Object Detection Using Convolutional Neural Network

D. Badrinarayana[1], A. Sai Vamsi Karthik[2]
[1, 2] Panimalar Institute Of Technology

Abstract- *Supervised machine learning based state-of-the-art computer vision techniques are in general data hungry. Their data duration poses the challenges of expensive human labelling, inadequate computing resources and larger experiment turn around times. Training data subset selection and active learning techniques have been proposed as possible solutions to these challenges. A special class of subset selection functions naturally model notions of diversity, coverage and representation and can be used to eliminate redundancy thus lending themselves well for training data subset selection. Our results show that diversity based subset selection done in the right way can increase the accuracy by upto 5 - 10% over existing baselines, particularly in settings in which less training data is available. This allows the training of complex machine learning models like Convolutional Neural Networks with much less training data and labelling costs while incurring minimal performance loss.*

I. INTRODUCTION

There is an ever-increasing amount of image data in the world, and the rate of growth itself is increasing. Infotrends estimates that in 2016 still cameras and mobile devices captured more than 1.1 trillion images. According to the same estimate, in 2020 the will increase to 1.4 trillion. Many of these images are stored in cloud services or published on the Internet.

In 2014, over 1.8 billion images were uploaded daily to the most popular platforms, such as Instagram and Facebook.

Going beyond consumer devices, there are cameras all over the world that capture images for automation purposes. Cars monitor the road, and traffic cameras monitor the same cars. Robots need to understand a visual scene in order to smartly build devices and sort waste. Imaging devices are used by engineers, doctors and space explorers alike.

To effectively manage all this data, we need to have some idea about its contents. Automated processing of image contents is useful for a wide variety of image-related tasks. For computer systems, this means crossing the so-called semantic gap between the pixel level information stored in the image files and the human understanding of the same images. Computer vision attempts to bridge this cap.

II. PROBLEM STATEMENT

Objects contained in image files can be located and identified automatically. This is called object detection and is one of the basic problems of computer vision. As we will demonstrate, convolutional neural networks are currently the state-of-the-art solution for object detection. The main task of this thesis is to review and test convolutional object detection methods.

In the theoretical part, we review the relevant literature and study how convolutional object detection methods have improved in the past few years.

In the experimental part, we study how easily a convolutional object detection system can be implemented in practice, test how well a detection system trained on general image data performs in a specific task and explore, both experimentally and based on the literature, how the current systems can be improved.

III. RELATED WORK

Chungkeun Lee1, H. Jin Kim1∗ and Kyeong Won Oh in his “**Comparison of Faster R-CNN models for object detection”** Object detection is one of the important problems for autonomous robots. Faster R-CNN, one of the state-of-the-art object detection methods, approaches real time application; nevertheless, computational time lies borderline of real time application, i.e. 5fps with VGG16 model in K40 GPU system. Moreover, computation time depends on model and image crop size, but precision is also affected; usually, time and precision have trade-off relation. By adjusting input image size in spite of downgrading performance, computation time meets criteria for one model. Therefore, selection of a model is one of the important problems when faster R-CNN based object detection system for an autonomous robot is constructed. In this paper, we convert several state-of-the-art models from convolution neural network (CNN) for image classification. Then, we compare converted models with several image crop size in terms of computation time and detection precision. We will utilize those comparison data for selecting a proper detection model in case a robot needs to perform an object detection task.

Zhirong W, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, Jianxiong Xiao work on “3D ShapeNets: A Deep Representation for Volumetric Shapes” 3D shape is a crucial but heavily underutilized cue in today's computer vision systems, mostly due to the lack of a good generic shape representation. With the recent availability of inexpensive 2.5D depth sensors (e.g. Microsoft Kinect), it is becoming increasingly important to have a powerful 3D shape representation in the loop. Apart from category recognition, recovering full 3D shapes from view-based 2.5D depth maps is also a critical part of visual understanding. To this end, we propose to represent a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid, using a Convolutional Deep Belief Network. Our model, 3D ShapeNets, learns the distribution of complex 3D shapes across different object categories and arbitrary poses from raw CAD data, and discovers hierarchical compositional part representation automatically. It naturally supports joint object recognition and shape completion from 2.5D depth maps, and it enables active object recognition through view planning. To train our 3D deep learning model, we construct ModelNet - a large-scale 3D CAD model dataset. Extensive experiments show that our 3D deep representation enables significant performance improvement over the-state-of-the-arts in a variety of tasks.

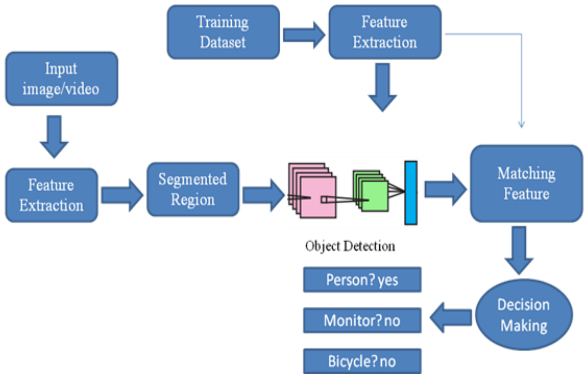
IV. PROPOSED SYSTEM

Deep models are neural networks with deep structures. The history of neural networks can be traced back to the 1940s. It was inspired by simulating the human brain system and the goal was to find a principled way to solve general learning problems. It was popular in 1980s and 1990s. In 1986, Rumelhart, Hinton, and Williams published backpropagation in Nature, and it has been widely used to train neural networks until now. In the following subsections, we will introduce the structure of multilayer neural networks, feedforward operation used to predict output from input, and backward propagation.

The computation units of neural networks are called neurons and are organized into multiple layers. Neurons in adjacent layers are connected with weights. However, neurons in the same layer are not connected. In feedforward operation, neurons in a lower layer pass signals to neurons in its upper layer. A neuron is activated if its received signals are strong enough. Similar to the brain, some connections between neurons are stronger, while some are weaker, indicated by different weights.

Architecture Diagram

In this section, we focus on the system workflow and the used methods in our system. We firstly extract the image features of object’s templates by deep learning after the data is preprocessed, which is a very important part of the image classification. Then we introduce why and how we use inception v4 network to extract image features. At the end of this part, we briefly introduce the speech recognition method we use.



Module 1: Pre-processing

The deep learning requires a lot of training data. If the system only learns a picture at a time, it will learn very slowly and inefficient. So we expand the data to improve the system learning speed. We expand one image to 20 images through translation, rotation and other affine changes. We achieve the data augmentation by using the Image Generator function which includes in deep learning package of keras. In order to reduce the impact of the noise, uneven illumination on the feature extraction, we normalize the expanded image so that the data has zero mean and unit variance, then we will get better features.

In recent years, as the emergence of large scale training data, people observed that the performance of *f* on test data got improved when the dimensionality of input data increased, which was called “blessing of dimensionality”, because larger training data required larger learning capacity. As illustrated, the performance of machine learning models with shallow structures (e.g. SVM and Boosting) gets saturated when training data becomes very large because of their limited learning capacity. They face the underfitting problem, i.e. their prediction accuracy on large-scale training data is not satisfactory.

Module 2: Deep Neural Network

It can extract more discriminable and powerful features by deep neural network, and the recognition system based on these features can learn faster and have a higher recognition rate. We mainly consider the recognition accuracy and the speed of feature extraction when we select the framework of the deep neural network. We tried to use VGG, GoogLeNet, ResNet and GoogLeNet inception v4, which have a good performance in object recognition, to extract image features, and do our comparative experiment, these networks were pre-trained on ImageNet.

According to the original publication, Fast R-CNN is more efficient to train than R-CNN, with nine-fold reduction in training time. The entire network (including the RoI pooling layer and the fully-connected layers) can be trained using the back-propagation algorithm and stochastic gradient descent. Typically, a pre-trained network is used as a starting point and then fine-tuned. Training is done in mini-batches of N images. R=N RoIs are sampled from each mini-batch image. The RoI samples are assigned to a class, if their intersection over union (see section 4.6) with a ground-truth box is over 0.5. Other RoIs belong to the background class.

Module 3: The Classifier Model Using Support Vector Machine (SVM)

Once features are extracted, a classifier can be trained to classify a test sample as a member of one of the known classes. In this work the images have been classified using linear Support Vector Machine (SVM). SVM is a supervised learning technique that seeks an optimal hyper-plane to separate two classes of samples. SVM performs classification by finding the hyper plane that maximizes the margin between the two classes as shown in Fig. 14 Support vectors are the data points that lie closest to the decision surface. They are the most difficult to classify. The basic idea of support vector machine is to find an optimal hyper-plane for linearly separable patterns.

SVM classification including a hyper plane that maximizes the separating margin between two classes SVM have some remarkable characteristics such as, its ability to learn independently of the dimensionality of the feature. This classifier has been chosen due to its robustness, simplicity and does not tend to over fit training data.

V. CONCLUSION

We propose an incremental object recognition system based on deep learning. The system can identify a variety of objects according to user’s needs. During the interaction with user, it will become more and more intelligent, and recognize more and more object classes. It has high performance by learning few samples, such as 5 images of an object class. The recognition rate is getting higher and higher, and it can recognize more and more object classes after learning the new samples. We have verified its powerful learning capability, high recognition rate, wide range of applications by testing it on challenging datasets. We also confirmed that object recognition rate of our incremental learning method will be higher with the interaction with user.

VI. FUTURE WORK

One of the strengths of convolutional networks is their inherent translation invariance. Yet, taking the context of the whole image into consideration could potentially create an even more precise system. We experimented with a geometry-based inference system, which alters the probabilities of object detections based on their geometric plausibility. In future we able to design mobile application which can be used for blind people for navigation.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, vol.86, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, vol. 25, pp. 1106–1114, 2012.

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, pp. 1409-1556, 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed,D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 1–9, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 770-778, June, 2015.

[6] Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. pp. 4278-4284, 2016.

[7] C. Papageorgiou T. Poggio, "A trainable system for object detection" Computer Vision, vol. 38, pp. 15-33, 2000.

[8] Agrawal P, Girshick R, Malik J. Analyzing the performance of multilayer neural networks for object recognition[M]. Computer Vision, pp. 329-344, 2014.

[9] Wohlhart, P., & Lepetit, V. Learning descriptors for object recognition and 3D pose estimation. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 3109-3118, 2015.

[10] Li, Fei Fei, R. Fergus, and P. Perona. "Learning generative visual models from few training examples: an

incremental bayesian approach tested on 101 object categories." IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 178-178, 2004.

[11] Polikar R, Upda L, Upda S S, et al. Learn++: an incremental learning algorithm for supervised neural networks[J]. Systems Man & Cybernetics Part C Applications & Reviews IEEE Transactions on, vol. 31, pp. 497-508, 2001.

[12] Franco A, Maio D, Maltoni D. Incremental template updating for face recognition in home environments[J]. Pattern Recognition, pp. 2891- 2903, 2010.

[13] Kim T, Kittler J, Cipolla R. Incremental learning of locally orthogonal subspaces for set-based object recognition[J]. Proc Iapr British Machine Vision Conf, pp. 559-568, 2006.