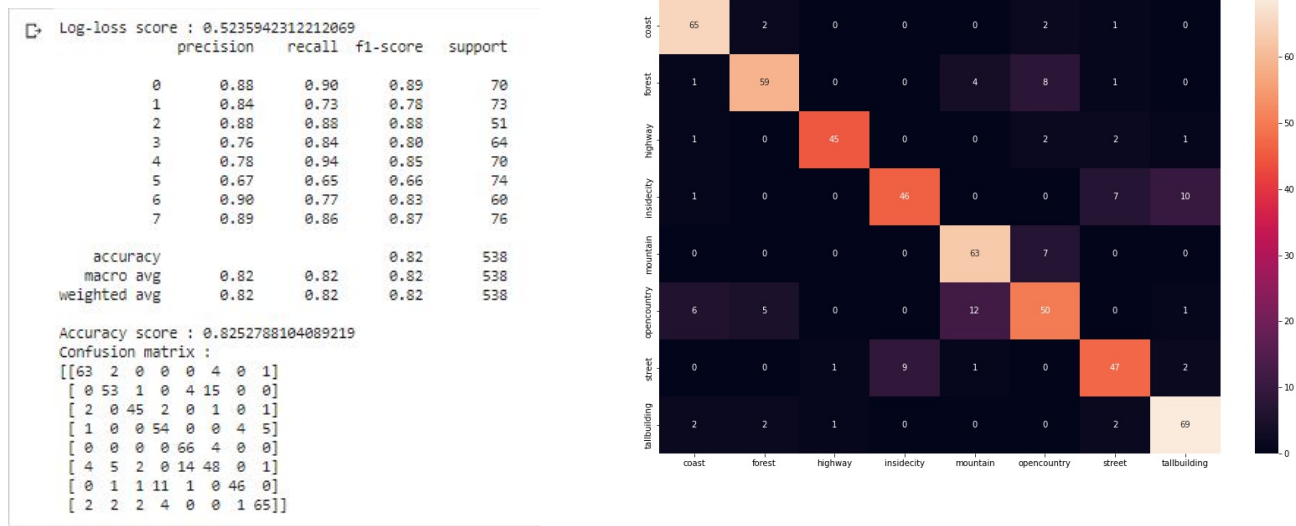


Figure 5: Model evaluation of Scene-Centered (VGG16)

## 9.2 Visualizing Predictions

### Correct Predictions:

The models successfully classified the scene for the images given in Figure 6.



Figure 6: Correct predictions

### Incorrect Predictions:

The models incorrectly classified the scene for the images given in Figure 7.



Figure 7: Incorrect predictions

## 10 CONCLUSION

The performance of both the models, object centric (VGG16+SVM) and scene centric (VGG16), were measured for the scene classification. The accuracy scores of both the models were calculated and compared. The Scene-Centric CNN got an accuracy score of 82.5% whereas the Object-centric CNN got an accuracy score of 81.9%.

Hence it can be concluded that the scene-centric model performed better than the object-centric. The reason being that scene-centric CNNs can capture more detailed scene information, namely local semantic regions and fine-scale objects, which is important to differentiate the scenes whereas object-centric CNNs are influenced by scaling because such models are originally pre-trained to detect objects on different datasets.












## References

- [1] Chong, Wang, David Blei, and Fei-Fei Li. "Simultaneous image classification and annotation." 2009 IEEE Conference on computer vision and pattern recognition. *IEEE*, 2009..
- [2] Fitzpatrick, Paul. "Indoor/outdoor scene classification project." Pattern Recognition and Analysis (2015).
- [3] Zhou, Bolei, et al. "Learning deep features for scene recognition using places database." (2014).
- [4] Zhou, Bolei, et al. "Object detectors emerge in deep scene cnns." arXiv preprint arXiv:1412.6856 (2014).
- [5] Herranz, Luis, Shuqiang Jiang, and Xiangyang Li. "Scene recognition with cnns: objects, scales and dataset bias." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
- [6] López-Cifuentes, Alejandro, et al. "Semantic-aware scene recognition." Pattern Recognition 102 (2020): 107256.

## Document Information

<b>Analyzed document</b>	31468_STC_Report.pdf (D119147067)
<b>Submitted</b>	2021-11-19 11:27:00
<b>Submitted by</b>	Rekha Kulkarni
<b>Submitter email</b>	rakulkarni@pict.edu
<b>Similarity</b>	15%
<b>Analysis address</b>	rakulkarni.pict@analysis.urkund.com

## Sources included in the report

<b>W</b>	URL: <a href="https://cupdf.com/document/adbms-seminar-report.html">https://cupdf.com/document/adbms-seminar-report.html</a> Fetched: 2021-11-19 11:35:00		2
<b>W</b>	URL: <a href="https://www.semanticscholar.org/paper/Scene-Recognition-with-CNNs%253A-Objects%252C-Scales-and-Herranz-Jiang/99864b6f513ee34032e20e9b8cb8db11a7ee2db2">https://www.semanticscholar.org/paper/Scene-Recognition-with-CNNs%253A-Objects%252C-Scales-and-Herranz-Jiang/99864b6f513ee34032e20e9b8cb8db11a7ee2db2</a> Fetched: 2021-11-19 11:35:00		1
<b>W</b>	URL: <a href="http://vision.stanford.edu/pdf/WangBleiFei-Fei_CVPR2009.pdf">http://vision.stanford.edu/pdf/WangBleiFei-Fei_CVPR2009.pdf</a> Fetched: 2021-11-19 11:35:00		2
<b>W</b>	URL: <a href="https://paperswithcode.com/task/scene-classification">https://paperswithcode.com/task/scene-classification</a> Fetched: 2021-11-19 11:35:00		1
<b>W</b>	URL: <a href="https://arxiv.org/abs/1801.06867">https://arxiv.org/abs/1801.06867</a> Fetched: 2021-11-19 11:35:00		2
<b>W</b>	URL: <a href="https://people.csail.mit.edu/khosla/papers/iclr2015_zhou.pdf">https://people.csail.mit.edu/khosla/papers/iclr2015_zhou.pdf</a> Fetched: 2021-11-19 11:35:00		3
<b>W</b>	URL: <a href="http://places.csail.mit.edu/places_NIPS14.pdf">http://places.csail.mit.edu/places_NIPS14.pdf</a> Fetched: 2021-11-19 11:35:00		2
<b>W</b>	URL: <a href="https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Herranz_Scene_Recognition_With_CVPR_2016_paper.pdf">https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Herranz_Scene_Recognition_With_CVPR_2016_paper.pdf</a> Fetched: 2021-11-19 11:35:00		2
<b>W</b>	URL: <a href="https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c">https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c</a> Fetched: 2021-11-19 11:35:00		1
<b>W</b>	URL: <a href="https://www.ripublication.com/ijaer18/ijaerv13n10_74.pdf">https://www.ripublication.com/ijaer18/ijaerv13n10_74.pdf</a> Fetched: 2021-11-19 11:35:00		1
<b>W</b>	URL: <a href="https://openaccess.thecvf.com/content_cvpr_2016/html/Herranz_Scene_Recognition_With_CVPR_2016_paper.html">https://openaccess.thecvf.com/content_cvpr_2016/html/Herranz_Scene_Recognition_With_CVPR_2016_paper.html</a> Fetched: 2021-11-19 11:35:00		1