OBJECT CLASSIFICATION USING DEEP LEARNING

FONG SOON FEI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical-Computer and Microelectronics System)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JUNE 2015

Dedicated to my parents, lecturer, and friends.

ACKNOWLEDGEMENT

First and foremost, I would like to express my appreciation to my project supervisor Dr. Usman Ullah Sheikh for his guidance and patient. My sincerest appreciation goes to Dr Sajad Farokhi who was PhD student of Computer Science for his support, help and technical advices throughout my project. Without his support and interest, my project cannot be completed.

Next, I would like to thank to INTEL MICROELECTRONICS (M) SDN BHD for offering and sponsoring me on this master taught course program. In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts.

Lastly, I would like to express my sincere appreciation and gratitude towards my family members for providing constant support and encouragement to me throughout this journey.

ABSTRACT

Object recognition is a process of identifying a specific object in an image or video sequence. This task is still a challenge for computer vision systems. Many different approaches of object recognition including the traditional classifier or deep neural network were proposed. The objective of this thesis is to implement a deep convolution neural network for object classification. Different architecture and different parameters have been tested in order to improve the classification accuracy. This thesis propose a very simple deep learning network for object classification which comprises only the basic data processing. In the proposed architecture, deep convolution neural network has a total of five hidden layers. After every convolution, there is a subsampling layer which consists of a 2×2 kernel to do average pooling. This can help to reduce the training time and compute complexity of the network. For comparison and better understanding, this work also showed how to fine tune the hyper-parameters of the network in order to obtain a higher degree of classification accuracy. This work achieved a good performance on Cifar-10 dataset where the accuracy is 76.19%. In challenging image databases such as Pascal and ImageNet, this network might not be sufficient to handle the variability. However, deep convolution neural network can be a valuable baseline for studying advanced deep learning architectures for large-scale image classification tasks. This network can be further improved by adding some validation data and dropout to prevent overfitting.

ABSTRAK

Pengenalan objek adalah proses mengenal pasti objek dalam imej atau video. Tugas ini masih merupakan satu cabaran untuk sistem penglihatan komputer. Pelbagai pendekatan berbeza untuk pengenalan objek termasuk pengelas tradisional atau “deep neural network” dibincangkan dalam tesis ini. Objektif projek ini adalah untuk melaksanakan “deep convolution neural network” yang digunakan untuk pengelasan objek. Selain itu, pelbagai seni bina dan parameter diuji untuk meningkatkan ketepatan klasifikasi. Tesis ini mencadangkan “deep learning network” yang mudah untuk pengelasan objek yang terdiri daripada hanya memproses data asas. Dalam seni bina yang dicadangkan, konvolusi dalam rangkaian neural mempunyai lima lapisan tersembunyi. Selepas setiap konvolusi, terdapat lapisan “subsampling” yang terdiri daripada kernel 2×2 untuk melakukan pengumpulan purata. Ini boleh membantu untuk mengurangkan masa latihan dan mengira kerumitan rangkaian. Sebagai perbandingan dan pemahaman yang lebih baik, projek ini juga menunjukkan bagaimana untuk menala parameter-parameter rangkaian untuk mendapatkan ketepatan yang lebih tinggi. Kerja ini mencapai prestasi yang baik pada dataset “cifar-10” di mana ketepatan yang diperolehi adalah 76.19%. Dalam pangkalan data imej yang mencabar seperti “Pascal” dan “ImageNet”, rangkaian ini mungkin tidak mencukupi untuk mengendalikan variasi yang terdapat . Walau bagaimanapun, DCNN boleh menjadi asas untuk mengkaji “deep neural network” untuk tugas pengelasan imej yang lebih besar. Rangkaian ini boleh diperbaiki dengan menambah beberapa data pengesahan dan untuk mengelakkan keciciran “overfitting”.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial intelligence
ANN	-	Artificial neural networks
CNN	-	Convolution neural network
DBN	-	Deep belief networks
DCNN	-	Deep convolution neural network
DNN	-	Deep neural network
GPU	-	Graphics processor unit
GUI	-	Graphical user interfaces
ILSVRC	-	ImageNet Large Scale Visual Recognition Challenge
MATLAB	-	Matrix laboratory
MLP	-	Multilayer perceptron
MSE	-	Mean squared error
PCA	-	Principal component analysis
PCA(Net)	-	Principal component analysis network
RBM	-	Restricted Boltzmann machines
SVM	-	Support vector machines

LIST OF SYMBOLS

w	-	weight
b	-	bias
C	-	cost function
η	-	learning rate

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Project Background

Object recognition is a process of finding and identifying a specific object in a digital image or video sequence. Humans can easily recognize an object in an image even through the object inside the image may vary somewhat in different sizes or scales, different vantage points and even partially obstructed from view. However, object recognition from an image or video is still a challenge for computer vision systems. Even with the help of smart algorithms and human assistants, a classifier in the computer is still unable to catch everything in an image (Sivic and Zisserman, 2003). Many approaches to the task have been implemented over multiple decades.

Object recognition task is successful if the network system is able to label the object based on models of known objects. For example, given an image containing one or more different objects with background, the network system is capable of assigning the labels to a set of regions in the image correctly as showed in Figure 1.1. The classification accuracy of the network system can be calculated by comparing the result with a set of labels corresponding to a set of objects known to the system. The object recognition has a very close relationship with segmentation. This is because if the network system is unable to recognize an object, segmentation cannot be done correctly, and without a good segmentation, object recognition cannot be done as well.

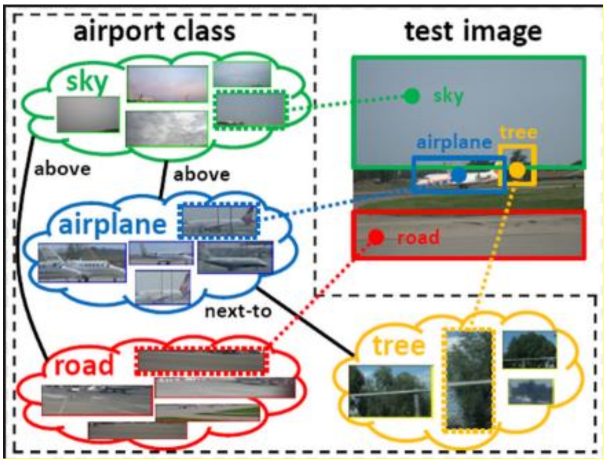

Figure 1.1 Object classification (Zhou *et al.*, 2013)

Machine learning is a set of algorithms that can learn and explore from the construction and recognize the patterns or objects from an input data. Therefore, machine learning can make accurate predictions for previously unseen data. Hence, machine learning can be used as a powerful tool to overcome the challenges in computer vision such as object recognition, natural language understanding, medical imaging, and web search/information retrieval. In the past few decades, machine learning shows that it can be used in many real-world applications and is successful in solving many artificial intelligence (AI) problems (Lee, 2010). For example, it has been successfully applied in practical speech recognition, effective web search, and face detection.

Machine learning gives a handful of labeled examples and able to do binary classification. For example, given ten images, five images of table with the label zero and another five images of not table with the label one. The algorithm of the system starts to learn and identify images of table. After the training process is done and when new images are fed to the network, the network is able to produces the correct label. In other words, the network produce output zero if the image contains a table, and output one if the image does not contain a table. Recently, deep architectures show a good way to do binary representations by extracting the important features and characterizing of the input distribution.

Deep learning also known as deep machine learning, deep structural learning or hierarchical learning is extension algorithms of machine learning that attempts to model higher level of abstractions in data by using complex architectures. The deep learning structural composed of multiple layers and multiple non-linear transformations is used for hierarchical feature (Schmidhuber, 2014). The neural network is shallow if the number of layers of units, regardless of their types, is usually at most two. A deep neural network is deep if it has multiple, usually more than three layers of units. In essence, a neural network is deep when the following two conditions are met. The first condition is the network can be extended by adding layers consisting of multiple units and second condition is the parameters of each layer are trainable (Bengio and LeCun, 2007). From these conditions, it should be understood that there is no absolute number of layers that distinguishes deep neural networks from shallow ones. Rather, the depth of a deep neural network grows by a generic procedure of adding and training one or more layers, until it can properly perform a target task with a given dataset. In other words, the data decide how many layers a deep neural network needs (Cho, Raiko, and Ihler, 2011).

Deep learning tries to move in this direction by capturing a good representation of input data by using compositions of non-linear transformations. A good representation can be defined as one that disentangles underlying factors of variation for input data. It turns out that deep learning approaches can find useful abstract representations of data across many domains (Ainsworth, 2006). Facebook is also planning on using deep learning approaches to understand its users. Deep learning has been so impactful in industry that MIT Technology Review named it as a top-10 breakthrough technology of 2013.

1.2 Problem Statement

Recently, AI has become one of the most important domain in computer science. Companies like Google, Facebook and Microsoft have also started to form their own research teams and making some impressive acquisitions. The goal of machine learning is to develop algorithms that can learn and recognize patterns or objects from complex data and make accurate predictions for previously unseen data (Lee, 2010). However, machine learning is not perfect yet and have some limitations.

First and foremost, the success of machine learning systems often requires a preprocessing of labeled data into a usable form before going through training phase. This allows the machine learning algorithm of choice to make sense of the incoming data. However, it is expensive to preprocess a large amount of data since it often requires significant human labour. Besides that, the performance of current machine learning algorithms depends heavily on the particular features of the data chosen as inputs. Furthermore, many real-world machine learning applications require a good feature representation to be successful. In contrast, deep learning always can perform well without having the need for preprocessing of input image.

Many existing machine learning algorithms using shallow architecture like support vector machines (SVM) which only have one hidden layer. Therefore, the internal representations learned by such shallow architecture are unable to extract some types of complex structure from input image because such system are simple (Bengio and LeCun, 2007). By contrast, deep learning architecture is able to extract these complex features and therefore object recognition by using deep learning with multi-layers of nonlinear processing are more efficient.

Lastly, deep learning method often require long training time as it consists of multi-layers with more than 1000 parameters in order to classify object with high degree of accuracy. Hence, difference approach like max-pooling is used to reduce the size of the feature maps in order to reduce the compute complexity and eventually reduce the training time. The high accuracy of classification is needed so that it can be used for application.

1.3 Objectives

First and foremost, the objective of this project is to train the multi-layer deep convolution neural network (DCNN) by using hierarchical features learning from labeled inputs without the need to preprocess the input image. Next, the purpose of the project is to classify an object with higher degree of accuracy by fine tuning the hyperparameters of the network. The last objective is to reduce the training time and compute complexity of the network by adding a subsampling layer after each convolution layer.

1.4 Scope

This research mainly focuses on how to train a DCNN system and then classify different objects into different classes correctly. In this work, each individual image inside the dataset contains only one object. Besides that, this study is limited to the software implementation using Matrix laboratory (MATLAB) and does not involved any hardware implementation. Next, the scope of this project is also limited to a still image. The segmentation and bounding box training are not covered in this research.

1.5 Contributions

The majority of this work shows how to implement a DCNN which is capable of extracting feature representations from a large amount of labeled data. Next, this work shows how neural network uses binary representation to classify an object into separate classes. Additionally, an attempt is made to optimize the hyperparameter of the DCNN to improve the performance. The DCNN model presented in this thesis is a very simple deep learning network which effectively extracts useful information for object classification. Adding average pooling to the network helps to simplify further on the calculation and reduces the training time. This proposed network structure can be a valuable baseline for the study of a more advanced deep learning architectures and be used for large-scale image classification tasks. Competitive results are also achieved on the Cifar-10 dataset. This constitutes an important generalization of deep learning to structured prediction and makes these models suitable for application.

1.6 Thesis Organization

This project report consists of six chapters. The first chapter reviews the introduction, problem statement, objectives, scope, and contribution of the project. The second chapter will discuss on related works. Chapter three will discuss the theories of neural network and some background on deep learning. Chapter four discusses the method and tool used in this project and how to implement a DCNN. Results and discussion will be discussed in chapter five and lastly chapter six includes the conclusion, future works and recommendations of this work.

REFRENCES

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3), 183-198.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8), 1798-1828.

Bengio, Y., and LeCun, Y. (2007). Scaling learning algorithms towards AI. *Large-scale kernel machines*, 34(5).

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*. 437-478. Springer Berlin Heidelberg.

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., and Ma, Y. (2014). PCANet: *A Simple Deep Learning Baseline for Image Classification?*,1404-3606.

Cho, K., Raiko, T., and Ihler, A. T. (2011). Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In *Proceedings of the 28th International Conference on Machine Learning*, :105-112.

de Jesus Rubio, J., Angelov, P., and Pacheco, J. (2011). Uniformly stable backpropagation algorithm to train a feedforward neural network. *Neural Networks, IEEE Transactions on*, 22(3), 356-366.

Deng, L., and Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*. 7(3–4), 197-387.

Epshtein, B., and Uliman, S. (2005). Feature hierarchies for object classification. In *Computer Vision, 2005. ICCV. Tenth IEEE International Conference on*. 1, 220-227.

Graves, A., Mohamed, A. R., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. 6645-6649.

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5), 5947.

Jha, S. N. (2010). *Nondestructive Evaluation of Food Quality*. (375). Heidelberg Springer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. :1097-1105.

Krizhevsky, A., and Hinton, G. (2010). Convolutional deep belief networks on cifar-10. *Unpublished manuscript*.

Kanan, C., and Cottrell, G. W. (2012). Color-to-grayscale: does the method matter in image recognition?. PloS one, 7(1), e29740.

Lee, H. (2010). *Unsupervised Feature Learning Via Sparse Hierarchical Representations*. Doctor Philosophy, Stanford University.

Maren, A. J., Harston, C. T., and Pap, R. M. (2014). *Handbook of neural computing applications*. Academic Press.

McCoppin, R., and Rizki, M. (2014). Deep learning for image classification. In *SPIE Defense+ Security*. 90790T-90790T. International Society for Optics and Photonics.

Mo, D. (2012). A Survey on deep learning: one small step toward AI. *Dept. Computer Science, Univ. of New Mexico, USA*.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. 1717-1724.

Schmidhuber, J. (2014). Deep Learning in neural networks: An overview. *Neural networks: the official journal of the International Neural Network Society*, 61-85.

Sivic, J. and Zisserman, A. (2003). A text retrieval approach to object matching in videos. In *Computer Vision, Proceedings, Ninth IEEE International Conference on*. 1470-1477.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11, 3371-3408.

White, R. L. (2000). Object classification as a data analysis tool. In *Astronomical Data Analysis Software and Systems IX*. 216, 577.

Zeiler, M. D., and Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks.,1301-3557.

Zhou, G. T., Lan, T., Yang, W., and Mori, G. (2013). Learning class-to-image distance with object matchings. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* 795-802.