Faculty of Electronics and Computer Engineering

PERFORMANCE EVALUATION ON QUANTIZED WEIGHT FOR CONVOLUTIONAL NEURAL NETWORK BASED OBJECT DETECTION

Mohd Hasbullah bin Putra

Master of Science in Electronic Engineering

2021

PERFORMANCE EVALUATION ON QUANTIZED WEIGHT FOR CONVOLUTIONAL NEURAL NETWORK BASED OBJECT DETECTION

MOHD HASBULLAH BIN PUTRA

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science
in Electronic Engineering

Faculty of Electronics and Computer Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2021

DECLARATION

I declare that this thesis entitled “Performance Evaluation on Quantized Weight for Convolutional Neural Network based Object Detection” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : MOHD HASBULLAH BIN PUTRA

Date : .....

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Electronic Engineering.

Signature :…….

Supervisor Name :PROFESSOR DR. ZULKALNAIN BIN MOHD YUSSOF

Date :……..

DEDICATION

To my beloved parents

“Life is like a bicycle, to keep the balance we need to keep going”

Abstract

A Convolutional Neural Network (CNN) based object detection is an emerging topic in the image processing field and has become the state-of-the-art in computer vision and machine learning. The traditional system in object detection uses a handcrafted feature extractor which is less robust in applications. By applying the CNN approach in the field, the accuracy of object detections can increase significantly. However, the use of deep CNN architecture model leads to high computation. In this research, a real-time CNN based object detection system is presented. The system is designed based on the modified You Only Look Once (modified-YOLO) architecture which is constructed with only 7 CNN layers. The grid cell parameter value of the system is varied to evaluate its effectiveness and ability in detecting small size objects upon deployment. The experimental results demonstrate that even with 7 convolutional layers, modified-YOLO can provide good detection accuracy and real-time operation achieving the best miss rate (*MR*) of 22.7% *MR*. Although the scores show an increase in the MR, the visual qualitative evaluation using randomly captured images indicate that the 7 layers modified-YOLO architecture with 11x11 grid cells can correctly and easily detect small objects. This makes the modified YOLO architecture which has been reduced in terms of complexity a suitable candidate for use in real-time operation. In order to further reduce the complexity of the CNN system, the trained floating-point weights are quantized. Three types of scalar quantization are used to quantize the CNN weights namely symmetric uniform quantizer, asymmetric uniform quantizer and non-uniform quantizer designed using *k*-means algorithm. The quantization reduces the storage and computation requirements. The quantitative results using the *MR* standard metric indicate that the non-uniform quantizer provides the best results compared to the other quantization methods. Using 6-bit precision non-uniformly quantized weights yields detection performance comparable to the CNN network using floating-point weights. Additionally, based on the qualitative results, the CNN network with 4-bit non-uniform quantization weights is able to detect the person objects correctly.

PENILAIAN PRESTASI PEMBERAT YANG DIKUANTUMKAN UNTUK PENGESAN OBJEK BERASASKAN RANGKAIAN NEURAL PELINGKARAN

ABSTRAK

Pengesan objek berasaskan Talian Neural Pelingkaran (CNN) adalah satu topik yang berkembang di dalam bidang pemprosesan imej dan telah menjadi kaedah terkini dalam bidang penglihatan komputer dan pembelajaran mesin. Kebanyakkan sistem tradisional dalam pengesanan objek menggunakan kaedah pengekstrak ciri yang tetap dan kurang berkesan ketika digunakan. Dengan menggunakan kaedah CNN, ketepatan pengesan objek telah meningkat dengan signifikan. Namun, seni bina CNN yang dalam telah menjurus kepada peningkatan pengiraan kompleks di dalam sistem. Dalam kajian ini, sebuah pengesan objek masa sebenar berasaskan CNN telahpun dibentangkan. Sistem tersebut direka berasaskan pengubahsuaian senibina "You Only Look Once" (modified-YOLO) yang terdiri daripada 7 lapisan CNN. Nilai parameter sel grid di dalam sistem tersebut juga telah dimanipulasi untuk menilai keberkesanan dan kemampuan sistem dalam mengesan objek yang kecil ketika digunakan. Keputusan kajian menunjukkan bahawa walaupun dengan hanya 7 lapisan CNN yang digunakan, "modified-YOLO" mampu memberi keputusan yang tepat dalam masa sebenar dengan pencapaian kadar sasar (MR) terbaik sebanyak 22.7%MR. Walaupun keputusan menunjukkan peningkatan MR, hasil penilaian kualitatif visual dengan menggunakan imej rawak yang ditangkap menunjukkan bahawa senibina YOLO dengan 7 lapisan CNN dengan parameter sel grid 11x11 mampu untuk mengesan objek kecil dengan tepat dengan mudah. Ini menjadikan pengubahsuaian senibina YOLO yang telah dikurangkan pengiraan kompleks sebagai calon yang sesuai untuk kegunaan masa sebenar. Untuk meneruskan pengurangan pengiraan kompleks dalam sistem, satu simulasi untuk menilai prestasi titik tetap CNN telah dijalankan dengan mengkuantumkan pengubahsuaian senibina YOLO. Tiga kaedah pengkuantuman skalar telah digunakan untuk mengkuatum pemberat CNN iaitu pengkuantuman seragam simetri, pengkuantuman seragam tidak simetri dan pengkuantuman tidak seragam yang menggunakan algoritma k-means. Pengkuantuman memberi kelebihan dengan mengurangkan pengiraan kompleks dan pada masa yang sama mengurangkan penggunaan memori. Keputusan pengkuantuman dianalisis secara kuantitatif menggunakan metrik standard MR menunjukkan pengkuantuman tidak seragam menghasilkan keputusan yang terbaik berbanding kaedah yang lain. Dengan hanya menggunakan kejituan 6-bit, kaedah ini menghasilkan pengesan yang setanding dengan titik apung CNN. Selain itu, berdasarkan keputusan kualitatif, rangkaian CNN yang sama mampu untuk mengesan objek dengan tepat.

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my sincere acknowledgement to my supervisor Professor Dr. Zulkalnain Bin Mohd Yussof from the Faculty of Electronics and Computer Engineering Universiti Teknikal Malaysia Melaka (UTeM) for his essential supervision, support, and encouragement towards the completion of this thesis.

I would also like to express my greatest gratitude to Dr. Sani Irwan bin Salim from Faculty of Electronics and Computer Engineering, co-supervisor of this project for his advice and suggestions in evaluation. Special thanks to Associate Professor Dr. Lim Kim Chuan, project leader for CREST grant for the financial support throughout this project.

Special thanks to my parents who consistently encourage me to finish this research work and become the main reason for me to keep on going. Thanks to all my friends, especially those who are in MLSP lab and Post Graduate lab for all the knowledge sharing and joy throughout the journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

c	-	Class
h	-	Height
w	-	Width
S	-	Number of Grid Cells
B	-	Number of Bounding Boxes

LIST OF PUBLICATIONS

1. Putra, M.H., Yussof, Z.M., Lim, K.C. and Salim, S.I., 2018. Convolutional neural network for person and car detection using yolo framework. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(1-7), pp.67-71.

2. Putra, M.H., Yussof, Z.M., Salim, S.I. and Lim, K.C., 2017. Convolutional Neural Network for Person Detection using YOLO Framework. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-13), pp.1-5.

CHAPTER 1

INTRODUCTION

1.1 Research background

Vision-based object detection is a hot research topic among the computer vision research community. In particular, the person and vehicle detection have a direct application in Advanced Driver Assistance System (ADAS), Intelligent Vehicle and Visual Surveillance System. Various methods have been proposed for person, car detection, or general object detection. However, majority of the techniques proposed to focus on achieving high detection accuracy at the expense of high computational complexity. Hence many of these methods are not suitable for real-time applications such as ADAS, which requires both high detection accuracy and high frame per second (FPS) performance.

Before the emergence of the Convolutional Neural Network (CNN), the Deformable Part Model (DPM) using handcrafted features such as Histogram of Oriented Gradient (HOG) has been the state-of-the-art object detector for many years. Inspired by the impressive performance demonstrated in image classification, CNN has been applied to object detection and achieves impressive results. Most notably, Girshick et. al. (2014) proposed the Regions with Convolutional Neural Network (R-CNN) framework for object detection and demonstrated state-of-the-art performance on standard detection benchmarks, for example in PASCAL VOC, with a large margin over the previous arts, which are mostly DPM based.

R-CNN uses a handcrafted selective search algorithm to generate object proposals and CNN classifier for detection tasks. The R-CNN is however computationally expensive

due to the forward pass computation required for each proposal. Girshick (2015) then proposed Fast Region-based Convolutional Network (Fast R-CNN), which reduces computational complexity by sharing convolutional features and pooling object proposals from the last convolutional layer. While Fast R-CNN achieves excellent detection accuracy, its speed is still limited by the bottleneck due to the object proposal generation. Then Girshick et. al. come out with series of proposal to overcome the intensive computation of CNN. Although the series of improvements introduced by the authors have achieved excellent object detection accuracy, it is computationally intensive and not suitable for use in real-time applications.

In order to meet the combined requirement of high object detection accuracy and real-time operation, a different approach of CNN-based object detection named You Only Look Once (YOLO) was proposed by Redmon et. al. (2016). In contrast to region proposal-based object detection algorithms such as Fast R-CNN, YOLO CNN-based algorithm predicts bounding boxes and class probabilities directly from full images in a single evaluation. Instead of having a consecutive pipeline of region proposals and object classification, YOLO implements a different approach by treating the overall input as a single regression problem. This reduces the latency at the expense of accuracy. The full 27 layers YOLO encounters a problem where the network is struggling to detect small sized objects.

However, all of the impressive performances are achieved by utilizing many layers and millions of parameters which leads to high computational and memory storage requirements. For example, the 8-layer AlexNet requires 244 MB of parameters and 1.4 billion floating point operations (GFLOP) to classify a single image, while the deeper CNN VGG-16 requires 552 MB of weight storage and 30.8 GFLOP per image Han et. al. (2015). This prohibits the deployment of deep CNNs on resource-limited embedded devices for example for mobile and portable devices especially involving real-time applications.

Embedded devices typically have constraints in terms of memory, energy and computation resource. Thus, in order to deploy deep neural networks to embedded platforms, some schemes should be applied to reduce the computation and storage requirements while preserving the deep CNN performance.

1.2 Problem statement

The early development of object detection methods used the selective search, which was deployed in most of the DPM based models and some of the early designed CNN models. The selective search method will slide through the whole image pixel by pixel. However, sliding through pixels requires lots of computational power and this process causes bottleneck towards the whole process of computation. Furthermore, the objects to be detected can appear at any location and scale within the input image. Figure 1.1 depicts the example of selective search in detecting the object. In addition, the object boundaries might be less clear compared with other objects that exist in the image. Therefore, all object scales have to be considered in selective search (Uijlings, 2012). Thus, the whole system speed performance will be slow and not suitable to be implemented in a real time detection system. In order to overcome this problem, the selective search method is replaced with another method called the unified detection which is introduced by YOLO and suitable for the real time application. In addition, YOLO network architecture was further modified to enable the model to be deployed in a low power platform.

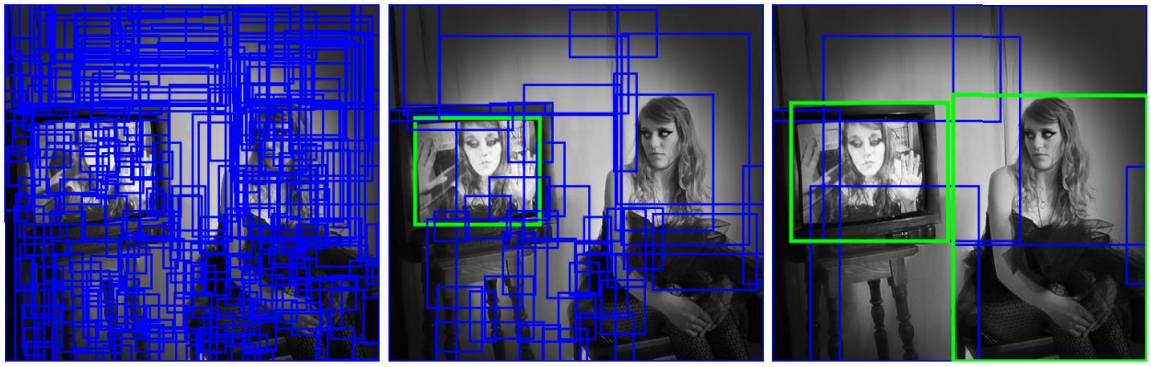

Figure 1.1: Selective search in detecting multiple sizes and location objects

On the other hand, the proposed CNN model YOLO struggles to detect small sized object which appears in the input image. Hypothetically, to remedy the problem, the grid cell parameter in the model is varied. This is to observe whether detection result towards small object can be improved. Theoretically, by increasing the grid cell parameter value should give advantages to detect small objects, but it will also increase the computation complexity of the system. Thus, in this research work, the optimum value of the grid cells parameter will be determined.

Apart from that, the existing CNN detection systems are computationally expensive and not suitable to be deployed on low powered platform such as FPGAs. Besides, the memory consumption of a CNN system is commonly very large. Therefore, in this research work, the quantization of the CNN weights is introduced, and the performance results of CNN using the quantized weights are compared to the floating-point CNN detection system. This is to ensure that the quantized CNN detection accuracy is still comparable after the quantization process.

1.3 Objectives

The objectives of this research work are as follows:

1. To design an object detection system using convolutional neural network.
2. To evaluate the performance of the designed object detection with other detection systems.
3. To compress the convolutional neural network weights using quantization schemes.
4. To evaluate the performance of the convolutional neural network using quantized weights and compare its performance with the floating point CNN object detection system.

1.4 Scopes

The scopes of this research work are as follows:

1. The input image for both training and testing are resized to 448 x 448 pixels.
2. The detection system detects persons for single object detection and combination of persons with car for multiple object detection.
3. The evaluation and datasets selection is based on ADAS standard.
4. The designed network architecture is trained to detect specific types of objects related to the driving assistance based on ADAS standard.
5. Only scalar quantization (SQ) will be used for weights quantization.

1.5 Research work contribution

In this research, the object detection system has been designed. The network architecture was inspired by Darknet and the network is compatible with the YOLO framework. Typically, a network for CNN will have a large number of layers, which is called Deep Neural Network (DNN). Having a greater number of layers will enhance the detection

capabilities of the CNN but will increase the computational power to run the system. Therefore, the proposed network architecture is designed to have less convolutional layers in order to ensure the network is capable to run in real-time. By reducing the number of layers, the expensive computations are reduced, and the FPS performance of the system can be boosted. Additionally, during the training phase of CNN, weights will require a lot of computation and can consume up to several days of training. To speed up the process, hardware accelerators like CUDA GPU is being utilized to further accelerate the training speed.

To enhance the accuracy of the detection, the grid cell parameter is varied in order to achieve optimum accuracy for the detection system. After the optimum parameter has been achieved, the yielded weights are then compressed using several quantization methods namely uniform quantization, asymmetric uniform quantization, and non-uniform quantization. The goal of weight compression is to reduce the network storage requirement while minimizing performance loss due to quantization.

Several network model compression schemes have been proposed in the literature review including network pruning, quantization, and low-rank matrix decomposition. The well-known fact partly motivates these previous works that the deep CNN network weights have significant redundancies as suggested by Cheng et. al. (2015). In our work, three different scalar quantization (SQ) techniques are used to quantize the network weights that are pre-trained with floating-point precision. No re-training is performed on the quantized network. The three SQ schemes applied are symmetric uniform quantizer, asymmetric uniform quantizer and non-uniform quantizer designed using the *k*-means clustering algorithm. All the weights in each layer of the YOLO networks are quantized. We then evaluate and compare the performance of the quantized networks using the three SQ schemes with a different number of quantization levels.

1.6 Thesis outlines

The rest of the thesis is organized as follows, Chapter 2 gives a brief review of the prior art research on object detection systems specifically on the convolutional neural network. Chapter 3 describes the proposed methods for research work. In the first part, the proposed network architecture is described alongside the fundamentals of CNN. Then the methods for quantization will be described in the last part of the chapter. Chapter 4 describes the result obtained from the research work. The results are being analyzed and discussed in this chapter. Lastly, Chapter 5 concludes the research work and explains the potential for future works of the research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the existing research on the fundamentals and the application of CNN. The first part of the chapter explains the basics of multi-layer perceptron which is the fundamental and commonly used in the Artificial Neural Network (ANN). Then the basic CNN model will be explained together with the evolution of the CNN design. Next, this chapter will also describe the commonly used network architecture in CNN. Then, the last part of the chapter will briefly describe the fundamental of quantization used in this research work.

2.2 Artificial Neural Network

Artificial Neural Network (ANN) is a system inspired by the biological neural network connection of the neuron cells in the brain. The brain consists of approximately 100 billion neurons, which each of the neurons is communicating with each other by sending electrochemical signals (Zhang, 2019). The neurons are connected through the junctions which are known as synapses. Each neuron is able to receive thousands of connections from other neurons, constantly receiving incoming signals until the signal reaches the cell body. If the accumulation of the signals surpasses a certain value of threshold, a response signal is sent through the axon. The example of biological neuron cells in the human brain is shown in Figure 2.1.