*Grenze International Journal of Engineering and Technology, Special issue*

# Indoor Outdoor Scene Classification using Deep Learning Neural Networks

K.Ramya[1], Dr. J. Augustin Jacob[2], D.Bharathi Shree[3] and S. Poornalakshmi [4]
[1, 3]PG Scholar, Department of ECE, Sri Vidya College of Engineering and Technology, Virudhunagar, India
[2]Associate Professor, Department of ECE, Sri Vidya College of Engineering and Technology, Virudhunagar, India
[4]Assistant Professor, Department of ECE, Sri Vidya College of Engineering and Technology, Virudhunagar, India

***Abstract*— Scene recognition is one of the most successful applications of image analysis and it has gained much attention in recent years. A scene recognition system is automatically identifying places by using place databases. Scene classification is act as an image retrieval task which consists of determining a match between the current scene and a previously visited location. Traditional neural networks such as probabilistic neural networks take a maximum time to classification task and performance is not as best. In order to improve the performance and classification speed ,the neural networks are used for scene classification. In recent times, Convolutional neural networks have been shown to achieve brand new performance on various classification tasks. In this project, indoor and outdoor scene classification technique based on Convolutional neural networks is presented. The two databases: places 205 and places 365 is used for scene classification in this proposed method.**

***Index Terms*— Neural Networks; Indoor Outdoor Image Classification; image retrieval; Deep learning; CNN.**

## I. INTRODUCTION

Scene classification involves automatically labeling an image based on its content like indoor or outdoor scenes. Scene understanding is important in scene classification, in order to acknowledge an object, its surrounds should be recognized first, but to recognized the environment, the objects should be recognized. But this is very complicated in image and video processing operations. Scene classification has gained much attention of researchers to reduce the search time and in addition scene classification plays a key role in content based image retrieval, automatic creation of photo albums and tourist information access. Due to the advancement in digital imaging, digital photography collection, image datasets have grown exponentially. Searching and retrieving an image in such a large collection of dataset takes enormous amount of search time.

Scene classification is a one type of image retrieval task that the image consists of various scenes[1]. Generally, an image retrieval system can be divided into two types, namely Annotation-Based Image Retrieval (ABIR) and Content-Based Image Retrieval (CBIR). The drawbacks of ABIR are time consuming due to manual image annotation and ambiguity raised by the synonymy and homonymy of the keywords. In CBIR, the image is addressed by own visual content and then content is represent the color ,shape, and other general information’s which is derived from image itself. Accuracy of CBIR systems mainly depends on the feature extraction techniques because the features carry sufficient information about images. Low level features such as color and

*Grenze ID: 01.GIJET.5.1.9*

textures are recognized as important features to classify an image as indoor or outdoor scene.

Many different types of neural networks used for scene classification. By using probabilistic neural networks[2] ,the scene is initially segmented using fuzzy *C*-means clustering (FCM) and features based on color, texture and shape are extracted from each of the image segments [3] [4][5] [6].And then the feed forward neural networks are used for the classification task, the GIST features extracted from the image. But this is not suitable for noisy images and it taken a long time for extraction process.[7].After, the wavelet decomposition method used for feature extraction and stored in a database as feature vectors and then for classification the back propagation neural network or resilient back propagation neural networks are used.[8] .Later, combination of two independent Neural networks scene classifiers that are connected hierarchically. The first classifier is designated to discriminate between indoor and outdoor scenes while the second classifier is used to classify between outdoor-natural and outdoor-urban scenes.[9][10][11].

Recently, the deep learning based convolution neural networks trained on image net used for place recognition that compares the response of feature layers from a CNN trained on Image Net and methods for filtering the subsequent place recognition hypotheses [12].as well as modified version of CNNs trained with external data and with two different architectures as feature extractors for scene recognition and it is used as A standard linear-SVM-based multi-feature visual recognition framework[13][14][15]. Another successful approach is NBNN and CNN integration frameworks which assures a momentous asset in terms of computational cost. It supports the fully convolutional network for automatize the multi-scale patch extraction process [16][17][18].Two different problems arising for patch extraction. This is overthrown through the principled way of using the multi scale architectures with scale specific networks to address scale-related dataset bias and combine scene and object knowledge [19],[20]-[23]. Another recent development using convolutional network (ConvNet) is extract general purpose features for each of these landmark proposals[24].Previous techniques and their performance shows a success of classification in indoor outdoor scene recognition. The necessity for accurate and detailed oriented classification simultaneously increases the need for changes, adaptations and innovations to deep learning algorithms. The proposed method using caffe deep learning technique and alexnet model for scene classification.

The structure of this article is as follows: the second chapter presents the history, structure and properties of deep learning technology. The third section contains the proposed method is given, and the last section contains evaluation and future studies.

## II. Deep learning

Deep learning is a part of machine learning which is based on learning data representations. Deep learning is an aspect of artificial intelligence that helps human beings to gain certain types of knowledge. In Deep learning ,"deep" denotes that having more than one hidden layer. Traditional neural networks contains 2-3 hidden layers while deep neural networks contains more than 150 layers. Deep learning such as deep neural networks and recurrent neural networks has applied on many fields like computer vision, speech recognition, audio recognition ,etc… But recently used deep learning models are based on artificial neural network[25][26]. It contains multiple layers between input and output layer.

Artificial neural network is also a deep neural network. Sometimes deep neural networks are harder to train compared with shallow neural network. It allows us to train an artificial intelligence (AI) to predict outputs, given a set of inputs. Both supervised and unsupervised learning can be used to train the AI. supervised learning uses labeled data and unsupervised learning uses unlabelled data. Deep learning achieves recognition accuracy at higher levels than ever before. The layer of computational nodes called "neurons" considered as a main element of deep learning neural networks. All neurons are connected with each other in the underlying layer. There are three types of layers considered as main layers. they are input layer, hidden layer, output layer which is explained in figure 1.

Deep learning algorithms try to make classification task perfectly through using above three layers. If classification can't be done perfectly previous layer makes an abstraction for next layer. Each successful layer makes abstraction to next layer which increases the recognition and classification power of deep learning algorithm. The successive layers called hidden layers. Deep learning algorithm acknowledges as finest leading algorithm compared to other algorithms based on feature extraction. The mostly used deep learning architectures are Unsupervised Pretrained Networks (UPNs), Convolutional Neural Networks (CNNs),Recurrent Neural Networks, Recursive Neural Networks.

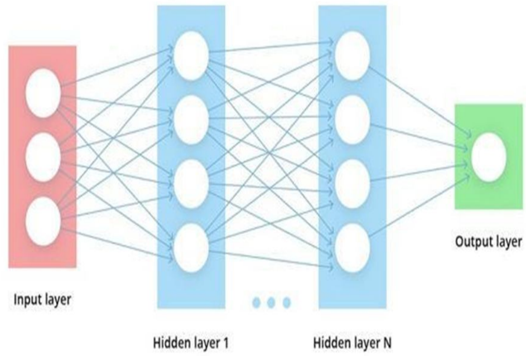


Fig 1:Structure of Deep learning

III. PROPOSED METHOD

*A. Scene classification using Alexnet*

Convolutional neural network is type of artificial neural network. A CNN uses a system much like a multilayer perceptron that has been designed for reduced processing requirements The removal of limitations and increase in efficiency for image processing results in a system that is far more effective, simpler to trains limited for image processing and natural language processing. The convolutional neural networks differed from shallow neural networks by layer architecture. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers and normalization layers.

convolutional neural networks accepts pixel values as inputs to the input layer. the first layer is a convolution layer to extract features from an input image. The Convolution layer uses a filter matrix over the array of image pixels and performs convolution operation to obtain a convolved feature map. The next layer is the ReLU layer which introduces non-linearity to the network. The third layer is known as Pooling layer. Pooling layers uses different filters to identify different parts of the image like edges, corners, body, etc.The fully connected (FC) layer in the CNN represents the feature vector for the input. This feature vector/tensor/layer holds information that is vital to the input. When the network gets trained, this feature vector is then further use for classification.

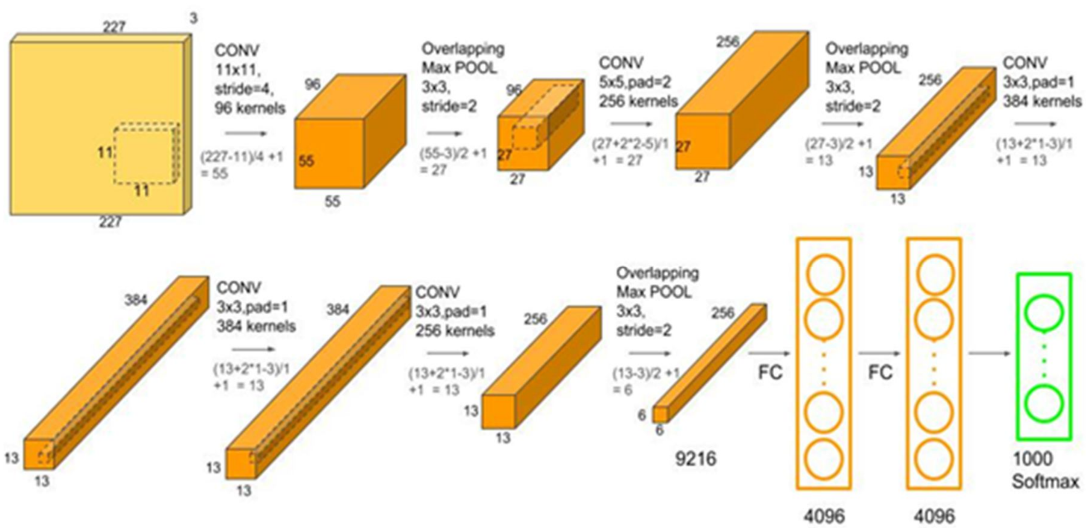


Fig 2: Alexnet architecture

Alex-Net model is the ideal model of CNN, which has three important advantages, i.e., remarkable performance, zero training parameters and high durability. From the above figure, Alexnet contains 5 convolutional layers and 3 fully connected layers. The input to AlexNet is an RGB image of size 256×256. This means all images in the training set and all test images have size of 256×256. Then input images are selected by extracting 227×227 patches from the 256×256 image that is taken away from the corner and center of the image.

The first layer is a Convolution Layer which contains 96 kernels of size 11x11x3 with stride 4 which are convolved with input image and generating 96 different feature maps. The width and height of the kernel are usually the same and the depth is the same as the number of channels. Therefore, the proper setting of the convolutional kernel is essential to extract the necessary and useful image features and improve the performance of the convolutional neural network. The output size of image after applied convolutional layer is 55×55. The 3×3 overlap pooling layers are applied after convolutional layer. The output image size using pooling layer is 96 kernels of 27×27. This overlap pooling layer is used to reduce considerable error rate compared with non overlapping pooling layers.And then second convolutional layer is applied after pooling layer .Finally the output of each layer obtained based on the following formula,

**Number of neurons=kernel*[K+1]**

**O/p = [(W−K+2P)/S]+1.**

Where, O/p is the output height/length, W is the input height/length, K is the filter size, P is the padding, and S is the stride.

The performance of the classification depends on the number of layers used in alexnet. Only 40% of accuracy acquired while using first two layers. If many layers used for classification performance will be high. The third, fourth and fifth convolutional layers are connected directly. The fifth convolutional layer is followed by an Overlapping Max Pooling layer, the output of which goes into a series of two fully connected layers which make the learning features of two channels cross-mixing to obtain a 4096-dimensional feature vector. Drop out technique is used to first two fully connected layers.That means, zero is located for two fully connected layers with ½ probability. The output of the last fully-connected layer is fed to a 1000-way softmax which produces a distribution over the 1000 class labels. Relu is applied after very convolutional and fully connected layer.

## IV. EXPERIMENTAL SETUP

Convolutional neural networks (CNNs) trained on the Places2 Database can be used for scene recognition as well as generic deep scene features for visual recognition. Initially a pertained places205 alexnet dataset is used which contains 2,448,873 images from 205 scene categories. AlexNet CNN trained on 205 scene categories of Places Database with 2.5 million images. Then the update version of place205 is places 365 that accommodates 1.8 million train images from 365 scene categories, which are used to train the Places365 CNNs. There are 50 images per category in the validation set and 900 images per category in the testing set. It comprised of all images of indoor and outdoor scenes including both natural and man-made environments under a wide variety of illumination conditions. The outdoor scenes consist of photographs of mountains, buildings, forests, sunrise/sunset, beaches, ocean, underwater fish and vegetation, animals, cars, boats, streets, and landmarks etc. Indoor scene varieties include photographs of photo studios, hotel rooms, departmental stores, shops, classrooms, offices, living rooms, museums, toilets, corridors, labs, auditoriums, kitchens, gymnasiums, bars, laundry rooms, arcades, storage houses, libraries, restaurants, closets, casinos, subway stations and garages etc. The figure 3 shows the some example images of indoor, outdoor scenes. In figure the above row represents indoor images and below row denotes outdoor images.

## V. EVALUATION

The evaluation of the indoor outdoor scene classification is based on four metrics: Accuracy, Sensitivity and specificity.

**Accuracy:**
**ACC = (TP+TN)/ (TP+TN+FP+FN)**
**Sensitivity:**
**SN = TP/ (TP+FN) = TP/P**
**Specificity:**
**SP = TN/ (TN+FP) = TN/N**

Fig 3: Example sample images of indoor (above row) and outdoor (below row) from places 205 dataset

## VI. EXPERIMENTAL RESULTS

In order to understand the scene classification the primary need is to realize that how the image is classified using Alexnet. For that purpose, the scrutinization of images has been predicted and as the result consequent images are classified. The size of the input image is 227×227 which is said to be mandatory in the case of Alexnet.

The simulation results of indoor outdoor scene classification presented in this section. For example, one sample image taken into consideration and the perspective bar chart is obtained with top 5 predictions. Here, the scene classification established based on the indoor, outdoor images.

The figure 4 depicts the indoor scene classification and the bar graph shows the highest probability .With depends on bar graph the input image is recognized as classroom with 55% probability.

Similarly when tried with other outdoor sample images the Alexnet predicts scenes with notable accuracy, it is shown in figure 5 .The image is predicted as mountain with probability 45%. Many sample images of indoor outdoor scenes from dataset that taken into consideration and output is obtained. Most of the images in the dataset are classified correctly using Alexnet model.

For example, Some correctly classified outdoor sample images are shown in figure6.

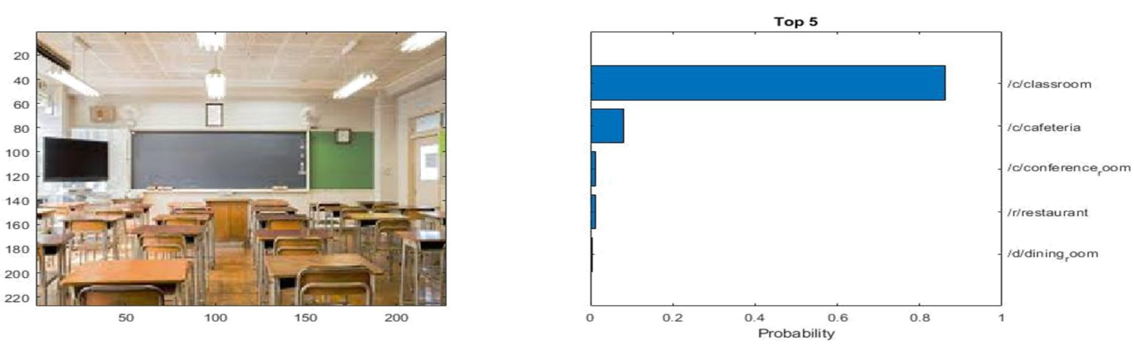


Fig 4:Simulation result of indoor scene image

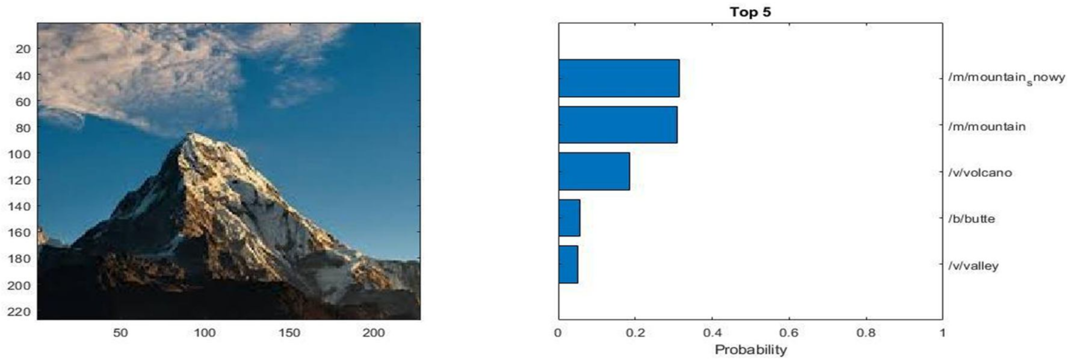


Fig 5: simulation result of outdoor scene image

Fig 6: Examples of correctly classified outdoor sample images

Fig 7: Examples of correctly classified indoor sample images

In addition the Alexnet model also classified many indoor scene images properly. Here some sample images of correctly classified indoor images are shown in figure 7.

Table I and Table II shows that the correctly classified indoor, outdoor sample images and its corresponding outputs .The table contains the original class name of indoor outdoor scenes and their top 5 probability of prediction. The comparison of accuracy for proposed method and previous methods are properly interpreted in the below Table III.

TABLE I: OUTDOOR CLASSIFIED IMAGES AND THEIR OUTPUTS

| **Category** | **Sample image** | **Top 5 Probability of prediction** | **Class name as categorized by Alexnet** |
|---|---|---|---|
| | **a** | Sky scraper-40%<br>Tower-35%<br>Office building-15%<br>Basilica-6%, Building racade-4% | Sky scraper |
| | **b** | Watering hole-50.5%<br>Hot spring-20%<br>Creek-15.5%<br>Rainforest-8%, River-6% | Watering hole |
| **Out door** | **c** | Cross walk-45%<br>Plaza-45%<br>Sky scrapper-5% | Cross walk |

| | | | |
|---|---|---|---|
| | | Market-3%, Office building-2% | |
| | **d** | Boat deck-50%<br>Harbor-22%<br>Ocean-20%<br>Coast-4.2%, Dock-1.8% | Boat deck |
| | **e** | Mountain-35%<br>Sky-25%<br>Volcano-20%, Vislet-14%<br>Coast-6% | mountain |

Table II: Indoor classified images and their outputs

| **Category** | **Sample image** | **Top 5 Probability of prediction** | **Class name as categorized by Alexnet** |
|---|---|---|---|
| **Indoor** | **a** | Pulpit-78%<br>Basilica-18% , Labbey-3%<br>Arch-2%, Church-1% | pulpit |
| | **b** | Dinette home-40%<br>Dining room-29%<br>Restaurant patio-21%<br>Patio-6%, Varenda-4% | Dinette home |
| | **c** | Cafetaria-46%<br>Restaurant-29%<br>Banquet-13.5%<br>Restaurant patio-9%, Foodcourt-2.5% | Cafeteria |
| | **d** | Book store-100%<br>Art gallery-0%, Super market-0%<br>Corridoor-0%, Art studio-0% | Book store |
| | **e** | Classroom-58%, Bonquet-22%<br>Conference room-20%<br>Cafeteria-7%,Conference enterance-3% | classroom |

TABLE III: ACCURACY COMPARISONS OF PREVIOUS AND PROPOSED METHODS

| **S.No** | **Methods** | **Accuracy(%)** | |
|---|---|---|---|
| | | **Indoor** | **Outdoor** |
| 1. | **Proposed method** | **98** | **96.8** |
| 2. | Probability Neural network | 94 | 90 |
| 3. | Resilient neural network | 90.8 | 90.8 |

## VII. CONCLUSION

Alexnet model for scene classification which was constructed using convolutional neural networks. Scene classification was done using a convolutional neural network model that did thousands of categorization of images. There are many publicly available datasets for scene classification. The places 205 and places 365 pretrained datasets are used for indoor outdoor scene classification. The proposed method using alexnet for scene classification has achieved highest accuracy compared with other methods. The efficiency of the proposed method is proved to be higher over existing methods in terms of classification rate.

REFERENCES


[1] C. Carson, M. Thomas, M. Belongie, J.Hellerstein, and J. Malik, "Blobworld: a system for region based image indexing and retrieval," in Third International Conference on Visual Information Systems , June 1999.

[2] Gupta, L., Pathangay, V., Patra, A., Dyana, and A., Das, S., "Indoor versus outdoor scene classification using probabilistic neural network," EURASIP Journ. on Adv.in Signal Processing, Special Issue on Image Perception 2007.

[3] Lalit Gupta, Vinod Pathangay, Arpita Patra, A. Dyana and Sukhendu Das "Indoor vs. Outdoor Scene Classification using Probabilistic Neural Network "

[4] OndrejDrbohlav,andMikeChantler, "Illumination invariant texture classification using single training images," In Texture 2005: Proceedings of the 4th international workshop on texture analysis and synthesis.

[5] Ju Han, Kai-Kuang Ma, "Rotation-invariant and scale-invariant Gabor features for texture

[6] imageretrieval" ,Image and Vision Computing,Volume 25, Issue 9, 1 September 2007, Pages 14741481.

[7] I E. Gordon, Theories of Visual Perception, 3rd ed. New York, USA: Psychology Press, 2004.

[8] C. Lu, P. Chung, and C. Chen, "Unsupervised texture segmentation via wavelet transform," Pattern Recognition , vol. 30, pp. 729–742, 1997.

[9] Waleed Tahir, Aamir Majeed "Indoor/Outdoor Image Classification Using GIST Image Features and Neural Network Classifiers "

[10] Amitabh Wahi,Sundaramurthy S "Wavelet - Based Classification of Outdoor Natural Scenes by Resilient Neural Network"'

[11] S.Md. Mansoor Roomi, R.Raja "Robust Indoor/Outdoor Scene Classification " "IIT Madras Scene Classification Image Database(SCID),

[12] IgnazioPillai, and Riccardo Satta, "ExploitingDepth Information for Indoor OutdoorScene Classification," IEEE conference, 2012.

[13] Zetao Chen, Obadiah Lam , Adam Jacobson and Michael Milford "Convolutional Neural Network-based Place Recognition "

[14] R. Raja, S.Md.Mansoor Roomi, D.Dharmalakshmi, "Classification of Indoor/Outdoor Scene".

[15] Li Zhou, ZongtanZhou, and DewenHu, "Scene classification using multi-resolution low level featurecombination," Nurocomputing 122(2013) 284–297, 2013.

[16] Fernando Rodriguez, and Guillermo Sapiro, "Sparse Representations for image classification:Learning discriminative and reconstructive non-parametric dictionaries," IMA Preprint Series #2213,2008

[17] Massimiliano Mancini, Samuel Rota Bul ` o, Elisa Ricci, and Barbara Caputo "Learning Deep NBNN Representations for Robust Place Categorization ".

[18] F. Mokhtarian and M. Bober, Curvature Scale Space Representation: Theory, Applications and MPEG-7 Standarization Kluwer Academic Publishers, 2003.

[19] Navid serrano, Andreas Savakis and Jiebo Luo, "A computationally efficient approach to indoor/outdoor classification," IEEE ,2002.

[20] Tao, L., Kim, Y.H., Kim, Y.T, " An efficient neural network based indoor-outdoor scene classification algorithm," In: Int. Conf. on Consumer Electronics (ICCE) pp. 317,318, 2010.

[21] Luis Herranz, Shuqiang Jiang, Xiangyang Li" Scene recognition with CNNs: objects, scales and dataset bias".

[22] Donald F.Specht, "Probabilistic neural networks," Neural Networks , vol. 3, pp. 109–118, March 1990.

[23] P. E. H. Richard O. Duda and D. G. Stork, Pattern Classification . John Wiley & Sons (ASIA) Pte Ltd, 2004.

[24] Niko Sunderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub,Edward Pepperell, Ben Upcroft, and Michael Milford "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free"

[25] John Wright, and y.yang, "Robust Face Recognition via Sparse Representation," IEEE trans. On Pattern Analysis and Machine Intelligence vol.31, 2009.

[26] Emine Cengil, Ahmet Ç I nar, Erdal Özbay"Image Classification with Caffe Deep Learning Framework ".

[27] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In Computer Vision and Pattern Recognition (CVPR) Workshops, pages 512– 519, June 2014.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR), 2015.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Computer Vision and attern Recognition (CVPR), pages 1–9, June 2015.

[30] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In Computer Vision and Pattern Recognition (CVPR), pages 1521–1528, June 2011.

[31] Torralba, B. C. Russell, and J. Yuen. Labelme: Online image annotation and applications. Proceedings of the IEEE, 98(8):1467–1484, Aug. 2010.