A Deep Learning Framework Using Convolutional Neural Network for Multi-class Object Recognition

Shaukat Hayat, She Kun, Zuo Tengtao, Yue Yu, Tianyi Tu, Yantong Du
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
e-mail: hayat.uestc@yahoo.com, kun@uestc.edu.cn, 2303629434@qq.com, 499676745@qq.com

***Abstract*—Object recognition is classic technique used to effectively recognize an object in the image. Technologies specifically in field of computer vision are expected to detect and recognize more complex tasks with help of local features detection methods. Over the last decade, there has been sustained increase in the number of researchers from various kind of disciplines i.e. academia, industry, security agencies and even from general public has caught an attention to explore the covered aspects of object detection and recognition concerned problems. It is further significantly amended by adopting deep learning model. In this paper, we applied deep learning to multi-class object recognition and explore convolutional neural network (CNN). The convolutional neural network is created with normalized standard initialization and trained with training set of sample images from 9 different object categories plus sample test images using widely varied dataset. All results are implemented in python tensorflow framework. We examine and compared CNN results with final feature vectors extracted from variant approaches of BOW based on linear L2-SVM classifier. Based on it, sufficient experiments verify our CNN model effectiveness and robustness with rate of 90.12% accuracy.**

Keywords-multi-class object recognition; deep learning; convolutional neural network; computer vision; bag-of-visual words (BOW); L2-SVM

I. INTRODUCTION

Technological advancements are being used by humans for assistance to complete their daily life activities. Dealing with objects having different characteristics includes a variety of shapes, sizes, surface material, orientation and degree of mobility are considered to be key issues [1]. Recognizing various kind of objects from images in daily life activities is one of the most encouraging applications of computer vision. Object detection and recognition can be defined; is to provide an image, text or video frame, to search out all the targeted locations with the application of specialized algorithm-based models and to represent each target with a specific category where it may belong. Humans can easily recognize and localizes objects within chaotic scenes in their daily life. In contrast, the computer cannot do easily and while doing so, it needs to be programmed well [2]. Over the last decade, there has been a sustained increase in the number of researchers from various kind of disciplines i.e. academia, industry, security agencies and even from the general public has caught an attention to exploring the covered aspects of object detection and recognition concerned problems [3].

Presently, Deep Neural Networks (DNNs) have been applied to various nature of problems and gets very effective outcomes. Specifically, Convolutional Neural Networks (CNN) model provides good results to handle related problems with natural language processing [4], computer vision[5], image segmentation[6] and classification[7].

Figure 1. Two randomly image samples of nine different object classes from Caltech-101 object categories

In this paper we propose a multi-class object recognition task considering DNNs approach. The recognition model is based on Convolutional Neural Network (CNN) by the most popular framework tensorflow. The image data used for CNN is set to nine kinds of various objects such as Airplane, Bonsai, Chandelier, Faces-easy, Hawksbill, Ketch, Leopard, Motorbike, and Watch. We use challenging dataset for object recognition, Namely, Caltech-101[8] object categories. This database contains from 31 to 800 images

978-1-5386-4991-6/18/$31.00 ©2018 IEEE

per category. It contains 9,144 images plus has one background category [9]. For this experiment we selected nine different object categories and split into a training set with 900 images and test set with 145 images. In general, the appearance of objects in images are more complex and have cluttered background scenes. The considered object classes for this experiment are shown in Figure 1.

Secondly, a regularization technique i.e. data augmentation and dropout are used to combat the potential overfitting and for improving the model accuracy respectively.

Finally, for weight initialization at every layer of the model, we employed normalized weight initialization approach defined by [27] is concisely described in section II. On the other hand, to accelerate the performance, there are several other deep learning techniques such as AlexNet[10],

GoogleNet [11] and Residual Networks (ResNets) [12] used for recognizing faces, scenes, objects and text character. Nevertheless, some of the classical techniques also used in different fields to improve the performance of recognition such as Bag-of-Visual Words (BOW) [13]. This classical image descriptor convoluted in feature extraction [14], using unsupervised machine learning (ML) approaches to construct codebooks like K-means clustering[15], spectral clustering[16], pooling clusters by local constrained linear coding [17], and utilization of fast minimum spanning tree [18].

According to [19] BOW method is extensively used and popular in computer vision field. In addition, various authors [20-22] threw light on the performance of the BOW techniques itself as well as with combination of R-CNNs features plus HOG-BOW based deformable part models and CCN augmentation with L2-Support vector machine (L2-SVM) considering other local features vectors descriptors outperforms with other feature vector algorithms. Therefore, in this research, we evaluated CNN model on nine different objects classes and compared the results with BOW and HOG-BOW different approaches trained on a novel dataset consisting of 5 different class objects of wild animals [13].

The rest of this paper is organized as follow: First, we introduce Neural Network generally and Convolutional Neural Network specifically in section II. Section III provides the experimental analysis. Results are provided in section IV. Conclusion and future work are presented in section V.

II. NEURAL NETWOK

The promising intellectual life is to be measured through learning ability. ML is now capable to learn and predict with a more sophisticated way to classify any unknown phenomenon from given datasets. Artificial neural networks (ANNs) is neuro-biologically inspired. The human brain is composed of complex multi-layers nerve cells in form of neurons. The convolutedness of real neurons is highly abstracted, but ANNs is governed through programming archetype based on some observational data and thus it permits the computers to learn and gives some predictable decision [23]. Many models are inspired by ANNs are Convolutional Neural Networks, Recurrent Neural Networks, Deep Belief Networks.

A. CNN- Convolutional Neural Network

The Convolutional neural networks (CNN) is deep learning method, recently it has stepped forward and vivid development in the field of computer vision such as an image segmentation, object detection, recognition, and captioning [24]. Indeed, it biologically inspired by the human brain. Convolutional nets and other related architectures under the deep learning umbrella are at best comparable to the neural networks exist in the human brain. Just like the structure and operations of the biological neurons in the human visual cortex through the deployment of hierarchical multi-layers networks, the similar way the deep neural network has to be reveal very effective to learn the various schemes of feature representation getting from training data. Its operations are automatic and features engineering tasks could be resolve with more fast and reliable way. They are quite able to find and effectively use specific peculiarities of image classes in case a massive training dataset is provided. In the early 90s, CNN was first launched for the purpose recognizing handwritten digits [25]. Later on, in 2012 a major development was made by releasing AlexNet [10]. There is a basic principle that needs to be considered in a special case for multilayer perceptron where every neuron is connected to the receptive field located in forward-face. Moreover, the neurons belong to each layer in the network shares the same weights.

CNN using back-propagation algorithm to carry on the learning process. CNN use stochastic gradient decent to update weighting filter and coupling coefficient. Along these, a non-linear task namely pooling (sub-sampling) and convolutional operations are utilized for recognizing the optimized features [26]. To accelerate the training process CNN used the most notable non-saturated activation function, Rectified Linear Unit (ReLU) for the task of category recognition.

For the task of object recognition, it may be divided into two main parts: object recognition and object detection. In this paper, we only focus on multi-class object recognition. However, extending existing recognition models to multi-class objects detection task needs to modify the architecture of the model. Tensorflow library provides full-support for training, testing, tuning and facilitate models deployment with well-documented examples for all these tasks. We successfully applied five layers CNN model for recognition with non-linear activation function Rectified Linear Unit (ReLU) for recognition purposes. Fig. 2 shows typical convolutional neural network architecture. We initialized biases with 0 value and for the initial scale of the weights initialization Wij at every layer, the heuristic approach suggested by [27] was adopted as standard initialization to design the proposed CNN model:

$$Wij \sim U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}] \quad (1)$$

where U represents uniform distribution in the interval $(-a, a)$, while n denotes the size of previous layer. This standard initialization in equation (1) and the following property mentioned below can cause the increase to variance:

$$nVar[W] = \frac{1}{3}$$

here n shows is the layer size and all layer sizes are to assume same. This will cause a change of the back-propagated gradient to be dependent on the layer.

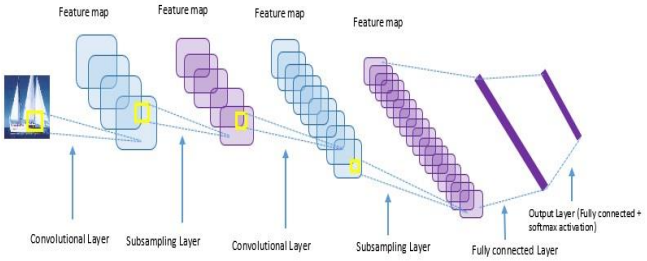

Figure 2. CNN- Convolutional Neural Network

Normalization factor is significant while initializing deep network and suggested initialization process in order to nearly satisfy the intentions to retain the activation changes and back propagated gradients difference one moves up or down the network [27]. So it is known to be normalized initialization as we implemented:

$$W \sim U[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}]$$

The network having multiple layers and each of which consists one or more planes. Each part in a plane gets input from a small neighborhood in the planes of the previous layer.

III. Experimental Analysis

A. Environments

The DNN framework consists of dual Xeon processor with 32GB RAM and GPU cards. Ubuntu 16.04 operating system with the 64bit environment is used to perform model training. We used Python programming language including its associative packages and libraries such as Tensorflow, Anaconda assists to recognize objects in image dataset. Tensor Flow framework is utilized for training the model which uses the convolutional neural network (CNN).

B. Dataset

900 color images, representing nine-classes of different objects, were collected from the (http://www.vision.caltech.edu/feifeili/Datasets.htm). The number of images for each category are summarized in Table I.

TABLE I. Image Number (Train/Test) for Each Object Category

Category	Airplane	Bonsai	Chandelier	Faces-easy
Image-No:	100/20	100/20	100/7	100/20
Category	Hawksbill	Ketch	Leopards	Motorbikes
Image-No:	100/4	100/14	100/20	100/20
Category	Watch		Total Train: 900	
Image-No:	100/20		Total Test: 145	

Caltech-101 [8] is not without shortcomings due to its diverse nature. These images are widely varied about the size and quality. Some of the images were in tidy position, some have uniform background while some of them having cluttered background. Most images are medium resolution, i.e. about 300 × 300 pixels. So it is really a complex dataset and then randomly divided into two separate parts for training and test images resulting in 900 training with 100 samples per category and 145 test images.

C. Regularization Techniques

Data augmentation has long been used for training CNN. In the field of computer vision, it is considered ubiquitous due to its ease to putting it into practice and usefulness. Image transformations steps such as mirroring or cropping of image-data can cause to increase model accuracy and robustness. We applied real-time data augmentation on images while training the model which could flip, resize and shear it.

A dropout technique is commonly used in image vision field while training deep learning model. Dropout is executed by setting concealed unit activations to zero with some settled likelihood during the training phase. All activations are kept while assessing the network model but the subsequent yield is scaled by the dropout likelihood. In this work we applied the dropout method in similar way.

IV. Results

In this experiment deep learning approach, CNN has been used to implement the overall recognition process. The fine- tuning of CNN model was done using backpropagation algorithm. We compare the performance of the proposed model with BOW and HOG-BOW variant approaches [13]. Subsequently, the performance of the algorithm for recognizing multi-class objects, using CNN method is statistically evaluated in terms of accuracy, loss and time efficiency and compared with the previously conducted study as shown in Table II on multi-class object recognition by [13]. Fig. 3-a shows the test accuracy of our proposed model using a different number of epochs on test subset of model development. We could figure out that after 160 epochs the accuracy of the testing set reaches to 0.91 while the training loss become decreases to 0.48. Dropout of 0.5 is

added between two layers, softmax is used as an activation function of the last layer.

It is noticeable (Table II) that CNN network return improved accuracy in comparison with final feature vectors extracted from variant approaches of BOW based on linear L2-SVM classifier. The proposed network trained with better training loss (Fig. 3–b).

The results show that our proposed CNN model is competitive when compared to the classical BOW method and require less computing time.

TABLE II. COMPARISON TO OTHER METHODS

Methods	Testing Accuracy
BOW-Color with Max-Pooling	84.00
BOW-Color with Sum-Pooling	82.40
BOW-Gray with Max-Pooling	82.00
BOW-Gray with Sum-Pooling	81.40
HOG-BOW-Gray with Sum-Pooling	82.60
HOG-BOW-Gray with Max-Pooling	78.40
HOG-BOW-Color with Sum-Pooling	73.20
HOG-BOW-Color with Max-Pooling	63.60
CNNs (proposed)	**90.12**

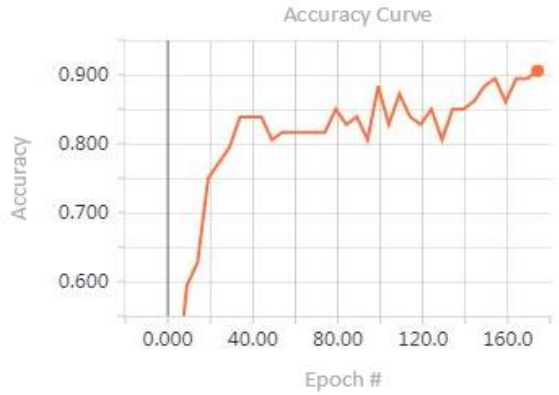

Figure 4-a. Performance curve of CNN configuration with different number of epochs- test accuracy

Figure 4-b. Training loss curve of CNN configuration with different number of epochs- training loss

V. CONCLUSION

To improve the performance of CNN model for multi-class object recognition, initial scale of the weights initialization for every layer a heuristic approach was adopted as standard initialization. The proposed model was further tuned and the recognition performance was improved. We use nine different object classes taken from wide varied image dataset caltech101 and deploy five layers Convolutional Neural Network model. We compared the performance of the proposed model with different classical bag-of-words (BOW) approaches trained on a novel dataset consisting of 5 different class objects of wild animals. Concluded from the results tthat performance of CNNs model with test accuracy rate 90.12% is much better than different classical BOW approaches.

In the future work, we will focus on other applications for object detection and recognition by using deep learning approaches.

ACKNOWLEDGEMENT

This work is supported by Sichuan Science and Technology support program under grant No. 2016GZ0073.

REFERENCES

[1] Martinez-Martin, E. and A.P.d. Pobil, *Object Detection and Recognition for Assistive Robots: Experimentation and Implementation*. IEEE Robotics & Automation Magazine, 2017. 24(3): p. 123-138.

[2] Zhang, Y., H. Wang, and F. Xu. *Object detection and recognition of intelligent service robot based on deep learning*. in *2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. 2017.

[3] Turaga, P., et al., *Machine recognition of human activities: A survey*. IEEE Transactions on Circuits and Systems for Video technology, 2008. 18(11): p. 1473-1488.

[4] Bhatnagar, S., et al., *IITP at SemEval-2017 Task 5: An Ensemble of Deep Learning and Feature Based Models for Financial Sentiment Analysis*. 2017.

[5] Szegedy, C., S. Ioffe, and V. Vanhoucke, *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. 2016.

[6] Girshick, R., et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[7] Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks*, in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. 2012, Curran Associates Inc.: Lake Tahoe, Nevada. p. 1097-1105.

[8] Fei-Fei, L., R. Fergus, and P. Perona, *Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories*. Computer vision and Image understanding, 2007. 106(1): p. 59-70.

[9] Li, F.-F., R. Fergus, and P. Perona. *Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories*. in *2004 Conference on Computer Vision and Pattern Recognition Workshop*. 2004.

[10] Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

[11] Szegedy, C., et al. *Going deeper with convolutions*. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[12] He, K., et al. *Identity Mappings in Deep Residual Networks*. 2016. Cham: Springer International Publishing.

[13] Okafor, E., et al. *Comparative study between deep learning and bag of visual words for wild-animal recognition*. in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2016.

[14] Wang, C. and K. Huang, *How to use Bag-of-Words model better for image classification*. Image and Vision Computing, 2015. 38: p. 65-74.

[15] Ye, P., et al. *Unsupervised feature learning framework for no-reference image quality assessment*. in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[16] Passalis, N. and A. Tefas, *Spectral Clustering using Optimized Bag-of-Features*, in *Proceedings of the 9th Hellenic Conference on Artificial Intelligence*. 2016, ACM: Thessaloniki, Greece. p. 1-9.

[17] Wang, J., et al. *Locality-constrained Linear Coding for image classification*. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010.

[18] Jothi, R., S.K. Mohanty, and A. Ojha. *Fast Minimum Spanning Tree Based Clustering Algorithms on Local Neighborhood Graph*. 2015. Cham: Springer International Publishing.

[19] Csurka, G., et al. *Visual categorization with bags of keypoints*.

[20] Coates, A., A. Ng, and H. Lee. *An analysis of single-layer networks in unsupervised feature learning*. in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011.

[21] Girshick, R., et al. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation*. in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[22] Razavian, A.S., et al. *CNN features off-the-shelf: an astounding baseline for recognition*. in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. 2014. IEEE.

[23] Sikora, R., et al. *Artificial neural network application for material evaluation by electromagnetic methods*. in *Neural Networks, 1999. IJCNN'99. International Joint Conference on*. 1999. IEEE.

[24] Schmidhuber, J., *Deep learning in neural networks: An overview*. Neural Networks, 2015. 61: p. 85-117.

[25] LeCun, Y., et al. *Handwritten digit recognition with a back-propagation network*. in *Advances in neural information processing systems*. 1990.

[26] Kang, L., et al. *Convolutional Neural Networks for Document Image Classification*. in *2014 22nd International Conference on Pattern Recognition*. 2014.

[27] Glorot, X. and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, T. Yee Whye and T. Mike, Editors. 2010, PMLR: Proceedings of Machine Learning Research. p. 249--256.