Scene Recognition using Places Dataset

Sharmila Agnal.A, Jenny Savani, Suchhanda Das, Preetam Swaraj, Ayushi Rathi

Abstract: ***Artificial Intelligence is the major breakthrough in the field of computer science. Moreover, with the help of computer vision and image processing, analyzing and understanding digital images has become very easy. In this paper, we are constructing an Artificial Intelligence model which will detect the scene with maximum accuracy. At first, the machine is trained with various categories of landscapes with each and every image from the collection of labeled pictures from database. Training includes dividing the image into fine sub-regions and analyzing histograms using external and internal features present inside each sub-region. For this, we are using Places database, which is collection of millions of scenic images. These images will be labeled with semantic categories which comprises of a large and diverse list of the types of landscapes found in the world. Also, by using CNN (Convolutional Neural Networks), we learn deep features for scene recognition tasks, and establish several scenic-centric categories. Scene recognition provides simplistic visual sense which helps in understanding relationship between foreground and background in an image. Lastly, we intend to design a system which can build the proposed model using a database consisting of all the possible categories of landscapes.***

Index Terms: ***Artificial Intelligence, deep features, image processing, scene recognition***

I. INTRODUCTION

In computer science, digital image processing is the use of computer algorithms to perform image processing on digital images. Image recognition is the process of identifying and detecting an object or a feature in a digital image or video. In machine learning the image recognition involves extraction of image features from image and using them as input to the model. In deep learning, CNN (Convolutional Neural Network) may be used. In CNN the machine automatically learns from the sample images and automatically identifies the learned featured in new images. CNN requires little pre-processing compared to other image classification algorithms[1]. Network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human input in feature design is huge advantage.

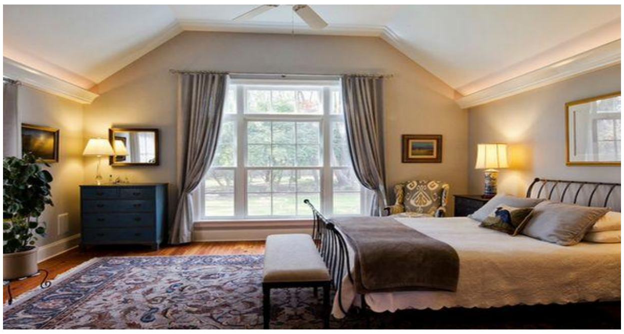

Fig. 1: Scene with different objects

When the above image (Fig. 1) is presented to anyone and asked what do they see? The answer will clearly be ‘A bedroom, probably inside a house’. Similar will be the answer of any visual recognition system. But this bedroom can be inside a house or in any resort/hotel. A visual system aims at identifying the object and recognizing event around that object. But along with knowing the object and event, knowing the place or context is equally important. A system with computer vision should be able to interpret the entire image or sight present in the frame. This image or sight consists of various objects as well as some sort of background, videlicet a ’scene’. A plethora of technologies have developed through which the system can identify the object and its surrounding, so that it can gain a certain level of understanding regarding what an image contains, just the manner a human does.

In order to have processing like a human brain, a system should be able to understand the entire environment depicted in the image. This gives a deep insight about what might have happened in past and also helps in predicting the results for future. Here is when scene recognition comes into the picture! It helps the system to build scene understanding by providing relationship between the different objects in an image and its surrounding as well.

Moreover, in order to recognize a scene just like a human brain, the system needs some primary knowledge regarding the possible landscapes. For this, we train the system by processing thousands of images of same category in order to cover all the possible scenes. Features extracted from these images are treated as points in high dimensional space and accordingly, database is generated. Also, to process and manage this large number of unknown inputs, we are using neural network. This neural network acts in the same manner as a human brain. It helps in learning the expected output from training datasets for a given input of images. Convolutional Neural Networks (CNNs) has successfully delivered tremendous results for object recognition tasks. In the same way it can be used for extracting diverse features from the scene[2]. In this paper, we are using Places

Revised Manuscript Received on April 06, 2019.

Sharmila Agnal.A, Computer Science & Engineering, SRM Institute of Science and Technology, Chennai, Country, India.

Jenny Savani, Computer Science & Engineering, SRM Institute of Science and Technology, Chennai, Country, India.

Suchhanda Das, Computer Science & Engineering, SRM Institute of Science and Technology, Chennai, Country, India.

Preetam Swaraj, Computer Science & Engineering, SRM Institute of Science and Technology, Chennai, Country, India.

.**Ayushi Rathi**, , Computer Science & Engineering, SRM Institute Of Science And Technology, Chennai, Country, India.

dataset, which is a collection of more than one million scene centric images. By using this dataset and CNN structure, we are comparing testing and training datasets. Moreover, we also carry out scene recognition using this deep learning model after the training of datasets. This method has come out to be effective and accurate while compared to existing systems.

A. Existing Systems

It is a challenging task to create a large-scale database of images and training. Current datasets mainly focus on accuracy and time span. There a lot of datasets developed till the date with different tools and algorithms. Most of them employ data collection scheme with Amazon Mechanical Turk. One such dataset is ImageNet, which offers collection of sorted images. This diversified dataset offers 50 million labelled full resolution images. While on the other hand, the small image datasets were used for general training and evaluations in computer vision algorithms. But again, the advancement in the computer vision more diverse and large-scale datasets. A dataset named Tiny Image was successful only with few applications because of the low resolution of images it provided. Another dataset named ESP dataset, which is not publicly available allows access of very few images and their labels. Another example can be LabelMe and Lotus datasets, which has only 200 categories of images along with outline and location of the object provided. In both of these datasets, the images are uploaded by the users of the datasets[3]. Also, Lotus is paid dataset.

B. Problem Statement

First and foremost, problem which arises with a large-scale dataset is of over fitting because of enlarged network. Size of the network increases due to increasing computational resources. Also, the process of training the dataset is very tedious and laborious. For some datasets low resolution of images makes it less suitable. Moreover, accuracy of identifying the correct image is what makes the system more effective. But most of the datasets lacks sense disambiguation and fails to identify the scene correctly. Few datasets take only user selected images as input and fail to cover images from entire internet [3].

Building a large-scale dataset is an intricate process and expansive as well. Because of this reason, few datasets are paid. Most of the dataset lags in versatility. They can be used for only certain applications.

II. METHODOLOGY

A. Proposed System

The strategy of Places is to cover all possible scenes encountered in environment. It aims at providing maximum accuracy and generating results within stipulated time period. It contains millions of images falling into diverse categories along with the labels. In order to make this large-scale dataset, we are using Convolutional Neural Network (CNN) which helps in achieving visual recognition.

B. System Architecture and Design

The following diagram of system architecture that we are proposing gives pictorial explanation about the process and modules involved in creating the database. For an example, an image and its histogram are shown, which is divided into small segments to extract the features out of it.

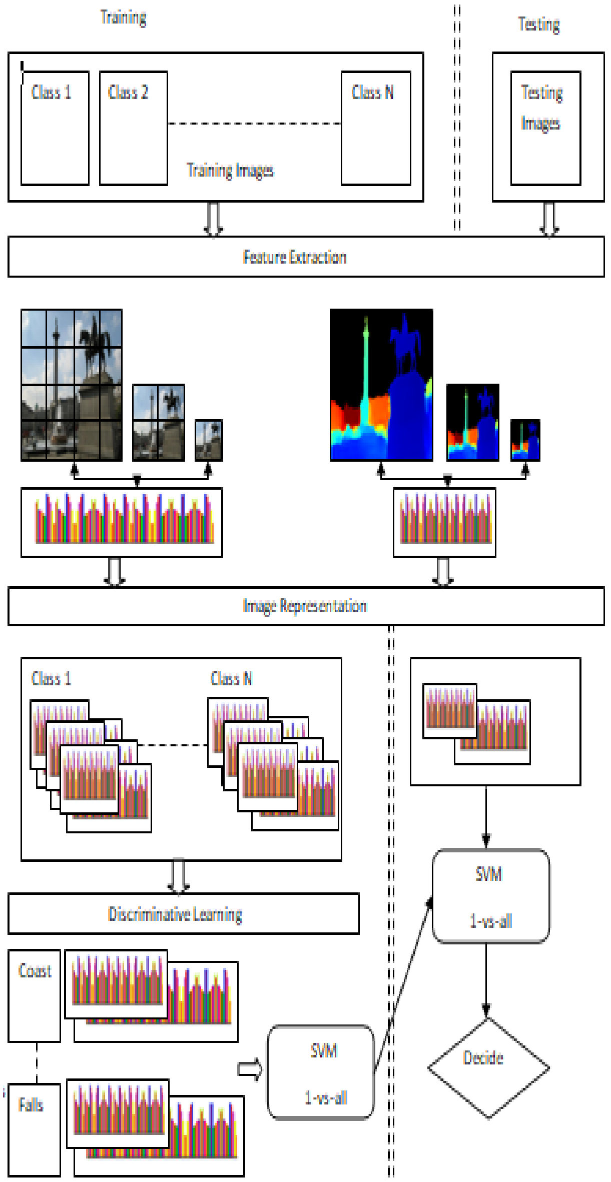

Fig. 2: System Architecture

Different modules of this system architecture are:

1. Training

In this step, we will use our data to incrementally improve our model's ability to predict. The training process involves initializing some random values attempting to predict the output with those values. We can compare our model's predictions with the output that it should produce, and adjust the values. This process then repeats each iteration or cycle of updating the values.

2. Feature Extraction

Features are the unique signatures of the given image or unique properties that defines an image. Features are extracted in order to differentiate between the images. Features of the respective images are then collected and forwarded for further processing.

3. Image Representation

After an image is segmented into regions[4], the resulting aggregate of segmented pixel is represented and described in different classes. External representation is chosen for shape characteristics. Internal representation is chosen for regional properties such as color and textures.

4. Discriminative Learning

After all the classes of features are created, the tags are then merged with the respective image class[5]. The test cases are then checked with all the classes and output is shown as per the best matching scenario.

C. System Requirement

A normal PC or laptop with Nvidia GPU of compute capability of 5 or higher is all what is required on hardware side. For software side we have used different tools and datasets mentioned below.

- Places Dataset
- TensorFlow
- CUDA
- Anaconda

1. Places Dataset

Selecting the dataset was the important part of this model. Since we are relying on CNN, the image set from which the machine is going to learn should contain variety of images. Places dataset given by MIT was the choice after considering its large size and collection distinctive images. The compressed version was used for this model but in theory the original dataset should also work on high configuration machines. Different pictures were classified in a single folder such as coast, mountains, etc. The machine was then trained on this dataset for 20000 steps to achieve the desired accuracy.

2. TensorFlow

TensorFlow is an open source machine learning framework by Google. With TensorFlow the process of acquiring data, training, predicting results and refining the results becomes easy to use. With TensorFlow the CNN can be implemented easily for image classification. Due to all these features TensorFlow was used for this work which allowed us to look into our works carefully with implementing the required features[6].

3. CUDA

CUDA (Compute Unified Device Architecture) is created by Nvidia which is a parallel computing platform and programming model. These are produced by the Graphical Processing Unit produced by them. This was required by TensorFlow so the computing of numbers can be done by the GPU for much faster and accurate results[7].

4. Anaconda

Anaconda is a Python distribution tool used for this project. It provided us with all the packages necessary in AI. The project was tested in Anaconda environment with imported libraries of TensorFlow. The training and result were produced in Anaconda environment itself.

III. Results

The practical implementation of the proposed system has been successful in providing 90% accuracy in identifying the correct scene. For sample purpose, the system was made to encounter 10000 images including various scenes of sea, mountains, desert, and forests. The images were identified correctly within stipulated time period. The histograms generated in each process were noted and evaluated.

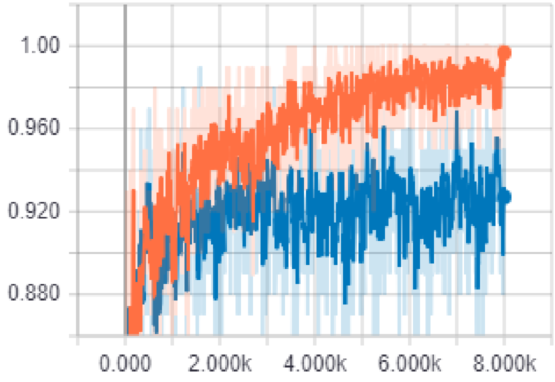

Fig. 3: Accuracy

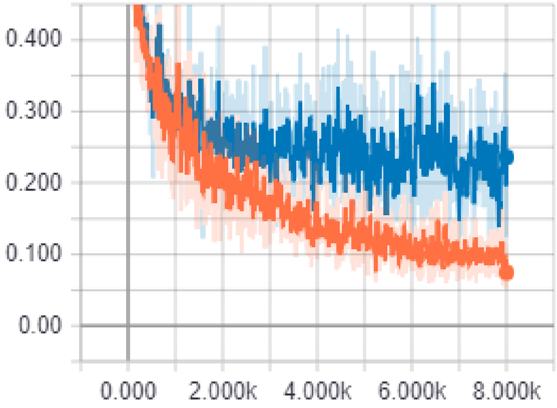

Fig. 4: Cross Entropy

The above two picture (Fig. 3 and Fig. 4) shows the accuracy and cross entropy of training and validation data (Fig. 5 and Fig. 6). The training data is represented in orange while the validation is represented in blue. We see that both the test and validation data are above 0.9 which means accuracy achieved is above 90%. The cross entropy has reduced with the steps which means that the machine will be able to show results of any test image faster.

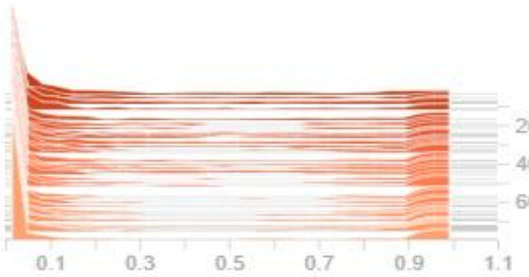

Fig. 5: Activation histogram of train data

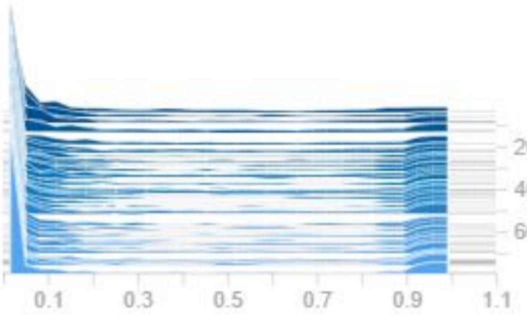

Fig. 6: Activation histogram of validation data

IV. CONCLUSION

With so much data at disposal, machines are becoming smarter day by day. With this much data available the machine learning algorithms are able to reach near-human semantic classification of visual patterns such as places and objects. This work allows even normal machines such as laptops to detect place scene within a fraction of second and high accuracy. For now, the system is optimizing the number of repetitions in order to accurately identify the scenes in the images. Places' meticulous structure and vast coverage of image world may help advance the understanding of human visual system collection. It can be concluded from the results obtained by successful implementation that system is able to categorize the given scenes impeccably. This work can lead to solving of other problems such as actions happening in the environment or spotting the inconsistent objects. If required in future, of the images in the dataset can be increased in order to make scene recognition more accurate nearly 99%. Hence, our proposed system is just an attempt towards making scene recognition more accurate and fast as possible. It can be improved and modified further as per the requirement. With time, the machine is only going to get smarter reaching human level of intelligence.

ACKNOWLEDGMENT

The authors would like to thank SRM Institute of Science and Technology for research work. The authors would also like to thank Ms. Sharmila Agnal for her guidance in the implementation of work.

REFERENCES

1. Bolei Zhou, Agata Lapedriza, Aditya Khosla,Aude Oliva, and Antonio Torralba, “Places: A10 Million Image Database for SceneRecognition”
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition”
3. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database”
4. Neetu Rani, “Image Processing Techniques: A Review”
5. Simon Lacoste, JulienFei Sha and Michael I. Jordan, “DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification”
6. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis and Jeffrey Dean, “TensorFlow: A System for Large-Scale Machine Learning”
7. Danilo De Donno, Alessandra Esposito, Luciano Tarricone, and Luca Catarinucci, “Introduction to GPU Computing and CUDA Programming: A Case Study on FOlD”
8. Christian Szegedy, Wei Liu, Yangqing Jia1, Pierre Sermanet and Scott Reed, “Going Deeper with Convolutions”
9. Li-Jia Li and Li Fei-Fei “What, where and who? Classifying events by scene and object recognition”
10. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, “Learning Deep Features for Scene Recognition using Places Database”

AUTHORS PROFILE

Sharmila Agnal.A is professor in SRM Institute of Science and Technology for CSE department. She has great knowledge in field of AI due to her teaching experience. Currently she is also involved in other research work with her students.

Jenny Savani is currently pursuing her B.Tech in CSE from SRM Institute of Science and Technology.

Suchhanda Das is currently pursuing her B.Tech in CSE from SRM Institute of Science and Technology.

Preetam Swaraj is currently pursuing his B.Tech in CSE from SRM Institute of Science and Technology.

Ayushi Rathi is currently pursuing her B.Tech in CSE from SRM Institute of Science and Technology.