Grenze International Journal of Engineering and Technology, June Issue

Indoor and Outdoor Scene Recognition

Nitin Kumar[1], Dr. Hitesh Singh[2], Ms. Tanya Varshney[3], Mr Vikrant Malik[4] and Dr Vivek Kumar[5]
[1-5]Department of Computer Science & Engineering, Noida Institute of Engineering and Technology, India
Email: ntnkmr4444@gmail.com, hitesh.singh.85@gmail.com, tanyavarshney03@gmail.com, vikrant.malik@niet.co.in, vivek.bansal1977@gmail.com

***Abstract*—In computer vision recognition of scenes is a long-time research problem. Scene can be defined as the real-time environment view which consists of a lot of views (like road, tree, building, parks, etc.) in a meaningful manner. The problem of scene recognition can be explained as assessment of labels such as "road", "building", "hall", "bedroom" or in an extra simplified way "Indoor scene" and "Outdoor scenes is classified on an input image based on the object or environment of the image. A huge amount of data is created and available every second in this growing era of digital data. Scene recognition is still a rising area that did not attain much success as compared to image recognition due to the vast variability of features in the scenic environment. Because of this reason, there is not a lot of work being completed these days in this area. This project focuses on the evaluation of problem statements using all the literature surveys completed lately and offering solutions for that problem statement. For resolving the proposed problem declaration ResNet and VGG variants are used ResNet18, 50 and 152 & VGG16, 19, and VGG19 with batch normalization are implemented. The entire scene recognition procedure is discussed in this report and the main motive is to form a foundation that can be further continued in proposing a new algorithm. Before deep learning the design and implementation of the scene recognition model depended on the low dimensional portrayal of the scene. The utilization of deep learning especially CNN for scene recognition has gotten extraordinary attention from the computer vision community.**

***Index Terms*— Scene recognition, image processing.**

I. INTRODUCTION

A. Theoretical Background

Visualizing an environment with a single glimpse is the ability of the human brain to process data and understand data at an impressive speed of just a few milliseconds. In addition to being exposed to the complex and rich diversity of natural images, one important factor for our brains is its ability to organize sequence in a hierarchical manner of increasingly complex operations, a structure that inspired Convolutional Neural Networks or CNN. In computer vision applications scene recognition is a long-established research problem and is widely used in autonomous systems and robotic systems. Recognition is the acknowledgment or identification of something. In artificial intelligence, we talk about scene recognition, object recognition, and image recognition all these techniques have so much in common, but the only comparison between them is when we focus on the known things like the objects, we call it object recognition and when we focus on the environment or nature in the image is called the scene recognition. Scene recognition is an area of visual recognition where we recognize and identifies the scene of the image means we identify what the image as a whole is about or where it is taken.

Grenze ID: 01.GIJET.8.2.18

We don’t classify a certain object in the picture but tries to recognize the whole scene of the picture. Scene recognition can be defined as a problem when evaluating labels such as “street”, “building”, “hall”, “bedroom” or more simply “Indoor scenes” and “outdoor scenes are done on an input image based on the content or nature of the image.

And if we talk about robots, they have made great advances, however, in artificial intelligence applications they are widely used. But knowledge of just image recognition is not sufficient for them, they need to understand more than this which will make it grow more. There are two different types of scenes available in nature, the first is a real-time and artificial environment. All-natural scenes are subject to real-time scenes, and all indoor scenes are subject to artificial natural scenes. A scene is an aggregate of semantically related visual entities in the case of a visual scene in an image and a scene in a video. In a belief network or another applicable graphical model, a scene could be modelled as a latent variable, where the visual entities are observable variables. In this case, recognition of a scene with a certain likelihood is used to influence the likelihood of an observed visual entity. In the case of videos, we can say that a scene is a collection of contiguous video shots means a collection of video frames that comprise a distinct concept in the video, that is they correspond to a semantic segmentation of the description of the video. Nowadays, the available visual data or the trouble with understanding and the input in recognition of the scene can be in a static layout as a picture and in a dynamic as video recording. Previously before deep learning the implementation and application of the scene recognition model was based on a very low-level dimensional representation of scenes. The use of deep learning, the convolution neural networks of spatial recognition has received a great deal of awareness in the community of computer vision.

Simply we can say that scene is background in an image. An image consists of the foreground which is a relevant visual object and the background. In the real 4 world, there is a correlation between background and foreground, which can be exploited to predict the likelihood of one based on the recognition of others. Some application of scene recognition is listed below. • In mobile phones, there be a rapid growth of camera-based applications, where understanding location is a key part. • Automatic mode cameras, can recognize scene categories and then adjust the camera parameters to take the best shot. • It is used to classify the image content into semantically meaningful entities in order to understand the image content. • Compared to object recognition the scene recognition performance has not achieved success to such an extent due to its wide variety of features because not much work has been done before. • We can use scene recognition in many applications from mobile robotics to navigation.

B. Objective for Research

Visually disabled people have their own style of doing things and lives an ordinary life with. But due to inaccessible infrastructure and social challenges they definitely face troubles. To navigate around places is the biggest challenge for a blind person, especially the one with complete loss of vision. They know the position of everything in the house therefore they can easily roam in their house without any help. But they have difficulty in recognizing scenes when they are in other places. So, we decided to make an indoor and outdoor scene 5 recognition System. We are interested in this project after we went through few papers in this area. As a result, we are highly motivated to develop a system that recognizes the scene in the real-time environment.

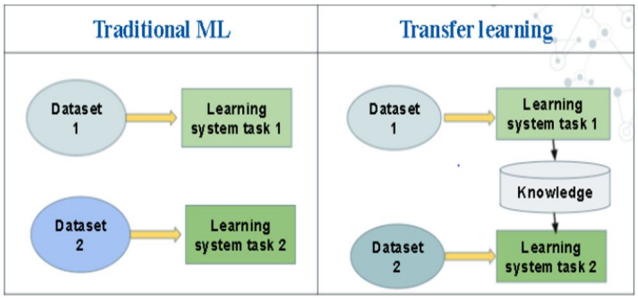
Scene recognition is still a developing area in computer vision there are multiple applications of scene recognition. Due to so much variability in the environment, this becomes quite difficult to recognize the background scene of the image. Still, we have so much more to explore in this area, as data is generated day by day at a rapid rate it becomes important to recognizing and understand the features in that data. Several recognition techniques work well with image data, but in the case of scene recognition, it is not giving promising results due to less variability.

C. Preliminaries

Transfer learning

This preliminary segment introduces the models and algorithms used in this project. For implementation, we used the transfer learning technique. So far, the algorithms traditionally designed for machine learning and deep learning techniques are work in isolation. For resolving particular tasks these algorithms are trained. The models are rebuilt from scratch once the feature-space distribution changes. Transfer learning is the concept of making use of knowledge obtained for one task to resolve associated ones and to overcoming the isolated learning paradigm. Transfer learning isn't a new idea, this is very particular to deep learning. The conventional techniques of training and building a machine learning models is different from using transfer learning principle following method. Traditional learning is remoted and takes place simply based totally on particular on training separate model on dataset and tasks. There will be no knowledge present which used from one model and transferred to

another. In transfer learning, for training the latest models we can reuse information from previously trained models.



Here we can utilize the understanding from previously learned tasks and perform them to newer tasks. Other than this we can generalizes the knowledge from the first task which considerably have greater data and use of its learning for the second task (with appreciably much less data). To resolve the issue of computer vision domain, where features like shape, edge, corners and intensity which are low-level features are shared all through tasks, to resolving this permit transfer of knowledge among tasks. As we've discussed, knowledge from a present task will acts as a further input while learning a new aim task. There are three types of transfer models. • Train the whole model whilst the dataset is large but isn't the same as the pre-trained version's dataset. • Train a few and leave other frozen when the dataset is massive similar to dataset of pretrained model or when we have small dataset different from the pre-trained dataset.

• Lastly, keep the full model freeze feed the dense layer according to the classifier wherein we've got the small dataset that is same as pre-trained dataset. We are only training some layers and leave others frozen where the subsampling triple layer which is extremely complex and having millions of parameters and extracts features, is left alone at the value obtained during the pretraining. While the parameters of fully connected layers are modified by the use of a small set of new images. Transfer learning is modifying just a few layers of the model, so it is going to take less time as compared to training all the layers, and it may be accelerated by a GPU. It enables fine-tuning a pre-existing model to work for tasks that do not have a lot of data available. Pretrained models are used in transfer learning techniques the choice of using the pre-trained model relies upon the size of the dataset and the number of computational resources available. These models include • Fine-tuning in this the final classifier layer of a network is swapped out and replaced with a SoftMax layer the right size to fit the current dataset while keeping the learned parameters of all other layers. • Freezing for smaller dataset it is common to freeze some first layers of the network, meaning the parameters of the pre-trained network are not modified in these layers. The other layers are trained on the new task as before. • Feature extraction is the loosest usage of pre-trained networks.

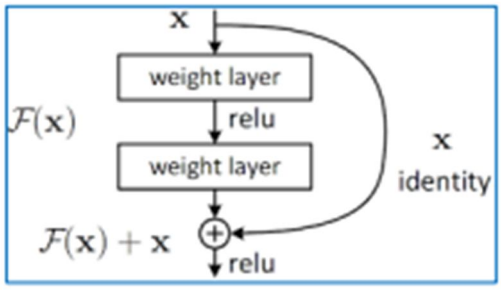
ResNet

ResNet is a short name for the residual network. It allowed us to train an extraordinarily deep neural network with more than 150 layers. ResNet is essentially a pre-trained CNN model, and it helps to train the model with computer vision tasks. In residual learning, we try to learn some residual rather than attempting to learn a few features, we can say to characteristics learned from the input of that layer residual perform subtraction of those characters. Training a simple deep convolutional neural network is very hard while this form of network is easier to train and additionally, the network solves the problem of degrading accuracy. As we are aware deep networks without adjustment are regularly afflicted by vanishing gradient which means as the model back propagates, the gradient gets smaller and smaller. But ResNet makes that change by using the skip connection. Learning can be made intractable by a tiny gradient. Skip connection is nothing but the process of jumping from one convolution layer to another by skipping many layers that are present in between them. This helps CNN to escape from the vanishing gradient problem and to give an identical mapping from input and output. The major idea behind implementing this is to have a mapping like.

$$X \rightarrow F(x) + G(x)$$

Here x represents the input, F(x) is output, and G(x) is the residual. When the input is exactly the same the output G(x) will be zero. The advantage of including the skip connection is that if any layer hurts the

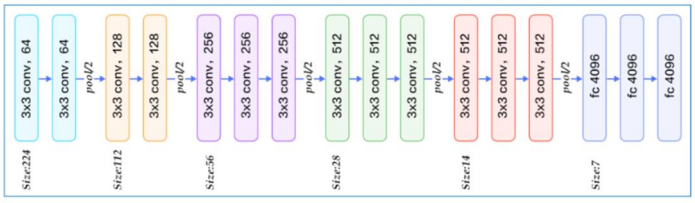
performance of the architecture, then it will simply be skipped by regularization. This results in training a very deep neural network without the problem caused by the vanishing gradient.



The model is confirmed to be better for image recognition because it contains a deep architecture with over 23 million trainable parameters. Constructing a model from scratch in which we need to acquire a very huge amount of data and train that ourselves, the use of a pre-trained model is a highly effective approach compared to that. ResNet gives a framework that permits training thousands of layers whilst preserving overall performance. ResNet boosts computer vision functions like object detection and face recognition due to the fact of its effective representational ability. ResNet-50 is a beneficial tool to use and stated for its outstanding generalization on the task of recognition having overall performance with fewer error rates. In ResNet models, all convolutional layers follow the same convolutional window of dimension 3x3, the number of filters will increase following the depth of networks, from 64 to 512 (for ResNet 18 and ResNet 34) from 64 to 2048 (for ResNet50 and ResNet 152). In all models, there is only one max pooling layer with pooling dimension 3x3, and a stride of two is applied after the first layer, consequently decreasing the resolution of the input is significantly restricted during the training process. The input size of the image in this networkis 224×224 at the very beginning we have a convolution layer that contains a kernel of dimension 7×7 and stride 2. At the end of all models, the average pooling layer is applied to replace fully connected layers. This substitute has some advantages. Firstly, there is no parameter to optimize in this layer, therefore it helps to minimize the model complexity. Secondly, this layer is extra native to enforce the correspondence between feature maps and classes. The output layer has a thousand neurons which are corresponding to the range of classes in the ImageNet dataset. Besides, a SoftMax activation function is applied on this layer to provide the probability that the input belonging to every class. ResNet18 is eighteen layers deep CNN, which is trained on more than one million pictures of the ImageNet database. In this, the images can be categories into a thousand object categories it is very beneficial and efficient in image classification. It has only 2 pooling layers one at the beginning and another at the end. Furthermore, it follows 3x3 convolution layers. It also follows a batch normalization process at every layer where the input will be normalized for each batch.

VGG network

VGG has been invented through the Visual Geometry Community. The fundamental contribution of VGG is that it is using small receptive fields in the layers. By increasing the depth of CNN, we can improve the classification accuracy of the network. Earlier Neural networks used the larger receptive field (7x7 and 11x11) while VGG smaller receptive field of multiple 3×3 ones after another. Having multiple small-size kernels stacked on top of each other is better than having one large-size kernel because as this depth of the network increases it enables the model to learn more complex features, and that too at a very low cost. There are 3 fully connected layers in VGG, the primary fully connected layers, every with 4096 nodes and the last one with a thousand nodes, one for each class, and finally a SoftMax classifier.



The input image dimension is fixed to (244×244) for the architecture. In the step of pre-processing, the suggested value of RGB is excluded from image every pixel. After pre-processing is done, the image is passed to a stack of the convolutional layer with small (3x3) receptive field. The size of the filter is set to (1×1) in few

configurations. The stride of the convolution operation is constant as spatial pooling performed through 5 layers of max-pooling. Max pooling is performed over a pixel window of size (2x2) with stride size set to 2. A fully connected layer configuration is always similar to the first fully connected layers, each with 4096 nodes and the 3 rd one with a thousand nodes, and SoftMax layer is the final layer. In VGG network ReLu activation function is connected to all the hidden layers. All configurations differ only in-depth and follow the universal pattern of architecture, from eleven weight layers to 19 weight layers. VGG16 is a lot deeper which consists of sixteen weight layers which include 13 convolutional layers with filter dimension 3×3 and a fully connected layer consisting of six filters same as the convolution layer. For spatial resolution is preserved after convolution, all convolution layers have stride and padding fixed to 1 pixel. All convolution layers are divided into a group of five and a maxpooling layer follows each. VGG16 andVGG19 are well-known in all variations of VGG. VGG19 is consists of nineteen layers. Pre-trained network which is trained on the ImageNet dataset having more than a million images, and we can load and use these pre-trained networks. Using these pertained networks, we are able to classify images into a thousand object categories.

II. Literature Review

Bolei Zhou et al. [3] They work on a problem that scene recognition has not achieved much success because the features are trained on ImageNet and are not compatible enough for the scene recognition task. For solving this problem, they introduced a place database which is a scene-centric database with 7 million label images of a scene with 476 place categories. Their method compared the diversity as well as density of available image datasets and show that their proposed dataset which is place is as denser as other available scene data, and it has more variety. Their introduced dataset is 60 times larger than the sun dataset using this they compare the accuracies with a different dataset. For learning the features, they use CNN and establish their new result on several other scene-centric datasets. After visualizing the deep features, the result, represents the difference in scenecentric and object-centric network internal representations Places-CNN performs much better than another available database.

Mouna Afif [1]. proposed a technique for indoor scene recognition using a common object as an intermediate semantic representation which is a probabilistic hierarchical model. They used an object classifier for associating low-level features to an object and a contextual relation to the associated object to the scene. By including the geometrical information from the threedimensional distance sensor, their model improved the performance of the current category-level object classifier. They used two different indoor environments and four different scenes or places for computing the probability distribution of their model. Seven objects are used to estimate the place probability and sliding window approach is used that consider 5 different window shape. There are some 8 limitations of their model it outperforms if testing and training data are not from the specific indoor environment.

Priya Singla, Rajesh Mehra [10] the issue focus is that in scene classification recognizing tall buildings and mountains is difficult. They applied a combination feature extraction algorithm which is a hybrid combination of two algorithms SIFT and HOG (HFCNN). From all the features descriptor shift and hog are reliable and most widely used for feature detection in scene recognition. Where scene invariant feature transform (SIFT) is a local feature extraction model used to describe the key points of an input image while (HOG) is a global feature extraction model that describes the whole feature vector of an image works in sliding window technique. This hybrid model performs better with CNN architecture, and it works for indoor and outdoor datasets they upload an image and convert it in grayscale and to remove distortion they use filters. Their model shows impressive results it improves the accuracy rate of the overall model by 96% and the time consumption became lower 0.15-0.20.

Bolei Zhou, Agata Lapedriza, et al. [4] This paper describes a database called place which is a repository of about 10 million scene pictures, labelled which has 434 categories of the semantic scene, which consisted of a wide environment which is the different environment of about 98% of the types of scenes which humans visualize in the world. They used CNN as a baseline for scene classification which is a place CNN that outperforms the existing previous approach. They also compared that CNN is used to learn the features for the classification of scenes.

Espinace, T. Kollar.al. [2] They proposed a new variable model which is a latent variable. This scene can be represented as a collective of location models these collections are organized in a sample which can be configured again. An 9 image is divided into some predefined regions set, and they can assign to each image region using the latent variable. For capturing the appearance of a picture, they use a bag of words (BOW). A spatial bow technique is generalized which depends on the model for BOW in region image. To train their model

they used discriminative and generative approaches. In the generative to estimate or train model parameter from category labelled collection of images, they use the Expectation-Maximization (EM). In the
discriminative approach, they used a latent structural
SVM (LSSVM). As compared to EM training, LSSVM is more sensitive to initialization. They achieve the best performance from their model which is a combined approach of LSSVM and EM.
S. N. Parizi, J. G. Oberlin and P. F. Felzenszwalb, E. [5] developing a reliable and robust model for automatic recognition of scenes is an important field in AI. They describe an ensemble technique delivered by research papers in the area of scene recognition and present some reviews on their findings in which they describe all the challenges people faced in this field and future research line. They even give an overview of all the available datasets from the image as well as scene recognition. Furthermore, they aimed to be a future guide in the field of visual recognition for selecting models.
Sheng Guo, Weilin Huang, et al. [9] A(LS-DHM) is proposed which is a locally supervised deep hybrid model for effectively enhancing & exploring convolution features for recognition of scenes. They find that local objects can be captured by convolution feature and the structure of scene image yield the importance of discriminating ambiguous scenes. Their model computes multilevel deep features. Their proposed model is a local convolution supervision LCS which increases the features of the image by label information propagation on the convolution layer. The fisher convolution vector FCV is used to encode convolution features which are the large-sized convolution maps. Then they 10 combined both the LCS and FCV in the LSDHM representation which achieved outstanding performance in the experiment.
Since the problem statement that is being considered in this project is all about the variability or wide variety of features that are an important process of scene recognition. And other than this addressing the problem of an indoor scene. Focusing on all the limitations of previously proposed models here this report comes with a common solution or model which will be able to recognize a high variability area that has a wide variety of features, for example, university or tourist place, hospitals, etc. For training the model we will be using CNN which will produce audio output by converting an input image into a string of text and text to audio. And due to lack of integrated framework or portable device which can help in understanding and recognizing visual scene for visually impaired people. And to address the problem of the indoor scene the model exploits local and global discriminative information. This model will work as a navigator or assistance system which will help people with partial blindness or fully visually impaired to navigate indoor scenes. • To help navigate in an indoor place for people with partial or full visual impairment the system will work as an assistive system to help them recognize the surrounding. For addressing the problem of indoor scene recognition, this model will exploit the local as well as global discriminative information.

III. Conclusion

Scene recognition is the process in which we evaluate labels or input image data. This project proposed a solution where the suggested model will help in navigating the scene of a large square area which will also be useful for visually impaired people. Since indoor recognition is being difficult due to its spatial properties this model helped in solving this problem and helped in describing the global as well as local discriminative information. The input data is trained on ResNet and VGG network and produces output in the form of audio. Different variations of ResNet and VGG networks are tested using the same dataset and performance analysis. The outcome that ResNet152 has the highest accuracy among all with 88.59% test accuracy. Compared with the existing model on the same dataset this model shows the promising result with prediction indoor as well as an outdoor scene with high accuracy.

References

[1] Mouna Afif, Riadh Ayachi, Yahia Said, Mohamed Atri, “Deep Learning Based Application for Indoor Scene Recognition”, Springer Science+Business Media, LLC, part of Springer Nature 2020(20 march 2020).
[2] Espinace, T. Kollar, A. Soto, and N. Roy,” Indoor Scene Recognition Through Object Detection”,2010 IEEE International Conference on Robotics and Automation Anchorage Convention District (May 3-8, 2010).
[3] Bolei Zhou, Agata Lapedriza, “Places: A 10 million Image Database for Scene Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.40, pp.1452-1464, 04 July 2017.
[4] Bolei Zhou, Agata Lapedriza, “Learning Deep Features for Scene Recognition using Places Database”, (NIPS) Neural Information Processing System, 8-11th december 2014, Montreal, canada.
[5] S. N. Parizi, J. G. Oberlin and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012,pp.2775-2782,doi: 10.1109/CVPR.2012.6248001.

[6] N. Sun, X. Zhu, J. Liu and G. Han, "Indoor scene recognition based on deep learning and sparse representation," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 844-849, doi: 10.1109/FSKD.2017.8393385.
[7] Alina Matei, Andreea Glavan, and Estefan´ıa Talavera, “Deep Learning for Scene Recognition from Visual Data: A Survey”, Lecture Notes in Artificial Intelligence: LNAI-LNCS. arXiv, 2020.
[8] Ariadna Quattoni, Antonio Torralba, “Recognizing indoor scenes”IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–420. IEEE (2009)
[9] Sheng Guo, Weilin Huang, Limin Wang, Yu Qiao,”Locally-Supervised Deep Hybrid Model for Scene Recognition” IEEE Trans. on Image Processing, 2017, arXiv:1601.07576v2
[10] Priya Singla, Rajesh Mehra, “Scene Recognition using Significant Feature Detection Technique”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020
[11] Sun, N. et al. “Fusing Object Semantics and Deep Appearance Features for Scene Recognition.” IEEE Transactions on Circuits and Systems for Video Technology 29 (2019): 1715-1728.
[12] L. Herranz, S. Jiang, X. Li, "Scene recognition with CNNs: objects, scales and dataset bias", Proc. International Conference on Computer Vision and Pattern Recognition (CVPR16), Las Vegas, Nevada, USA, June 2016
[13] J. Gao, J. Yang, J. Zhang and M. Li, "Natural scene recognition based on Convolutional Neural Networks and Deep Boltzmann Machines," 2015 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 2015
[14] A. Quattoni and A. Torralba, "Recognizing indoor scenes," IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009.
[15] Boyi Hong, Cong Jin, Nansu Wang, Yajie Li, and Hongliang Wang “The application of transfer learning for scene recognition”, International Conference on Image and Video Processing, and Artificial Intelligence, 1132107 (27 November 2019).