Inception Classification and Object Detection based Joint-CNN for Indoor Scene Classification

Yanling Tian, Weitong Zhang, Qieshi Zhang, *Member, IAENG,* and Gang Lu*

***Abstract*—While convolutional neural network (CNN) has been successfully used in many fields including single-label scene classification, it is vital to note that real world scenes generally contain multiple semantics and multi-label, especially in the indoor scene classification due to its content complexity. At the same time, most approaches try to make the network much deeper to make sure that they can extract more detail information. However, the deeper network will cause a lot of problems such as the increase of computational costs and network costs and so on. In order to solve these problems, this paper presents a novel framework which called Joint-CNN based on the proposed special label extraction and network structure. Extensive experiments on various data sets show that our method has enhanced the performance on MIT indoor67 and SUN397 data sets.**

***Index Terms*—scene classification, CNN, multi-label, indoor scene.**

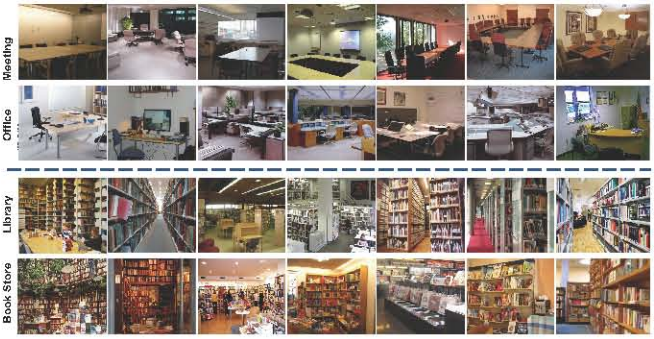

Fig. 1. Image examples from the MIT indoor database. We show two pairs of scene categories (i.e., Office and Meeting room, Book Store and Library). As we can found, they are easily confused.

I. Introduction

SCENE classification is a very important and challenging research topic and application technology in computer vision field. With the development of technology, more and more problems have appeared. Traditional feature-based and semantic-based approaches are no longer so popular. With the emergence of more and more data, classification approaches based on deep learning are widely used in scene classification, especially convolutional neural network (CNN) [1]. Although the accuracy is already high, the classification problems in specific situations cannot be solved, such as the complicated indoor scene classification problem. There are mainly two problems as follows: The first problem is label ambiguity. In general, scene usually consists of several objects. With scene categories number grows, label ambiguity has become another crucial issue in large-scale classification. Single-label-based scene classification method, is no longer suitable for complicated scenes classification. We cannot determine what kind of scenes the image with a few objects belongs to. For example, as Fig. 1 shown, library is very similar to bookstore and they both contain identical representative objects such as books, bookshelves. The second problem is the calculation cost. With the exponential growth of digital images storage, different objects, scenes, actions and attributes exist in an image. The traditional approaches to learning multi-label scenes deeply is to learn their own filters and thresholds for each independent scene. These techniques, although working well, fail to explicitly exploit the label dependencies in an image. Deeper CNN [1] is more conducive to getting visual information and image structure, which causes that more and deeper networks will be created, their calculation speed is slow, and their calculation cost increase.

These challenges provide a major impetus for us to develop a new network for indoor scene classification, by making two major contributions:

1) Propose an Object Detection CNN (OD-CNN) architecture to calculate objects' probabilities and generate a local vector;
2) Modify the Batch Normalization Inception (BN-Inception) [2] model to generate an Inception Classification CNN (IC-CNN) to classify the images with OD-CNN. We combine them into a new network, Joint-CNN.

The rest parts are structured as follows: In Sec. II, we introduce related work including image label and scene classification. Then Sec. III refers to the details of our proposed method. Experiments are in Sec. IV, we will make a comparison with state-of-the-art methods. Lastly, the conclusion is summarized in Sec. V.

II. Related Work

In this section, we briefly review previous works, and then we present related work on two aspects: one is image label, another is scene classification.

Manuscript received Dec. 8th, 2017; revised Jan. 28th, 2018. This research was supported by Natural Science Basic Research Plan in Shaanxi Province of China under Grant No. 2017JM6103, 2017JM6060, Shaanxi Natural Science Foundation Project 2017JM6101, 2017JQ6077, Teaching Reform and Research Project of Shaanxi Normal University under Grant No. 17JG33, the Fundamental Research Funds for the Central Universities GK201703060, and NSFC 61701289

Y. Tian and W. Zhang are with the Key Laboratory of Modern Teaching Technology, Ministry of Education (Shaanxi Normal University), and the School of Computer Science, Shaanxi Normal University, Xi'an, China. (yanl.tian@foxmail.com, weitong.zhang@foxmail.com)

Q. Zhang is with the Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China (qieshi.zhang@ieee.org)

⋆Corresponding author: G. Lu is with the Key Laboratory of Modern Teaching Technology, Ministry of Education (Shaanxi Normal University), and the School of Computer Science, Shaanxi Normal University, Xi'an, China. (goforlg@126.com)

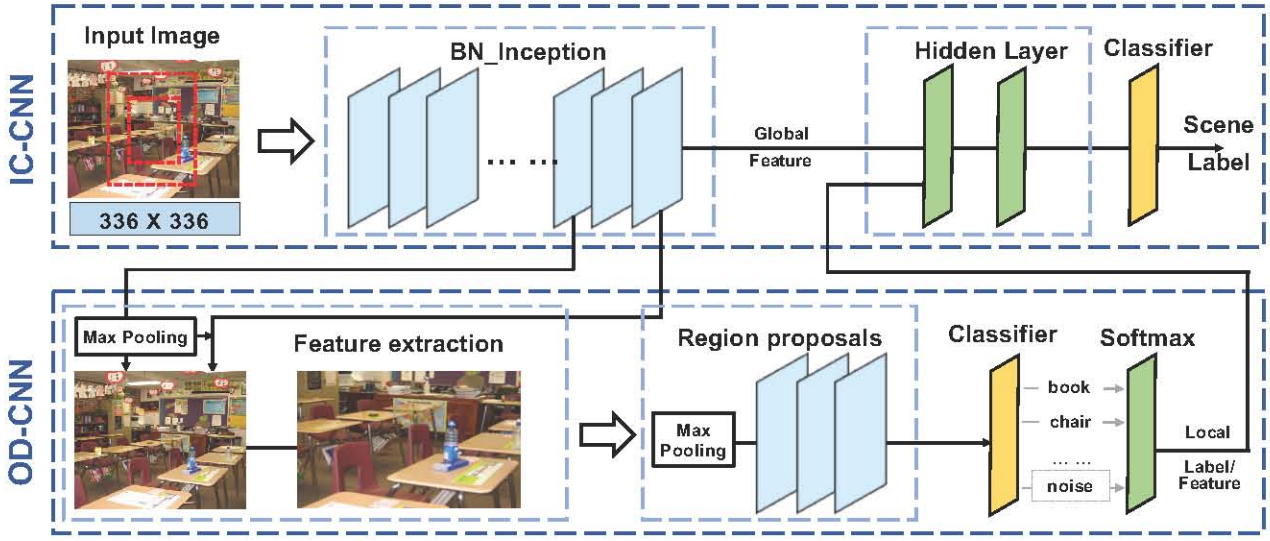

Fig. 2. Joint-CNN. We propose a Joint-CNN, which is composed of IC-CNN and OD-CNN. The OD-CNN picks some image features from IC-CNN, and generates image local feature. While IC-CNN extracts image global feature and then injects into the hidden layer to classify with local feature.

A. Image Label

In scene classification, image labels are some attribute information such as content, category. In the early classification methods, most of the images were based on a single-label, that is to say, only labeling the category which they belong to. However, with the advent of large data sets, scene categories in the data set have increased, and the complexity of pictures in the class has increased. The difference between classes is reduced and the overlap between classes is increased. The image single-label cannot reflect the accurate information of the picture, so the single-label classification method has been limited. So many people did some measures such as object detection to extract more information in the picture as multi-label [3], [4], [5], [6]. Current state-of-the art methods show that multi-label classification task is receiving increasing attention. R. Girshick *et al.* combined region proposals with CNN, called R-CNN [7] to localize and segment objects. J. Wang *et al.* [8], in order to explicitly exploit the label dependencies in an image, they proposed CNN-RNN framework.

However, these methods do not take label ambiguity too much into account when classifying the scenes. We propose a model called OD-CNN, which solves our problem by combining target detection and Faster R-CNN.

B. Scene Classification

Scene classification, is a very important field and challenge in computer vision, many researchers are constantly exploring. In previous work, Initially, LiLJ *et al.* [9] proposed Object Bank (OB), which identified the scene category by identifying objects in the scene. Then, Juneja *et al.* [10] proposed a Bag of Parts (BOP) feature based on high-level semantics to solve the problem of similar information of images, and to extract rich and diverse image semantic information. Recently, deep convolutional networks have been exploited for scene classification by Zhou *et al.* [11]. After this, more and more networks are proposed to solve this problem.

In this paper, our method is different from these previous work. We tackle scene classification problems, such as scene category ambiguity and large computation cost, with deep learning and large indoor scene data set. In addition, we design a new model to extract deep image information as multi-labels.

III. Joint-CNN Framework

In order to solve above problems, we propose a framework called Joint-CNN. It consists of OD-CNN and IC-CNN. Fig. 2 shows the overall framework.

A. Object Detection (OB)

1) Feature Extraction: In order to classify objects and scenes, we apply convolutional network to model this process. According to Fast R-CNN, we optimize it to make it more suitable for our network structure. Joint-CNN is applied so that we can share a common set of convolution layers to get feature maps and reduce the calculation cost in object detection [12] significantly. We extract multi-scale features from interval layers of the Joint-CNN as shown in Fig. 2. Given an image to enhance the detection capability for dense and small targets. We add max pooling layers on the lower layers to carry normalize multiple feature maps and use local response normalization (LRN). In our method, in order to assist the scene to be better classified basing on Joint-CNN, we simplify the neural network structure of object detection and combine it with multi-scale features effectively. The features can be pre-computed before region proposal generation and detection process.

2) Region Proposal: From the method Faster-RCNN, we get enlightenment that a pithy and targeted ConvNet can make the features extracted performs better. We use interval layers to activate the approximate regions and localize the objects. Region proposal has a great relationship with the category and complexity of the scene. So, we obtain about 450 regions on average from the layers we choose. The activated neural units are extracted from the layers not only

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

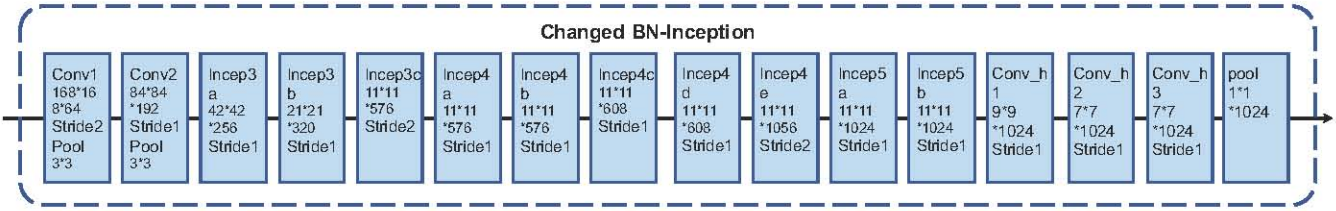

Fig. 3. Changed BN-Inception. We change BN-Inception architecture in order to extract the large-scale image information, and this is the network details.

cover almost the image, but even a large part of them overlaps each other. We calculate Intersection over Union (IoU) among the bounding boxes of region proposals and select a model as a benchmark for adjusting other bounding boxes when IoU meets the condition. The small bounding boxes, when compared to the large ones, are interference that affects the final possible scores. So, we eliminate the useless and overlapping parts to prevent from high false positive rate. We consider IoU$\geq$0.5 between the bounding boxes of a proposed region and rest of the regions is negative. In the end, it approximately keeps nearly half of regions with a larger area.

3) Fine-Tuning for Local Feature: Since a large number of untreated local features patches from hidden layer, an efficient and streamlined classifier is needed to decide which object they belong to [13]. Up to now, methods as Support Vector Machine (SVM), K-means, the trained models as LBP [14] (Local Binary Pattern), neural networks [15] and so on, are accepted by researchers more and more. For the sake of substantial data, the pre-trained convolutional neural network is selected as a fine-turned classifier [16]. The disambiguation clustering methods which based on the stochastic gradient descent of the convolutional neural network is defective. In the process, we found that there are many unrecognized but useful patches wasted except labels contained. We do experiments on this MIT Indoor67 [17], there exist around a million remaining patches, and 200 object labels. So, in the classifier's network, we replace the normal output layer with a new output layer and retain the hidden layers without changing. It is worth mentioning that the purpose of this is that one extra label set is needed to represent those which are discarded. An accurate object label classifier increases the robustness especially for a complex scene with many noisy labels.

B. IC-CNN

Basic Network: With the development of deep learning, more and more CNNs (such as AlexNet [18], VGGNet [19], GoogLeNet [20], ResNet [21], etc.) are used for scene classification, especially indoor scene. In the proposed model, BN-Inception is used as the base model. In addition to its accuracy, it cannot only reduce the number of training parameters but also improves its calculation ability and speed. Based on this, we optimize and modify BN-Inception model, and after that, further classify the scene images with the multi-label which is generated from OD-CNN. It plays well in the scene classification.

BN-Inception, which is shrinking down the representation size so dramatically, doesn't seem to lower the performance that saves you a lot of computation [22]. As our base model, there are two convolutional layers, ten inception layers and a max pooling layer. In addition, BN is applied to the activations of convolutional layers, following by the Rectified Linear Unit (ReLU) [23] for non-linearity. The effects of inception model have been improved, especially the convergence speed is faster by adding sigmoid activation function model after BN operation.

As shown in the Fig. 3. In the original BN-Inception model, in order to make it good at large-scale image classification, it is necessary to increase the width of the network. So we take 336$\times$336 images as input and increase three convolution layers after inception model. At the same time, the image local features (which can also called multi-label) extracted from IC-CNN, are provided for OD-CNN. Then, the feature vector $p1$ and another feature vector $v1$ are injected into the hidden layer $H1$, where the feature vector $p1$ is the global feature extracted by the exchanged BN-Inception, and the feature vector $v1$ is the local feature that it is transformed from the corresponding probabilities of the object model generated by the OD-CNN network via the hidden layer $H0$. Finally, another hidden layer $H2$ is adding into our network, connected with the softmax layer to predict the category of the scene, as Fig. 4 shown.

C. Training

We have two relatively independent pre-training models, one is OD-CNN and the other is IC-CNN. In the first place, for OD-CNN training, we use appropriate technical methods to train its parameters. Here we divide the data set into training set and test set, we consider the intersection over union which is over 0.5 between the bounding boxes of

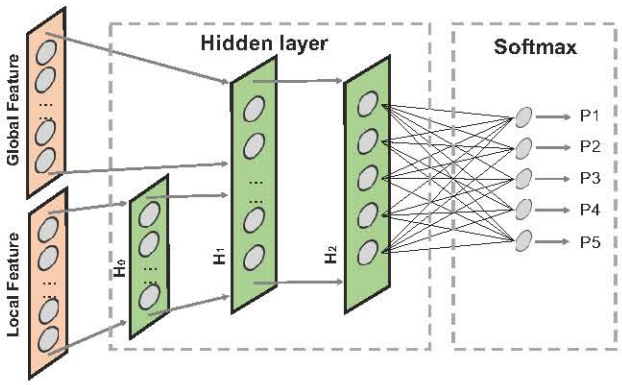

Fig. 4. This figure illustrates the feature technique. We get image local feature from $H0$. Then $H2$ combine the global feature from IC-CNN and local feature via $H1$ and $H2$. Finally, we get the classification result after a fully connected layer.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
CLASSIFICATION ACCURACY OF DIFFERENT PRE-TRAINED MODELS ON THE TEST DATA SET

Model network architecture	TOP1 Accuracy
Normal BN-Inception	88.9%
IC-CNN	89.2%
Normal BN-Inception + object detection	89.7%
IC-CNN + object detection	89.8%
Normal BN-Inception + OD-CNN	89.8%
Joint-CNN(IC-CNN + OD-CNN)	90.1%

a proposed region and ground-truth as positive and rest of the regions as negatives. Moreover, the same with OD-CNN training, IC-CNN uses a random gradient adjusted to fine-tune our parameters. While training the hidden layer, we define the cost function as follows:

$$L(\theta) = \frac{1}{M}\sum_{i}^{M} L_\theta(f^i) + \gamma R(\theta), \tag{1}$$

where, $i \in \{i, \ldots, N_s\}$ is scene label, θ is a parameter of three hidden layers $H0$, $H1$ and $H2$. f is testing feature, $L_\theta(f^i)$ is data error from the last layer, $\gamma R(\theta)$ is a regularization term that penalizes large weights to improve generalization. We use stochastic gradient descent (SGD) to optimize the network to find the weight θ by minimizing the loss $L0$ over the data. Finally, we update the parameters with the last equation as follows:

$$V_{t+1} = \mu V_t - \alpha \bigtriangledown L'(\theta_t), \tag{2}$$

$$\theta_{t+1} = \theta_t + V_{t+1}, \tag{3}$$

where μ is the momentum and α is the learning rate. $Vt+1$ is the gradient of the model parameter θ.

IV. Experiment

We obtain a better experimental result by training and adapting with fine turning on Places2. Then, the data which have been experimented on large-scale data sets to prove that our proposed method is effective. In this section, we evaluate the generalization of our model and comparison results among recent methods. Applying them on two other data sets: MIT Indoor67 and SUN397 [24], which are used as standard benchmarks for scene recognition.

From Table I, we can see that Joint-CNN outperform both normal BN-Inception and BN-Inception with object detection. After fine-tuning both IC-CNN and OD-CNN, the performance of our method is further improved.

From Table II, the best comparable result is found by Joint-CNN. With AlexNet [11] and VGGNet-16 [25], proposals are generated using which are then projected into the feature map or deeper network structure of CNN to extract feature. Then, some methods like VSAD [26] and LS-DHM [27] use refining features to train and classify binary SVM or more complex classifiers. We use our object proposal strategy instead of the method in Fast R-CNN net and use modified BN-Inception as main network for scene classification. Our Joint-CNN detector is approximately 10 times faster than Fast R-CNN as we only classify around 200 regions instead

TABLE II
COMPARISON OF JOINT-CNN WITH OTHER METHODS ON THE MIT67 AND SUN397 DATA SETS

Method \ Data set	MITIndoor67	SUN397
ImageNet-VGGNet-16 [18]	67.6%	51.7%
Places205-AlexNet [25]	68.2%	54.3%
Places205-GoogleNet [25]	74.0%	58.8%
Places205-CNDS-8 [25]	76.1%	60.7%
Places205-VGGNet-16 [25]	81.2%	66.9%
LS-DHM [27]	83.8%	67.6%
VSAD [26]	84.9%	71.7%
Joint-CNN	86.9%	72.4%

of thousands of regions presented on SUN and MIT 67 data sets. We found that how object detection performance varies with the complexity of the data set and different number of region proposals. The combination of appropriate object detection (OD-CNN) technique and Corresponding network framework (IC-CNN) presents a good scene classification effect.

V. Conclusion

In this paper, we propose a novel framework JOINT-CNN for indoor scene classification by using the vectors generated from OD-CNN and IC-CNN to classify the scene images. In OD-CNN, we try to detect more objects' information as soon as possible to generate image labels. And in IC-CNN, we consider the global feature from our exchanged BN-Inception model and the local feature from OD-CNN as hidden layer's input, and then make a classification. By doing this, we get improving performance compared with state-of-the-art approaches.

References

[1] Y. Lee, C. Y.cun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[2] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Int'l Conf. on Machine Learning (ICML)*, pp. 448–456, 2015.
[3] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," *IEEE Int'l Conf. on Computer Vision (ICCV)*, pp. 309–316, 2009.
[4] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *arXiv preprint arXiv:1312.4894*, 2013.
[5] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," *European Conf. on Computer Vision (ECCV)*, pp. 316–329, 2008.
[6] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: Single-label to multi-label," *Computer Science*, 2014.
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. on Pattern Analysis & Machine Intelligence (T-PAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017.
[8] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2285–2294, 2016.
[9] R. W. Zurek and L. J. Martin, "Interannual variability of planet-encircling dust storms on mars," *J. of Geophysical Research Planets*, vol. 98, no. E2, pp. 3247–3259, 1993.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[10] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 923–930, 2013.
[11] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Int'l Conf. on Neural Information Processing Systems (NIPS)*, pp. 487–495, 2014.
[12] E. T. Rolls and M. A. Arbib, "Visual scene perception," 2003.
[13] Q. Zhang and S. Kamata, "Improved optical model based on region segmentation for single image haze removal," *Int'l J. of Information and Electronics Engineering (IJIEE)*, vol. 2, no. 1, pp. 62–68, 2012.
[14] C. Chen, "Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors," *Remote Sensing*, vol. 8, no. 6, p. 483, 2016.
[15] T. Sun, Y. Wang, J. Yang, and X. Hu, "Convolution neural networks with two pathways for image style recognition," *IEEE Trans. on Image Processing (TIP)*, vol. 26, no. 99, pp. 4102–4113, 2017.
[16] Q. Zhang and S. Kamata, "A novel color descriptor for road-sign detection," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Science*, vol. E96-A, no. 5, pp. 971–979, 2013.
[17] A. Quattoni and A. Torralba, "Recognizing indoor scenes," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 413–420, 2001.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Int'l Conf. on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.
[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.
[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
[22] Q. Zhang and S. Kamata, "A novel color space based on rgb color barycenter," *IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1601–1605, 2016.
[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Int'l Conf. on Machine Learning (ICML)*, pp. 807–814, 2010.
[24] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.
[25] L. Wang, C. Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," *Computer Science*, 2015.
[26] Z. Wang, L. Wang, Y. Wang, B. Zhang, and Q. Yu, "Weakly supervised patchnets: Describing and aggregating local patches for scene recognition," *IEEE Trans. on Image Processing (TIP)*, vol. 26, no. 4, pp. 2028–2041, 2017.
[27] S. Guo, W. Huang, L. Wang, and Q. Yu, "Locally supervised deep hybrid model for scene recognition," *IEEE Trans. on Image Processing (TIP)*, vol. 26, no. 2, pp. 808–820, 2017.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)