

A Case Study on Scene Recognition Using an Ensemble Convolution Neural Network

Bongjin Oh *, Junhyeok Lee **

* Software Contents Laboratory ETRI, Daejeon Korea

** Department of Research KPST, Daejeon Korea

bjoh@etri.re.kr, jun@kpst.co.kr

Abstract— this paper proposes architecture to recognize scene images based on an ensemble of two convolution neural networks. A convolution neural network is used to train massive scene images, and the other convolution neural network is used to extract objects from the scene images. The object lists are stored according to scene classes, and used as a clue to decide the top-1 and top-5 classes during scene image recognition stage.

Keywords— scene recognition; ensemble deep neural network; object detection; convolution neural network; deep learning

I. INTRODUCTION

Convolution Neural Network (CNN) has been the most useful for object classification since The AlexNet won the 1st place in ImageNet 2012[1]. CNN is a type of Artificial Neural Network effective in areas such as image recognition and classification. Many CNN based networks such as the GoogleNet, the VGG16(also called OxfordNet) and the ResNet were proposed, and now computer vision shows better object classification accuracy than human [2, 3, 7, 8]. CNN is also used for object detection such as the Faster-RCNN and the YOLO [4, 5]. Object detection includes classification and localization of object(s) in each image. The Faster-RCNN has dramatically increased both of accuracy and speed of object detection in ImageNet 2015.

Scene recognition is also one of the hallmark tasks of computer vision, allowing definition of a context for object recognition. Places-CNNs trained a very large scale scene

images with 1.8 million images including 365 scene classes for scene recognition [6]. Although computer vision reaches human-level object recognition, it's difficult for computer to recognize scene classes with human intelligence yet.

Here we introduce an Ensemble Convolution Neural Network (CNN) architecture comprises of a CNN based object classification network and a CNN based object detection network to train a large scene dataset. Experimental analysis shows better result than Places-CNNs.

This paper is organized as followings. Section II describes the overall proposed an Ensemble CNN in detail, and the experimental analysis is explained in Section III. Finally, Section IV concludes our works and describes future works.

II. THE EMSEMBLED DEEP NEURAL NETWORKS

A. Overview

The Figure 1 shows the overall architecture of the proposed ensemble DNN for scene recognition with a large scale scene dataset. The Ensemble CNN also consists of scene training stage and scene test stage.

B. Training stage of the Ensemble CNN

Training stage is composed of large scene dataset, Object Generator and Scene Trainer modules.

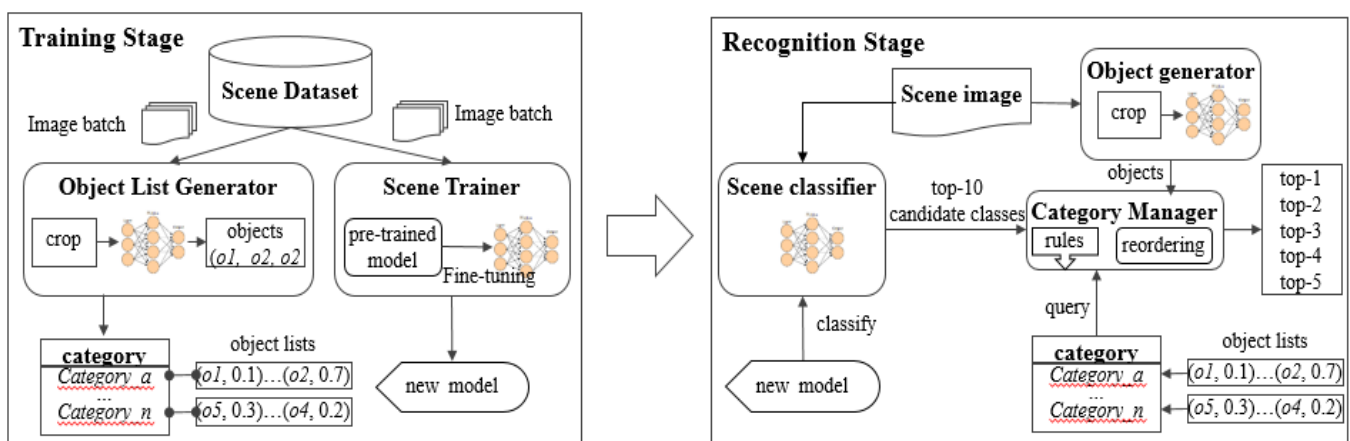


Figure 1. The proposed Ensemble Convolution Neural Network architecture

Scene dataset stores a lot of labeled scene images as a form of LMDB to train, validate and test. Mini-batches of dataset are inputted to both of Object Generator and Scene Trainer.

Object Generator multiply crops the input image, and classifies them using CNN based network, Resnet-200 is used in this paper, and accumulates the top-10 objects to list of the input image's category. Each object has count ratio to represent strength between objects and scene categories. The scene image contains objects with high count ratio may belong to the specific categories.

Scene Trainer trains the scene images to learn hyper parameters, weights and bias of neural networks, using CNN-based network. Scene Trainer fine-tunes pre-trained ResNet-200 of ImageNet 2012 on scene dataset. It is typically accepted that a well initialized CNN shows better accuracy.

C. Scene recognition stage of the Ensemble CNN

Object lists and learned parameters are used to decide top-5 scene categories at scene recognition stage.

When user requires classifying a scene image, CNN based Scene Classifier analyses the input image, and generates top-10 candidate scene categories.

CNN based Object Generator extracts objects from the input image at the same time with scene classification. And then, Class Manager regulates the order of candidate categories using the extracted objects according to decision making rules described as following steps.

- i. Class Manager queries object list for candidate categories from object list repository
- ii. The categories which the length of their object list is shorter than 2 are ignored that is, the order of the category is kept as it is
- iii. For other categories, Class Manager compares the extracted objects and object list of the top-5 categories
- iv. Class Manager reorders candidate categories according to the number of objects contained to object list for each categories
- v. If there are categories have no object, then Class Manager replaces them with alternative categories with matched objects from low top-5 candidate categories randomly

III. EXPERIMENTAL ANALYSIS

A. Environments

A Deep learning workstation consists of dual Xeon processors, 128GB RAM, and three Titan X GPU cards are used to accelerate training speed. MIT's 168GB-size Places365-challenge training dataset is used to generate object lists and learn hyper parameters of scene classifier.

Torch framework is used to training dataset after changing learning rate factor of all convolution layers (for both weight and bias) to 0 and learning rate factor of linear layer to 10 for weight and 20 for bias. This makes learning speed of linear layer faster by more 10 times. Pre-trained ResNet-200 (ImageNet 2012) are re-trained with the training dataset of Places365-Challenge on our system for 3 weeks.

TABLE 1. SYSTEM ENVIRONMENTS FOR EXPERIMENTAL ANALYSIS

Class	Items	Details
HW	CPU	Xeon processor dual socket (16 cores)
	RAM	DDR4 - PC17000 ECC 128GB
	GPU	NVIDIA Maxwell TITAN X * 3
SW	OS	Ubuntu 14.02
	CNN	ResNet-200 over Torch
	Training Parameters.	- Mini-batch of 192 (64 per GPU) - base learning rate: 0.01 - number of data loading threads: 16 - number of epoch: 20
	Training Dataset	Places365-Challenge (8 million images, 365 scene categories, ~ 40,000 images per category)

IV. RESULTS

The proposed Ensemble CNN got the better accuracy than those of Places365-CNN networks with Places365 validation dataset for both the top-1 class and top-5 classes. The Ensemble CNN got 14th rank at the Scene Recognition field of ImageNet 2016 Challenge.

TABLE 2. COMPARISON OF SCENE RECOGNITION PERFORMANCE

CNN Model	Top-1	Top-5
Places365-AlexNet	53.31%	82.75%
Places365-GoogleNet	53.39%	84.01%
Places365-ResNet	54.65%	85.07%
Places365-VGG	55.19%	85.01%
Proposed CNN	57.27%	87.71%

V. CONCLUSION

This paper proposes an Ensemble CNN to train and recognize a large scale scene images. The experimental results show the kinds of objects in the scene images are useful as an important key to distinguish scene classes among ambiguous images during scene recognition. The hybrid CNN outperforms the Places365-ResNet for both top-1 and top-5 accuracy by 3%.

ACKNOWLEDGMENT

This research was supported by a grant (Research Project ID: 17TLRP-B079209-04) from Transportation & Logistics Research Program funded by Ministry of Land, Infrastructure and Transport of the Republic of Korea.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. Hinton., "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, pp. 1106-1114, 2012.
- [2] C. Szegedy, W. Liu, and Y. Jia et al. "Going Deeper with Convolutions," arXiv:1409.4842, 2014.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, 2015.
- [4] S. Ren, K. He, R. Girshick, and K. Sun, "Faster-RCNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Neural Information Processing Systems*, 2015.
- [5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," arXiv:1506.20640, 2015.
- [6] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An Image Database for Deep Scene Understanding," arXiv:1610.02055, 2016.
- [7] O. Ryssakovsky, J. Deng, and L. Fei-Fei et al., "ImageNet Large Scale Visual Recognition Challenge," arXiv:1409.0575, 2014.
- [8] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, 2014



Bong-Jin Oh received B.S. and M.S. degrees in computer science from Pusan National University, Busan, Korea in 1993 and 1995 respectively, and the Ph.D. degree from Chungnam National University, Daejeon, Korea in February 2012. Since 1995, he has been with the Electronics and Telecommunications Research Institute (ETRI), where he develops home network middleware, data broadcasting middleware and big-data processing platform. His current research areas are IoT middleware, data broadcasting middleware, IPTV, big data analytics, and deep learning frameworks.



JunHyeok Lee received B.S degree in computer science from Korean National Open University, Korea in 2008. Since 2010, he has been the CEO of Korean Platform Service Technology (KPST) which is a company specializing in embedded software, image processing, and deep learning technology, and is engaged continuously in technology development and R&D research and development with Daedeok Innopolis Venture Association and research institutes. His research interests are artificial intelligence (machine learning), embedded emulators, authoring tools, modelers, platform middleware technology, image processing, and retrieval technology.