International Journal of Gender, Science and Technology
..ISSN: 2040-0748
UGC Care Group I Journal
Vol-11 Issue-01 June 2022

# SCENE UNDERSTANDING OF NATURE

**ALANKARAM SHOBITHA LAKSHMI1, Dr. BONTHALA VAMSEE MOHAN2**

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali
#2Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

**ABSTRACT_** It is our ultimate goal to understand an image by assessing its attributes and then classifying them under their relevant scenes. A dataset of 6000 trained and 3000 test photos was obtained and used to form convolutional neural networks, which were then tested using a variety of images from various categories. There are six classes in the dataset that provide a global view of nature. A convolution and pooling layer is used to train our model, followed by a fully connected layer. The dataset is then tested and sorted into scene categories after the model has been trained.

Our ability to obtain more accurate outcomes with CNNs (convolutional neural networks) is unmatched by any other method. It has also reached a level of visual recognition that is close to that of a human. Other image classification techniques require a lot more pre-processing than CNN. Through automated learning, the network learns how to optimise the filters or kernels. CNN has a huge advantage because of the lack of prior information and human interference in feature extraction.

## 1.INTRODUCTION

The ultimate goal of computer vision advancement has been to create machines with human-like vision. This project's ability to establish a context for object recognition relies on a thorough comprehension of the scene. For computer vision, scene understanding has a key position since it allows for new insights to be gained through the simultaneous perception, analysis, and interpretation of dynamic scenes in real time. Any image should be comprehensible and decipherable by computers in the same way that it is understandable by humans. Google Photos, intelligent robots, human-computer interaction, self-driving cars, and smart video surveillance all use scene recognition. Basically, it's a representation of the real world with a variety of surfaces and things in a meaningful arrangement.

We can achieve more accurate results using CNN than with any other method. It has also reached a level of visual recognition that is close to that of a human. There are 6000 photos in the training set and 3000 images in the testing set of our dataset, all of which have been categorised into six different categories. Image classification techniques that use convolutional neural networks require less pre-processing than others. Through automated learning, the network learns how to optimise the filters or kernels. CNN has a huge advantage because of the lack of prior

information and human interference in feature extraction.

To train the given dataset, we will apply a convolutional neural network (CNN) to it, like a series of convolution and pooling layers (max pooling) followed by a fully connected layer until we get a precise image with reduced features that helps us classify the images of predicted sets under the appropriate scene. So that during the testing phase, the input image that is being tested will be convoluted into layers and undergo classification thus labelled to a certain relevant class by the supervised learning method that we use.

## 2.LITERATURE SURVEY

Recently, scene recognition has received a lot of attention because of its quick growth. It is a necessary stage in a variety of applications, such as robot navigation and map construction. Deep learning techniques such as Convolution Neural Networks (CNN) have been developed to improve scene representation. A new scene-centric database called Places with over 7 million tagged photographs of scenes was introduced and new methods were proposed to compare the density and diversity of image datasets and showed that Places is as dense and diverse as previous datasets. For scene recognition tasks, they use a CNN to learn deep features and build new benchmarks on a number of scene-centric datasets. [3] They were able to see the differences between object-centric and scene-centric networks by visualising the CNN layers' responses [3]. When it comes to recognising a scene, one must know both the scene and its items, according to Luis Herranz, Shuqiang Jiang, Xiangyang Li (2016), who stated that scenes are formed of things.

There are two main issues to address when it comes to combining scene-centric and object-centric knowledge (i.e. Places and ImageNet) in convolutional neural network (CNN) architectures: A previous project with Hybrid-CNN demonstrated this.

They found that adding ImageNet didn't assist much, so they came up with a different strategy that took scale into consideration and resulted in large recognition gains. To avoid dataset bias and poor performance, researchers examined the responses of ImageNet-CNNs and Place-CNNs at various scales. They discovered that both networks operate at different scale ranges and thus utilising the same network for every scale would be a mistake. According to their findings, the level of recognition accuracy varies greatly depending on the scale, and a combination of ImageNet-CNNs and Places-CNNs, chosen with care, can raise the level of SUN397's state-of-the-art recognition accuracy to as high as 66.26% [4]. Piotr Wozniak is a Polish cyclist. Her

name was Hadha Amiri. Recep Tayyip Erdogan An approach based on deep neural networks proposed by Bogdan Kwolek(2014) was developed for indoor place recognition by retraining VGG-F, a previously trained convolutional neural network, using transfer learning.

Training and evaluation of the network have been done on a dataset of 8000 photos that were taken in sixteen rooms. They have made the data available on their website for anyone to use. According to the results of their experiments, they found that the proposed method outperformed the more common BoW techniques in loop-closure. The FC-6 layer of the VGG-F also outperformed a linear SVM-based classification system. More local structural and discriminative information is implicitly encoded in images when using traditional dictionary-based features (such as BoW and spatial pyramid matching).

To increase the efficiency For scene detection and adaption, Guo-Sen Xie; Xu-Yao Zhang; Shuicheng Yan; Cheng-Lin Liu(2015) proposed combining CNN with dictionary-based models (DA). CNN models (such as AlexNet and VGG Net) are used to build two dictionary-based representations: the MLR and the CFD.

By first grouping all representative parts of a single image and then all parts of all images, the weighted spatial and feature-space spectral clustering employed in MLR is an effective two-stage clustering method for generating a class-mixture or a class-specific part dictionary. Once the multiscale picture inputs have been processed using the component dictionary, they are then used to generate a mid-level representation. Fisher vectors are generated in CFV using a multiscale and scale-proportional Gaussian mixture model training technique. Connectedness elements of fully connected MLR are combined with supplementary information from CFV and CNN.

On scene recognition and DA problems, state-of-the-art performance can be attained.

In addition to being complimentary to GoogLeNet and/or VGG- 11 (trained on Place205), their proposed hybrid representation (from VGG net trained on ImageNet) also makes an important discovery. GoogLeNet-based multi-stage feature fusion was proposed by Pengjie Tang, Hanli Wang, and Sam Kwong (2016). (G-MS2F). The three outputs that correspond to the three components of the proposed model are fed into the product rule, which then generates the final choice for scene recognition. On the scene recognition datasets Scene15, MIT67, and SUN397, the proposed model outperforms a number of state-of-the-art CNN models and achieves recognition accuracy of 92.90 percent, 79.63 percent, and 64.06 percent, respectively [7].

**3.PROPOSED SYSTEM**

A total of six categories designated (0:buildings, 1: woodland, 2: glacier, 3: mountain), are taken for scene classification when the relevant libraries are imported. This is done by taking a training set of 6000 photos, followed by a testing set of 3000 validation images. CNN models are constructed using a series of convolution and pooling layers, then a fully linked layer. Images of 128*128*1 in size are fed sequentially into the convol layer with activation function 'Relu,' and into the pooling layer before entering the fully-connected layer, where flatten and dense layers with activation function "Softmax" are used to classify images based on output from convolutional layers. The prediction set of 1531 photos is used to test the entire model, and the images selected at random for prediction are classified correctly into one of six categories and then given an output of the scene to which the image belongs. The above-mentioned scenario classification has an accuracy rate of 81%.

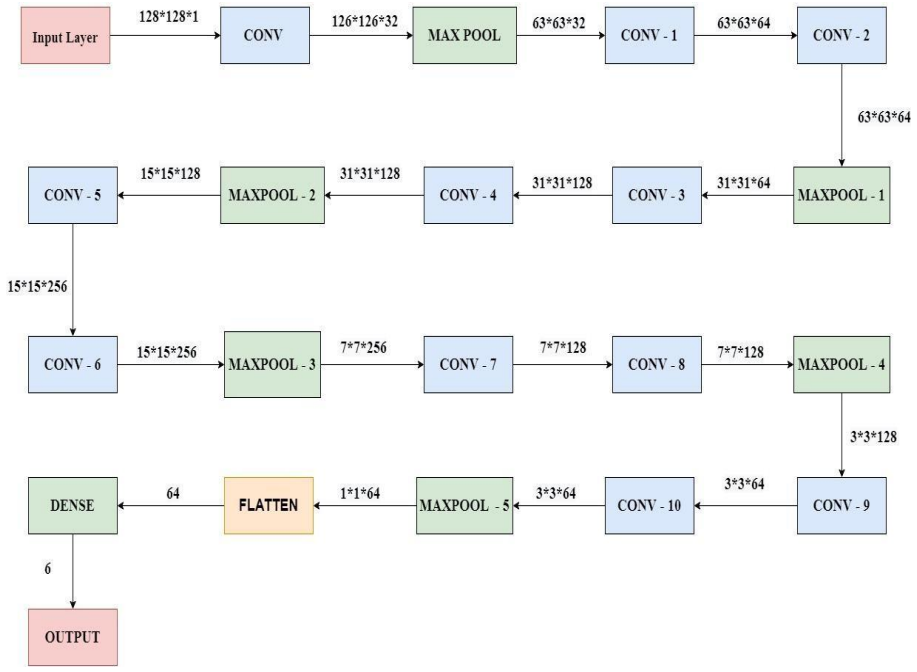



**Fig 1: Proposed System**

### 3.1 ALGORITHM

- Considered 6 categories labeled as - 0:buildings, 1: forest, 2:glacier, 3:mountain,4:sea, 5:street for scene classification.
- A training set of 6000 images and a testing set of 3000 validation images are takento train the model.
- The input image of size 128*128*1 is fed into the convol layer with activation function - 'Relu' and pooling layer
- A CNN model is built with a series of convol and pooling layers followed by a fullyconnected layer. sequentially and then into the fully connected layer where the flatten and dense layers with activation function - 'Softmax' are applied.

- The overall model is then tested by the prediction set consisting of 1531 images, where the images chosen randomly are correctly classified among given 6 categories giving an output of the respective scene to which the image belongs.
- Accuracy and experimentals results were obtained for the classified images.

**ARCHITECTURE**

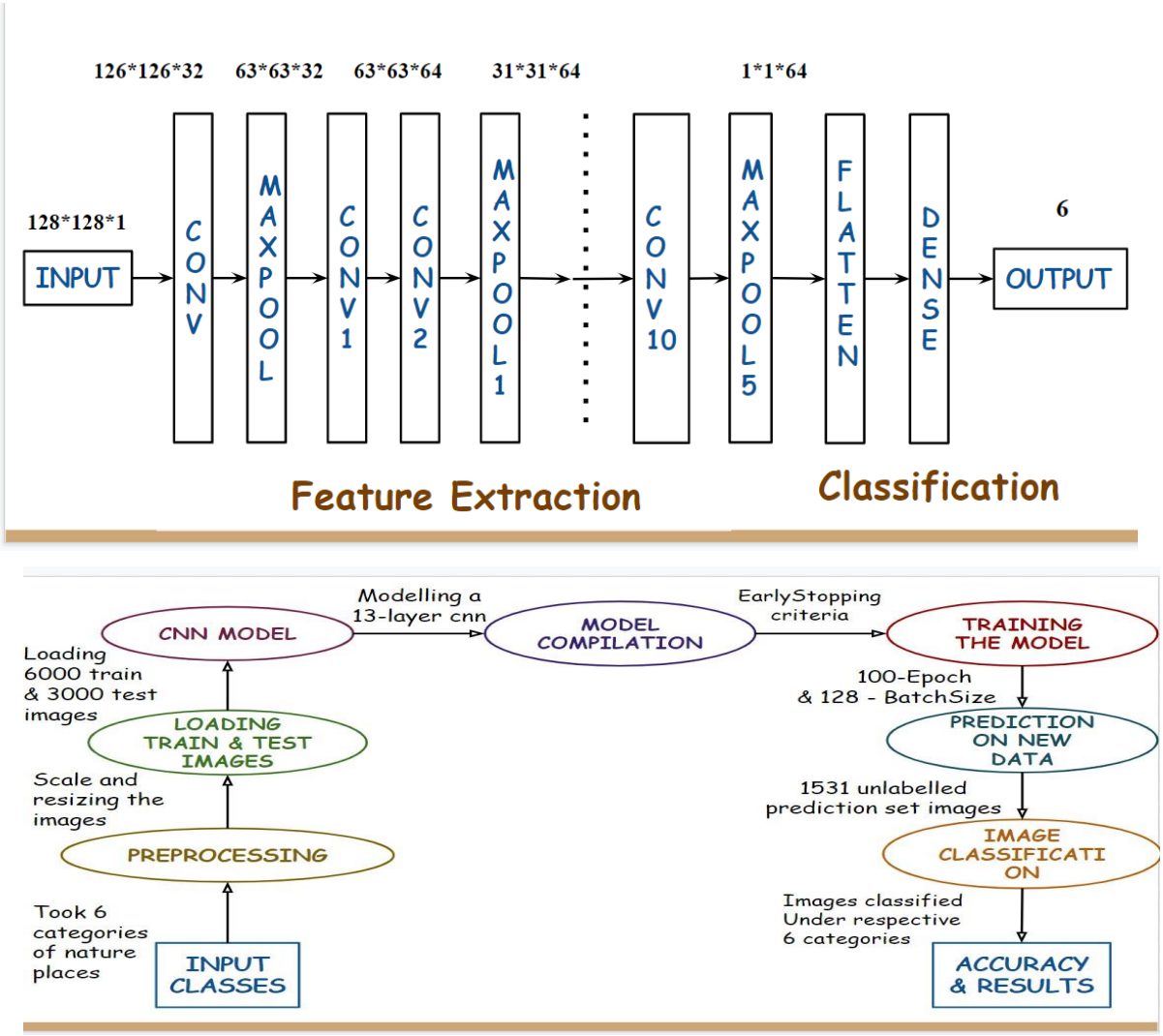



**Figure 2: Architecture of proposed system**

**3.2 IMPORTING MODULES:**

**Cv2** is a library of Python bindings designed to solve computer vision problems. cv2.imread()method loads an image from the specified file. If the image cannot be read (because of missing file, improper permissions, unsupported or invalid format) then thismethod returns an empty matrix.

**TensorFlow** is a Python library for fast numerical computing created and released by Google. It is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow.

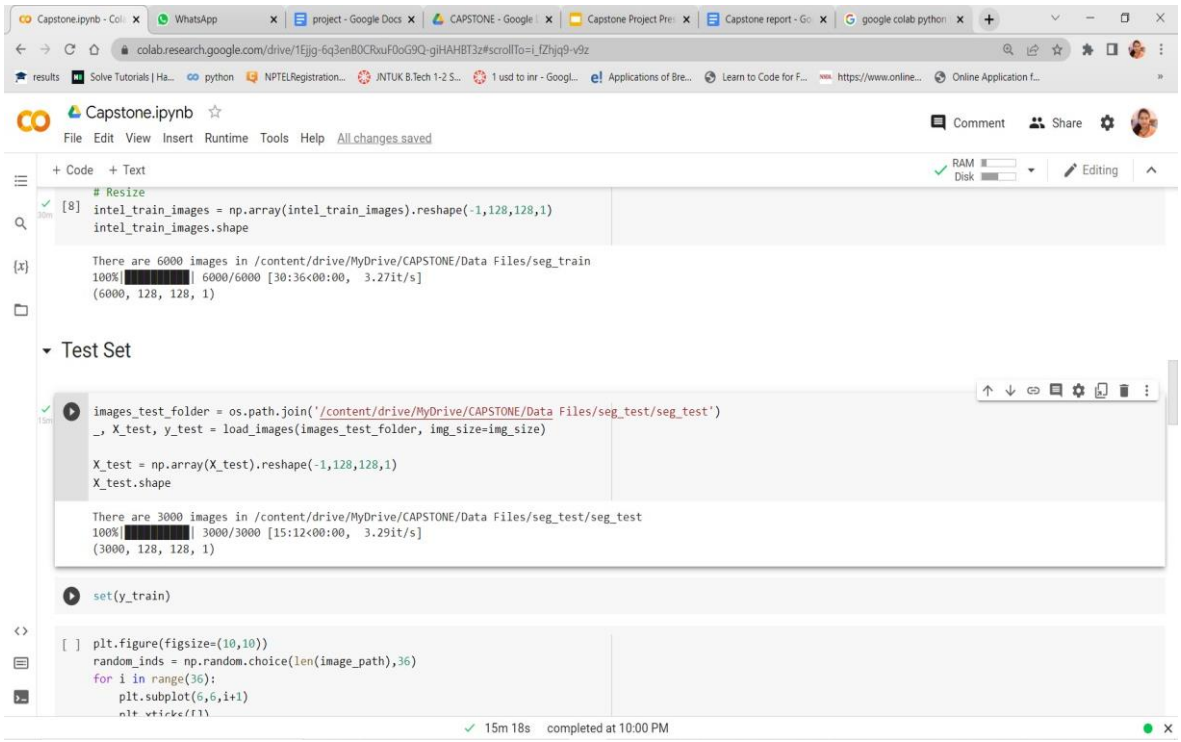
**Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

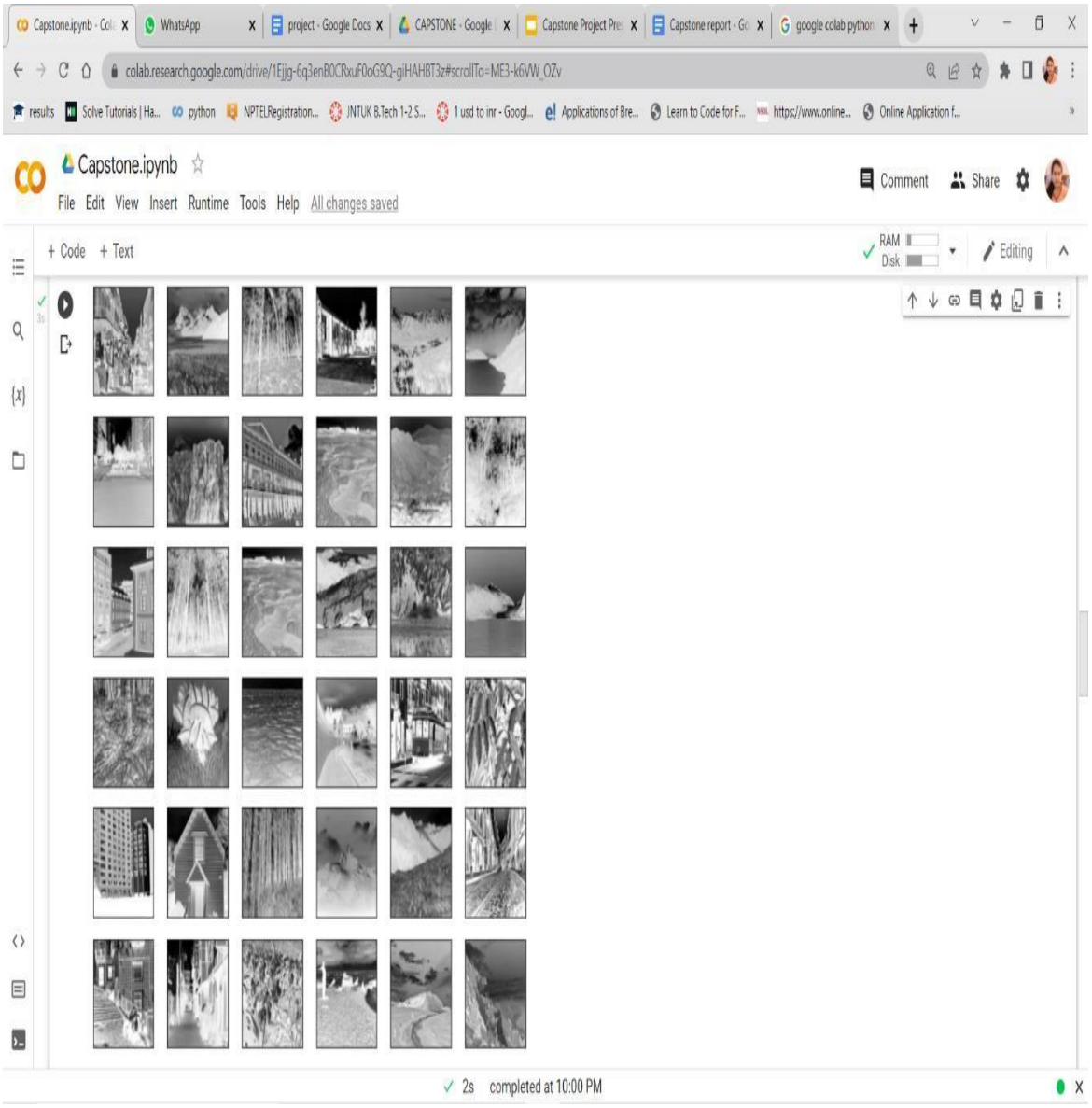
**Keras** is an open-source high-level Neural Network library, which is written in Python is capable enough to run on Theano, TensorFlow and modular for facilitating faster experimentation with deep neural networks

## 4.RESULTS AND DISCUSSION

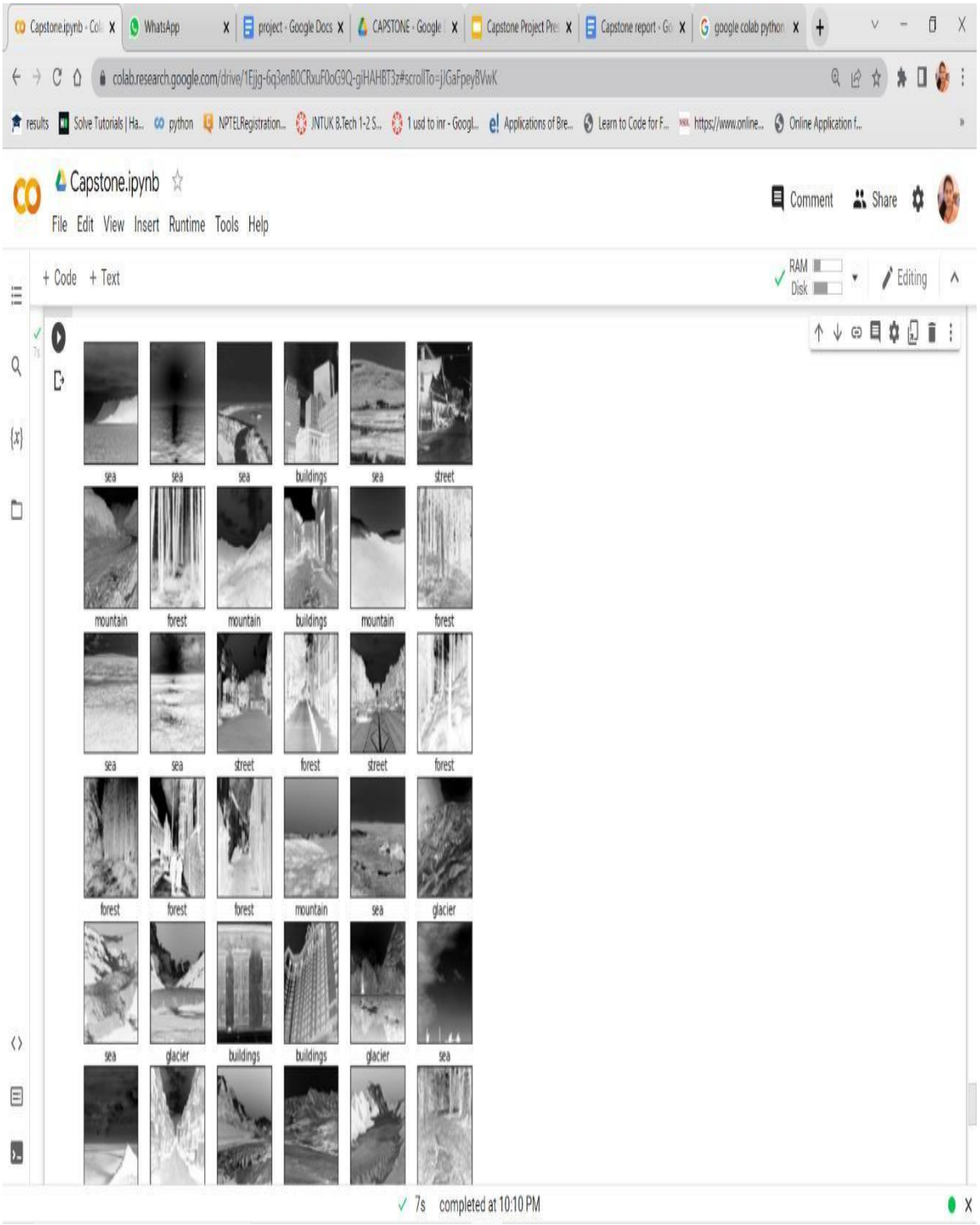
### Test Model Set

### Reading Random Unlabelled Images from Drive

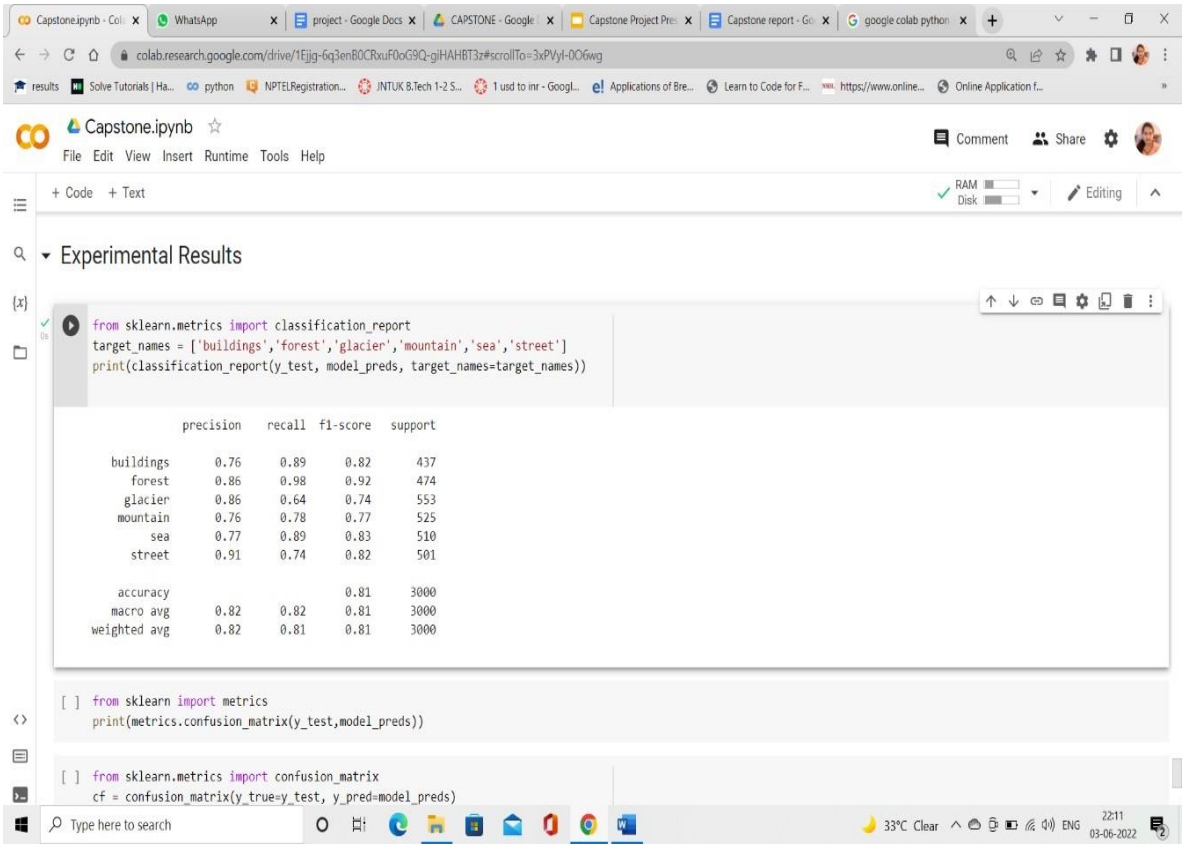



Capstone.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
2s completed at 10:00 PM

**Labelling Images By Model**



## Experiment Results




## Accuracy Score

## 5.CONCLUSION

Organizing different surfaces and objects into a coherent whole is what Scene Understanding is all about. We used a dataset of scene-centric photos comprised of six categories of various natural sites throughout the world to develop a model of convolutional neural networks and reach the highest level of accuracy in nature scene categorization. In this case, the result is obtained from a prediction set of photos that have been categorised according to their particular scene classifications, such as mountains, streets, woods, oceans, buildings, or glaciers. The model's total accuracy is 81 percent. 6000 photos were utilised to train the model, while 3000 images were used to test it, all of which were labelled according to six different categories. Afterwards, we used 1531 photos to categorise the various categories based on their labelling. In addition, the correctness of each category is determined. 81 percent of the model's accuracy is accounted for.

## REFERENCES

[1] Liang, M.; Hu, X. Recurrent convolutional neural network for object recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.

[2] Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark andState of the Art. Proc. IEEE 2017, 105, 1865–1883. [CrossRef]

[3]. Xia, G.S.; Tong, X.Y.; Hu, F.; Zhong, Y.; Datcu, M.; Zhang, L. Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation. arXiv 2017, arXiv:1707.07321.

[4]. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. ISPRS J. Photogramm. Remote Sens. 2016, 117, 11–28. [CrossRef]

[5]. Chen,S.; Tian, Y.L. Pyramid of Spatial Relatons for Scene-Level Land Use Classification. IEEE Trans. Geosci. Remote Sens. 2014, 53, 1947–1957. [CrossRef]

[6]. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.

[7]. Zhang, X.; Du, S. A Linear Dirichlet Mixture Model for decomposing scenes: Application to analyzing urban functional zonings. Remote Sens. Environ. 2015, 169, 37–
49. [CrossRef]

[8]. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. Science 2006, 313, 504–507. [CrossRef]

[9]. Liu, J.; Shah, M.; Kuipers, B.; Savarese, S. Cross-view action recognition via view knowledge

transfer. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3209–3216

[10] A. Dragan, N. Ratliff, S. Srinivasa, Manipulation planning with goal sets using constrained trajectory optimization, in: Int. Conf. on Robotics and Automation (ICRA), 4582–4588, 2011.

of a moving object from range video, in: Int. Conf. on Robotics and Automation (ICRA), 2617–2622, 2014.

[13]. D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Int. Conf. on Computer Vision (ICCV), 2650–2658, 2015.

[14]. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Computer Vision

[11]. D. Mart´ınez, G. Alenya, C. Torras, Planning robot manipulation to clean planar ` surfaces, Engineering Applications of Artificial Intelligence (EAAI) 39 (2015) 23–32. 8

[12]. F. Husain, A. Colome, B. Dellen, G. Alenya, C. Torras, Realtime tracking and grasping

and Pattern Recognition (CVPR), Conf. on, 3431–3440,2015.

[15]. F. Husain, H. Schulz, B. Dellen, C. Torras, S. Behnke, Combining Semantic and Geometric Features for Object Class Segmentation of Indoor Scenes, IEEE Robotics and Automation Letters (RA-L) 2 (1) (2016) 49–55.
[7] F. Husain, B. Dellen, C. Torras, Action Recognition based on Efficient Deep Feature Learning in the Spatio-Temporal Domain, IEEE Robotics and Automation Letters (RA-L) 1 (2) (2016) 984–991.