INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

A novel approach for indoor-outdoor scene classification using transfer learning

A. Yashwanth
aineeryyashwanth@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Shaik Shammer
shaik.shammer786@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

R. Sairam
sairam90056@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

G. Chamundeeswari
easwari.harshith@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

ABSTRACT

Scene understanding and analysis has gained significant importance and widely used in computer vision and robotics field. Classification of complex scenes in a real-time environment is a difficult task to solve. Convolution Neural Networks (CNNs) is a widely used deep learning technique for the image classification. But the training of CNNs is not an easy task since it requires large scale datasets for training. Also, the construction of CNN architecture from scratch is a complex work. The best solution for this problem is employing transfer learning which gives the desired result with small scale datasets. A novel approach of Alexnet based transfer learning method for classifying images into their classes has been proposed in this paper. We selected 12 classes from publicly available SUN397 dataset out of which 6 are indoor classes and the remaining 6 are outdoor classes. The model is trained with indoor and outdoor classes separately and the results are compared. From the experimental results, we found that the model exhibited an accuracy of 92% for indoor classes and 98% for outdoor classes.

Keywords— *Indoor, Outdoor, Scene classification, Transfer learning, CNN, Alexnet*

1. INTRODUCTION

In recent times, the indoor and outdoor scene recognition methods are vividly implemented in hand held 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, the scene classification algorithms help robots in recognizing the type of environment in which they are working. Also, there are many pictures being clicked by photographers at different places across the world every day. Due to this, the collection of images are has increased enormously. Recovering images or pictures from this large compilation of databases is a highly time consuming and complex task. Scene classification can be implemented to make this procedure easier. From research works in recent times, scene recognition has attained great success using deep learning methods. Researchers are inspired to handle the regular image recognition exercises related to computer vision and improve the accuracy of classification with different models.

In this paper, we classified images into 6 indoor classes and 6 outdoor classes. The indoor and outdoor scene recognition continues to be an open field for research when we aim for context-aware sensing and seamless positioning. Convolution Networks are a development of deep learning which can be efficiently implemented in image classification models.

1.1 Convolution Neural Networks

Convolutional Neural Networks are a type of directed acyclic graphs. Such kind 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. These neurons are clubbed closely such that the input of neurons at layer $l + 1$ is obtained from the output of neurons at layer l, that is.

$$a^{(k+1)} = f(W^k a^k + b^k)[17]$$

Here, $W^{(k)}$ represents a weight matrix of layer k, $b^{(k)}$ represents a bias term, and f is the activation function. The activation of layer k is represented by $a^{(k)}$.

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 which 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.

2. RELATED WORK

In the past few years, several improvements have been proposed to learn rich features from color images. One such approach is to use image region proposals for training Convolutional Neural Networks (CNN's) and another approach is to explore contextual information between different image segments that is to differentiate the multiple images according to their classes and properties. This classification of superpixels at multiple scales had been practised in the past [1] and there is another approach is to train a network end-to-end by attaching a sequence of deconvolution and unspooling layers [2].

In recent times, there are many improvements and developing 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 [3] or scene categorization has been broadly carried out in various environments. Authors, Szummer and Picard [4] demonstrated that the classification performance can be improved by measuring the features on sub-blocks, categorizing these sub-blocks, and then merging these results in a way similar to stacking for scene classification. Kim et.al. [5] proposed an approach for indoor/outdoor scene classification using Edge and Color Orientation Histogram (ECOH) with SVM classifier. Chen et.al. [6] Suggested a method for in-door scene understanding by RGB-Depth images. Zou et.al. [7] Presented a potential approach for scene classification based on the joint representation of local and global spatial features. Mana Shahriari and Robert Bergevin [8] demonstrated a dual-stage Convolutional Neural Networks (CNN) for outdoor-indoor scene classification. Arun Nehru et.al [9] suggested a novel approach by using a linear combination (fusion) of the global descriptor (GIST) and Local Energy based Shape Histogram (LESH) descriptor with Canonical Correlation Analysis (CCA). Yanxiang Chen et.al. Proposed a method for indoor scene understanding with the help of monocular RGB-D [10] images. In addition to that, they conceptualized indoor scene recognition as a global optimising framework which comprised of segmentation, support inference, multi-object recognition and scene classification. Shaopeng Liu et.al. [11] Proposed an approach for scene classification with the help of ResNet [12] based transfer learning model

3. PROPOSED APPROACH

The proposed model deals with the scene understanding and classification that intents to understand the activations from images of various public scene environments with the help of transfer learning. A well known pretrained network, 'Alexnet' [13] has been chosen for our proposed transfer learning model. The following diagram shows the workflow of the proposed model.

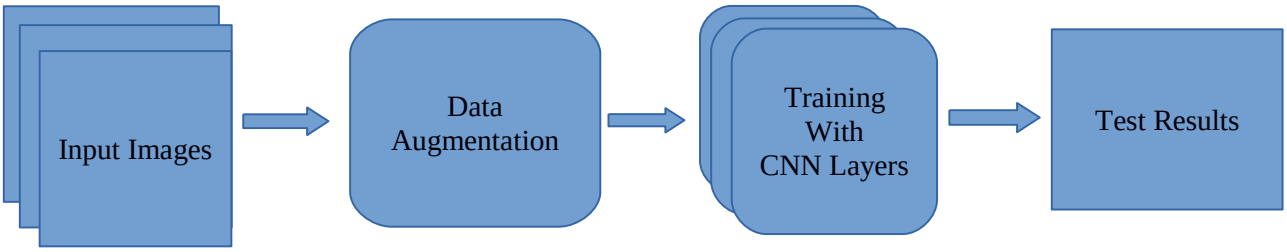

Fig. 1: Overview of the proposed system

3.1. Data augmentation

In order to train our customized dataset on a pre-trained network, we need to preprocess the images before inputting them for training. This process of preprocessing the images before training is called Data Augmentation or Image Augmentation. Data augmentation is required because the trained network will have some predefined parameters for input images. We cannot change the parameters of the input layer because the change may have a negative effect on other layers of the network. Data augmentation is the best solution to address this problem. We can resize or rescale the images according to the parameters of the input layer of the network with the help of data augmentation.

The following figure shows an example of how data augmentation is done on an image.

Fig. 2: Sample images before data augmentation

Fig. 3: After data augmentation

3.2. Activations

As we discussed earlier, in a CNN architecture the output of one layer is passed as input to the next layer. Convolution layers are made of weights and biases that are used to filter an input image. The output of a convolution layer is a set of filtered images. This output is called the activations of that layer. These activations are a 3-D array, where the third dimension is often called a channel.

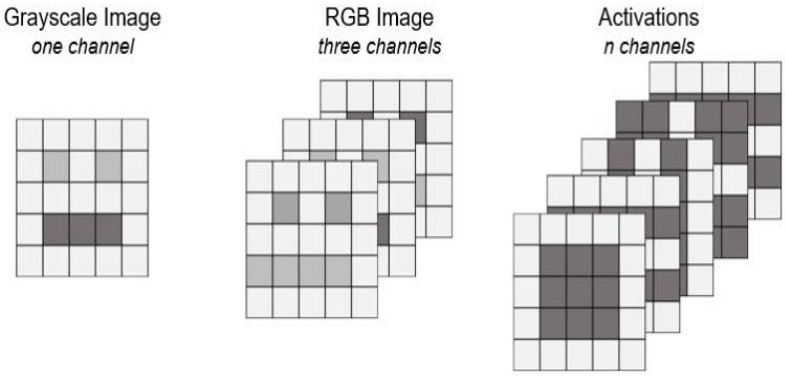

Fig. 4: Channels of activations

There is one output channel for each filter in a convolution layer. A convolution layer can have hundreds of filters, so each layer can create hundreds of channels. You will visualize each channel as a grayscale image.

$$a^{(k+1)} = f(W^k a^k + b^k)[17]$$

Here, W $^{(k)}$ represents weight matrix of layer l, b $^{(k)}$ represents bias term, and f is the activation function. The activation of layer k is represented by a $^{(k)}$

The following figures show the activations of an image from one of our chosen classes at different layers for different channels.

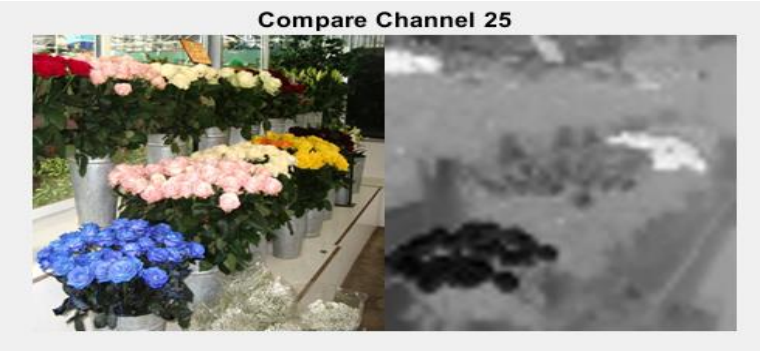

Fig. 5: Activation of 25th channel of 1st CNN layer

Fig. 6: Activation of the 90th channel of 3rd CNN layer

The activations from the first and third convolution layers are displayed. In most of these activations, the flowers are still recognizable. As the image progresses through the network, it will look less like the input image and more like features used to represent the image.

Convolution layers process images from left to right. This filter finds edges from dark to light. The dark color in the image indicates negative activation and the light color represents the positive activation. The grey color indicates nothing interesting.

3.3. Training process

The pretrained network, Alexnet [13] has 25 layers arranged in the series fashion. The 1st layer is a data layer for image input followed by a series of convolution layer, rectified linear unit, normalization layer, max pooling layer ad dropout layer in repetition. The last layer is a classification layer that classifies the images into their classes. The 23rd layer is a fully connected layer that contains fully connected neurons whose count is equal to the number of classes of images that are needed to be classified.

The Stochastic Gradient Descent with Momentum (SGDM) [14] algorithm has been used as an optimiser in training options.

The stochastic gradient descent algorithm oscillates towards the path of steepest descent towards the optimum. By adding the momentum parameter, we can reduce this oscillation [15]. The stochastic gradient descent with momentum update is

$$\theta\ell+1=\theta\ell-\alpha\nabla E\ (\theta\ell)\ +\gamma\ (\theta\ell-\theta\ell-1),\ [14]$$

Where ℓ indicates iteration number, $\alpha>0$ indicates learning rate, θ represents the parameter vector, and $E\ (\theta)$ determines the loss function, γ shows the contribution of the preceding gradient step to the current iteration. You can define this value by using 'Momentum' name-value pair argument. The value for the momentum parameter is given as 0.95 for the training progress.

Initial learning rate defines the rate at which the network learns the features. Longer the initial learning rate, higher is the time taken for training. If the value is less, then the network cannot learn the features properly and we may not get optimal results. Hence, it is very important to choose the correct learning rate that suits our experiment. The default value for the initial learning rate for sgdm solver is 0.01. But we have 0.001 as the value to the initial learning rate for our network.

After the training options were given, the layers in the pretrained network are adjusted according to our dataset and requirements. The weights and parameters of the inside layers are left without changing as they do not require any adjustments and can suit our needs. The number of neurons in a fully connected layer is changed from 1000 to 6 for both indoor and outdoor training network. The existing classification layer or otherwise the output layer is replaced with a new classification layer.

Once the training options are given and layers are adjusted, then the network is trained with our dataset. We have chosen 12 classes from publicly available SUN397 [16] dataset out of which |6 are indoor classes and the remaining 6 are outdoor classes. The model is trained with indoor and outdoor classes separately.

The following figure s shows the progress of training of both indoor and outdoor classes.

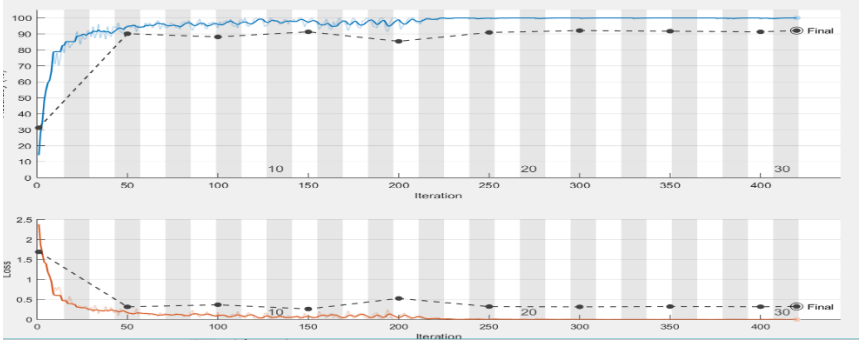

Fig. 7: Outdoor network training progress

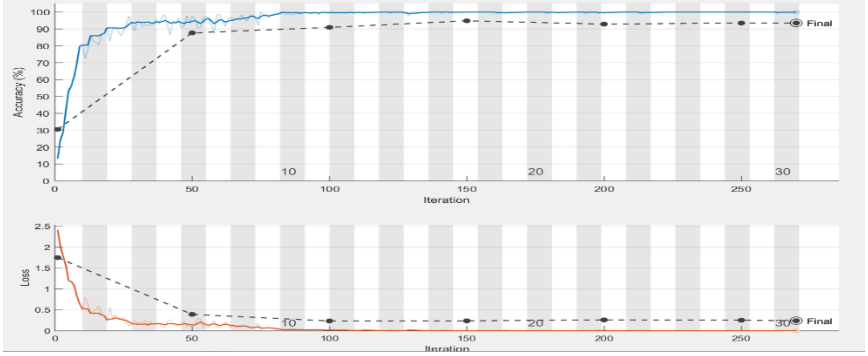

Fig. 8: Indoor network training progress

The 1st graph shows the accuracy of training progress and the 2nd plot shows the bini match loss during the progress of training. The dark line indicates smoothed training, the light colored line indicated training and the dotted line indicates validation.

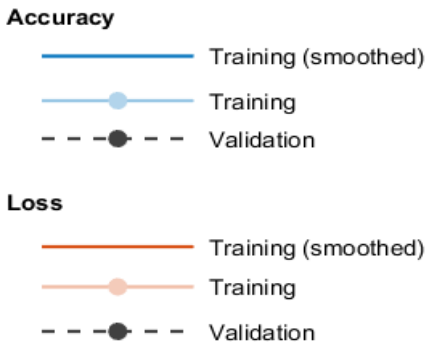

Fig. 9: Accuracy and loss

4. EXPERIMENTAL RESULTS

After training the networks separately for indoor and outdoor classes, it is observed that the accuracy of indoor classes is better than that of outdoor classes from the chosen dataset. The dataset used for the purpose is publicly available SUN397 [16] dataset. Out of 397 different classes in the dataset, 6 classes are selected for indoor environment and 6 are chosen for the outdoor environment, where all classes are public places. The sample images from each class are given in followed figures.
The 6 indoor classes are Airport terminal, Classroom, Florist shop, Gymnasium, Library and Church.

Fig. 10: Sample images from each class in indoor classes

The 6 outdoor classes are Amusement park, Market, Botanical Garden, Gas station, Crossway and Temple.

Fig. 91: Sample images from each class in outdoor classes

The following performance measures are used to determine the reliability of the proposed model.

Accuracy, A = $(tp + tn) / tp + fp + tn + fn$
Precision, P = $tp / tp+fp$
Recall, R = $tp / tp+fn$
F- Measure, F = $2*P*R / P+R$

where tp is truly positive, tn is true negative, fp is false positive and fn is a false negative. These values for each class are obtained from the confusion chart.

The following images show the confusion chart of indoor and outdoor scene classes respectively.

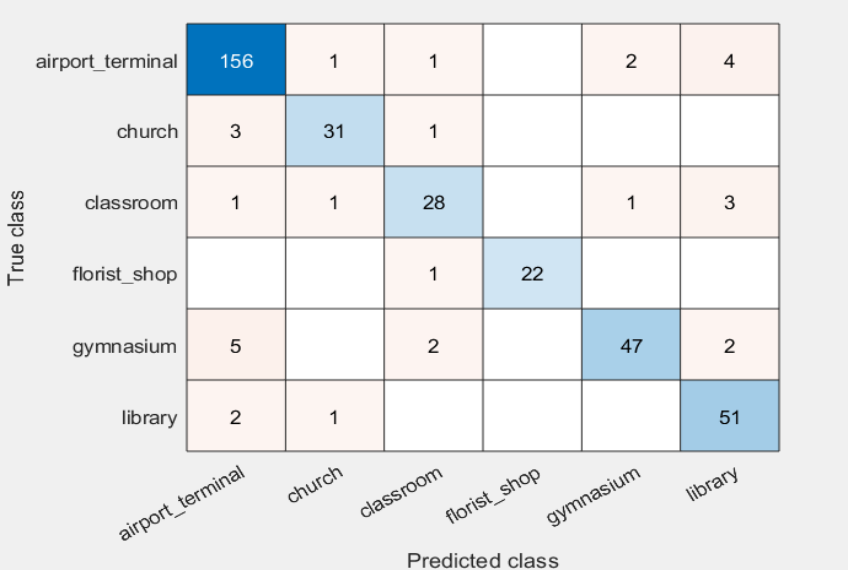

Fig. 10: Confusion chart for indoor classes

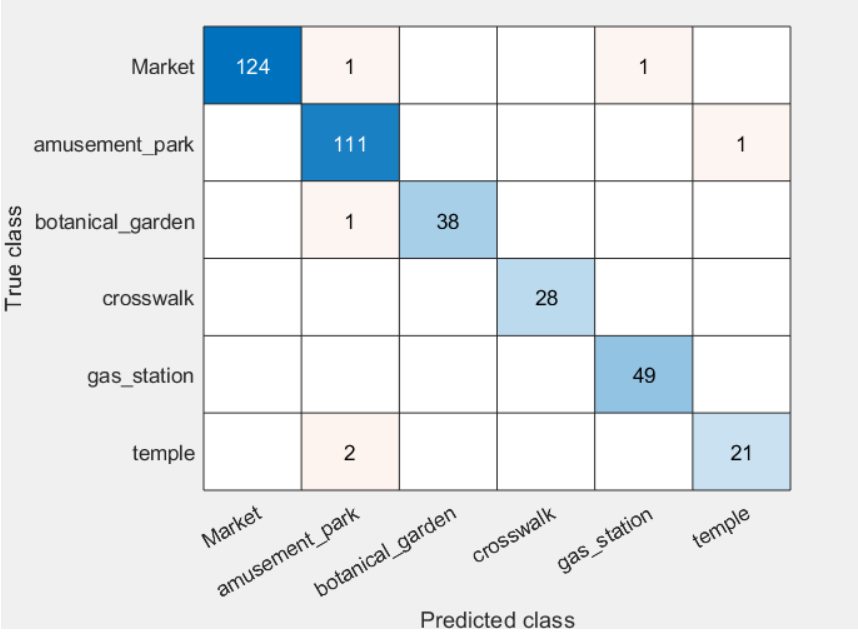

Fig. 11: Confusion chart for outdoor classes

The diagonal of the matrix represents the images that are classified correctly into their respective classes. The actual categories are represented in rows and the predicted classes are indicated by columns.

After evaluating the different performance measures of the network, it is observed that the average accuracy of all indoor classes is 92.5% or 0.925, precision is 94.2% or 0.942, recall is 96.1% or 0.961 and F-measure is 95% or 0.95. For outdoor classes, the average accuracy of all classes is 97.8% or 0.978, precision is 98% or 0.98, recall is 99% or 0.99 and F-measure is 98.3% or 0.983.

With these values of performance measures, we say that the network is more efficient for outdoor classes than that of indoor scenes.

5. CONCLUSION

In the paper, a model based on transfer learning is proposed to classify the indoor and outdoor scene environments. From the experiment results, it is observed that the 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. The further interest is building a new CNN architecture from scratch for indoor-outdoor scene recognition that recognizes both indoor and outdoor classes with high efficiency.

6. REFERENCES

[1] C. Farabet, C. Couprie, L. Najman, Y. Lecun, Scene parsing with multiscale feature learning, purity trees, and optimal covers, in Int. Conf. on Machine Learning (ICML), ACM, New York, NY, 2012, pp. 575–582.

[2] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in Int. Conf. on Computer Vision (ICCV), 2015.

[3] Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 524 531

[4] Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on, IEEE (1998) 42–51

[5] Kim, W., Park, J., Kim, C.: A novel method for efficient indoor-outdoor image classification. Journal of Signal Processing Systems 61(3) (2010) 251–258

[6] Chen, Y., Pan, D., Pan, Y., Liu, S., Gu, A., Wang, M.: Indoor scene understanding via monocular RGB-d images. Information Sciences 320 (2015) 361–371

[7] Zou, J., Li, W., Chen, C., Du, Q.: Scene classification using local and global features with collaborative representation fusion. Information Sciences 348 (2016) 209–226

[8] Shahriari, M., Bergevin, R.: A two-stage outdoor-indoor scene classification frame-work: Experimental study for the outdoor stage. In: Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on, IEEE (2016) 1–8

[9] J. Arun Nehru, A. Yashwanth and Shaik Shammer.: Canonical Correlation-based Feature Fusion Approach for Scene Classification. In 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017

[10] Yanxiang Chen , Daru Pan , Yifei Pan , Shengzhong Liu , Aihua Gu , Meng Wang, : Indoor scene understanding via monocular RGB-D images

[11] Shaopeng Liu, Guohui Tian, Yuan Xu: A novel scene classiÞcation model combining ResNet based transfer learning and data augmentation with a filter

[12] https://en.wikipedia.org/wiki/ResNet

[13] https://en.wikipedia.org/wiki/Alexnet

[14] https://www.mathworks.com/help/deeplearning/ref/trainingoptions.html

[15] Murphy, K. P. Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge, Massachusetts, 2012.

[16] SUN Database: Large-scale Scene Recognition from Abbey to Zoo. J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[17] Farzad Husain, Babette Dellen, Carme Torras: Scene Understanding using deep learning

