PERFORMANCE ANALYSIS ON IMAGE CLASSIFIER ON STM32 MICROCONTROLLER BOARD

MUHAMMAD AIMAN AKMAL BIN MOHD SHAIFULLIZAN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

PERFORMANCE ANALYSIS ON IMAGE CLASSIFIER ON STM32 MICROCONTROLLER BOARD

MUHAMMAD AIMAN AKMAL BIN MOHD SHAIFULLIZAN

This report is submitted in partial fulfilment of the requirements for the degree of Bachelor of Electronic Engineering with Honours

Faculty of Electronic and Computer Engineering
Universiti Teknikal Malaysia Melaka

2022

UNIVERSITI TEKNIKAL MALAYSIA MELAKA
FAKULTI KEJUTERAAN ELEKTRONIK DAN KEJURUTERAAN KOMPUTER

BORANG PENGESAHAN STATUS LAPORAN
PROJEK SARJANA MUDA II

Tajuk Projek : Performance Analysis on Image Classifier on STM32 Microcontroller Board

Sesi Pengajian : 2021/2022

Saya MUHAMMAD AIMAN AKMAL BIN MOHD SHAIFULLIZAN mengaku membenarkan laporan Projek Sarjana Muda ini disimpan di Perpustakaan dengan syarat-syarat kegunaan seperti berikut:

1. Laporan adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan laporan ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (✓):

☐ **SULIT***	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
☐ **TERHAD***	(Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan.)
☐ **TIDAK TERHAD**	

Disahkan oleh:

(TANDATANGAN PENULIS) (COP DAN TANDATANGAN PENYELIA)

Alamat Tetap: KM3329, Jalan Bayu 5, 78000 Alor Gajah Melaka

SANI IRWAN BIN MD SALIM
Pensyarah Kanan
Fakulti Kejuruteraan Elektronik & Kejuruteraan Komputer
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya
76100 Durian Tunggal
Melaka

Tarikh : 20 June 2022 Tarikh : 20 June 2022

*CATATAN: Jika laporan ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali tempoh laporan ini perlu dikelaskan sebagai SULIT atau TERHAD.

DECLARATION

I declare that this report entitled “Performance Analysis on Image Classifier on STM32 Microcontroller Board” is the result of my own work except for quotes as cited in the references.

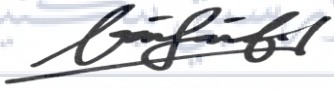
Signature : ……

Author : MUHAMMAD AIMAN AKMAL BIN MOHD SHAIFULLIZAN

Date : 20 June 2022

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 Bachelor of Electronic Engineering with Honours.

Signature : 

Supervisor Name : DR. SANI IRWAN BIN MD SALIM

Date : 20 June 2022

DEDICATION

I dedicate my dissertation work to my lecturer and many friends. A special feeling of gratitude to my loving supervisor, Dr Sani Salim whose words of encouragement and push for tenacity ring in my ears. I also dedicate this dissertation to my many friends and family who have supported me throughout the process. I will always appreciate all they have done for helping me to master about deep learning.

ABSTRACT

With the advancement in deep learning in the past few years, machine learning had improved in the one of develop complex models for image classification based on the characteristic of the image data. The use of deep learning usually been associated with big computers with fast CPUs and GPUs. Inevitably, it will consume a lot of electrical power and it very costly to build a high-performance machine learning workstation. To overcome this issue, deep learning framework is developed to be executed on a microcontroller board. This project is to analyse and evaluate the performance of image classifier implemented on the STM32 Microcontroller board. This project will use Teachable Machine online tool and STM32Cube.AI with FP-AI-VISION function pack to create an image classifier running on the STM32H747I-DISCO board. The deep learning model (image classification) are trained using Teachable Machine web application. The training model will be flashed to the STM32H747I-DISCO board for testing process. The STM32H747I-DISCO board performance while running deep learning system are analysed in term of the classification accuracy based on type of model had been make. The result for this project is successfully running image classification on STM32 microcontroller board with accuracy up to 100%, inference 577ms and 1.6 frame per second.

ABSTRAK

Pembelajaran mesin (Machine Learning) telah diinovasikan untuk menghasilkan sebuah model pengelasan imej yang lebih kompleks berdasarkan ciri-ciri yang terdapat pada objek tersebut dan ianya dipanngil Pembelajaran dalaman (Deep Learning). sistem ini biasanya dibangunkan menggunakan komputer canggih yang mempunyai CPU dan GPU untuk memberikan sistem pemprosesan yang pantas dan penyimpanan data yang besar. Tidak dapat dielakkan, ianya akan menggunakan tenaga elektrik yang banyak serta kos pembinaan yang tinggi. Bagi mengatasi isu ini, Pembelajaran Dalaman (Deep Learning) boleh diggunakan pada mikropengawal. Projek ini, adalah bertujuan untuk menganalisis prestatsi sistem pengelasan imej apabila diletakkan pada papan mikropengawal STM32. Teachable Machine iaitu aplikasi diatas talian akan digunakan untuk menghasilkan model pembelajaran dalaman (Deep Learning) bagi pengelasan imej bersama STM32Cube.AI dengan penambahan pek FP-AI-VISION dan model ini akan diletakkan pada STM32H747I-DISCO. Projek ini telah berjaya menjalankan klasifikasi imej pada papan mikropengawal STM32 dengan ketepatan sehingga 100%, inferens 577ms dan 1.6 bingkai sesaat.

ACKNOWLEDGEMENTS

First and foremost, Alhamdulillah, and thanks to Allah, the Almighty, for giving me to complete this Projek Sarjana Muda (PSM) successfully. Without the guidance of our Almighty, this project would not have been completed on time. This thankful is also gratitude to FKEKK from Universiti Teknikal Malaysia Melaka (UTeM) for the financial support. In this short period, there are a lot of obstacles and challenges that we must overcome. Throughout this project, I would like to express our utmost gratitude toward my project supervisor, DR. SANI IRWAN BIN MD SALIM for providing me with the title and all the other information regarding the title proposed for my Projek Sarjana Muda (PSM) and all their germinal ideas and guidance throughout this project completion. Without my supervisor, i would not fully grasp the concept of this Projek Sarjana Muda (PSM), which could not have hindered our progress. Besides, i would like to thank our family, colleagues, and lab assistance that always give a helping hand whenever necessary either directly or indirectly. Finally, I would like express my gratitude to everyone who has helped and supported me throughout Projek Sarjana Muda (PSM).

TABLE OF CONTENTS

Declaration

Approval

Dedication

Abstract **i**

Abstrak **ii**

Acknowledgements **iii**

Table of Contents **iv**

List of Figures **vii**

List of Tables **viii**

CHAPTER 1 INTRODUCTION **1**

1.1 Project Background 1

1.2 Problem Statement 3

1.3 Objectives 3

1.4 Scope of work 4

1.5 Thesis Outline 4

CHAPTER 2 BACKGROUND STUDY **6**

2.1 Introduction 6

2.2 Deep learning on microcontroller 7

2.3 Convolution Neural Network Deep Learning model in image classification 8

2.4 Modelling image classification using teachable machine 10

2.5 Transfer learning 11

2.6 TensorFlow Variant 12

2.7 Quantization of deep neural networks on microcontrollers 14

2.8 Summary of background study 14

CHAPTER 3 METHODOLOGY 16

3.1 Introduction 16

3.2 Flowchart 17

3.3 Detail Description of the Methodology Flowchart 19

3.4 Software Requirement 23

3.4.1 Teachable Machine 23

3.4.2 STM32Cube IDE and X-Cube-AI version 6.0.0 command line tool 24

3.4.3 FP-AI-VISION1 version 3.0.0 25

3.4.4 STM32CubeProgrammer 26

3.5 Hardware Implementation 27

3.5.1 STM32H747I-DISCO board 27

CHAPTER 4 RESULT AND DISCUSSION 28

4.1 Result analysis from Teachable Machine 29

4.1.1 Result of classification accuracy 29

4.1.2 Output model size based on types of models 33

4.2 Result analysis when converting TensorFlow Lite model to C-code 34

4.3 Result analysis when running deep learning model on STM32H747I-DISCO 37

4.3.1 Classification accuracy based on variable characteristic 38

4.3.1.1 Types of models 38

4.3.1.2 Distance of object from board camera 41

4.3.2 Comparison classification accuracy from Teachable Machine and STM32H747I-DISCO board 42

4.3.3 Delay of classification (inference) when different object changed 43

4.3.4 Frame per second of camera display 43

CHAPTER 5 44

5.1 Conclusion 44

5.2 Future Works 45

REFERENCES 46

LIST OF FIGURES

Figure 3.1 Detailed Process of Research Methodology 18

Figure 3.2 Distance of camera to the wall 22

Figure 3.3 Teachable Machine setup to train data 24

Figure 3.4 Adding X-cube-AI on STM32Cube IDE 25

Figure 3.5 Type of application folder inside FP-AI-VISION1 26

Figure 3.6 User Interface of STM32Cube Programmer 26

Figure 3.7 STM32H747I-DISCO board 27

Figure 4.1 Memory usage graph on STM32Cube IDE 34

Figure 4.2 Memory usage graph after activation buffer on STM32Cube IDE 35

Figure 4.3 Validation of the converting model 35

Figure 4.4 Output when using stm32-ai command line 36

LIST OF TABLES

Table 4.1 Result from Teachable Machine by different number of sample and epochs 30

Table 4.2 Under the hood Teachable Machine 30

Table 4.3 Accuracy result from Teachable Machine based on quality of image samples 31

Table 4.4 Result of accuracy Teachable Machine based on types of models 32

Table 4.5 Output model size based on types of models 33

Table 4.6 Result of different method of converting model 37

Table 4.7 Accuracy of classification for model with image sample 1080p 39

Table 4.8 Accuracy of classification for model with image sample 480p 39

Table 4.9 Accuracy of classification for model with natural shapes 40

Table 4.10 Accuracy of classification based on camera distance 41

Table 4.11 Comparison of accuracy between Teachable Machine and microcontroller board 42

CHAPTER 1

INTRODUCTION

1.1 Project Background

Machine learning can be defined as a data analytics technique that teaches computers to do what comes naturally to humans and animals such as how they learn from experience. Machine learning use computational methods to “learn” information directly from data without relying on a predetermined equation as a model so that the system can gradually improving output data accuracy. This output data can be used as a key technique for solving problems in computer and science areas, such as computational finance and biology, Image processing and computer vision, Automotive, aerospace, manufacturing and in Natural language processing. However, in line with technological development, the function of machine learning has been used to develop deep learning.

Deep learning is a type of machine learning that allows machines to learn from data that been stored. It involves the use of computer systems known as neural networks[1]. What can be conclude is deep learning are computers learning system that think using structures modeled same as the human brain so that deep learning can analyze images, videos, and unstructured data in ways machine learning can't easily do. Deep learning project can be built either by making code such as python and MATLAB or use deep learning application such as teachable machine and tinyML where is available online (webtools)[2] or offline (downloaded application)[3].

Electronic device with microprocessor such as Raspberry Pi[4] and Jetson Nano[5] is mainly used to running deep learning project like image classification or pose project. This is because this minicomputer got a fast CPUs and GPUs with big RAM size where is suitable to save data and running deep learning project using it. But the using of this minicomputer will consume a lot of electrical power and it very costly to build this high-performance machine learning workstation[6].

However, to overcome this issue today deep learning can be run using microcontroller where is cheaper than minicomputer[7]. Microcontrollers are typically small, low-powered computing devices that are embedded within hardware that requires basic computation[8].

Also, by implement deep learning in microcontroller, billions of microcontroller devices can be boosted to be more intelligent. With this improvement, household appliance and IoT devices can be use efficiently without relying on expensive hardware or reliable internet connections, which is often subject to bandwidth and power constraints and results in high latency. Normally for machine learning, user will have to string all raw data to the cloud which could contain confidential or private

information, with microcontroller it helps preserve privacy since no data leaves the device[9].

1.2 Problem Statement

The use of deep learning usually been associated with big computers with fast CPUs and GPUs. However, it will consume a lot of electrical power and it very costly to build a high-performance machine learning workstation. To reduce this issue, deep learning can be executed on microcontroller board. However, microcontroller got a less memory RAM value than minicomputer while the memory space needed to running deep learning application is bigger based on the deep learning application type. But by using Tensorflow Lite as a project model data compiler, it can reduce the file size by quantized the deep learning model. So that the memory space needed to run deep learning application can be reduced.

Because of this issue, analysis of the performance when running deep learning using microcontroller needs to be investigated if whether the deep learning performance using microcontroller are measured to determine is equal or better than device such as minicomputer.

1.3 Objectives

1. To design and develop image classification with Convolution Neural Network deep learning model using Teachable Machine to executed on STM32 microcontroller board.
2. To analyze the performance when running deep learning system on STM32 microcontroller board in terms of image classification accuracy.

1.4 Scope of work

1. Training variant of deep learning model based on some characteristics such as image shapes, values of epochs, number of image sample and resolution of image sample using teachable machine.
2. Convert the model from teachable machine into optimized C code for STM32 MCUs
3. Integrate the new model into the FP-AI-VISION1 to run live inference on an STM32 board with a camera
4. The software that will be used for this project is:
    a. Teachable Machine
    b. STM32Cube IDE
    c. X-Cube-AI version 6.0.0 command line tool
    d. FP-AI-VISION1 version 3.0.0
    e. STM32CubeProgrammer
5. The hardware that will be used for this project is:
    a. STM32H747I-DISCO Board
6. Analysis of performance:

    The analysis will be made when running image classification (deep learning) on STM32H747I-DISCO Board in term of system performance speed and accuracy of image classification based on variant type of model and the distance of camera to the object and also the frame per second of the board.

1.5 Thesis Outline

This report is divided into five chapters. The first section is an introduction, in which the project summary, problem statement, objectives, and scope of work are

explained. Chapter 2 includes information about the project that can be found in reference books, on the internet, in journals, or from other sources of information. Chapter 3 will discuss the robot motion limitation on the maximum and minimum angle of rotation for each degree of freedom. Chapter 4 will discuss the results and discussion in greater detail, while Chapter 5 will conclude this project and make some recommendations for future work.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

In this chapter, a review will be conducted to explore and gather more resourceful information and data that is relevant to this project. Multiple research paper, journals, and online resource such as E-books will be used to conduct a comprehensive study on multiple theories and background studies. Generally, the study will focus on the analysis running image classifier (deep learning) on microcontroller. Therefore, this required some background study on the evolution of deep learning in the microcontroller field. Through the procedure for running deep learning in microcontroller using teachable machine online application, detail information will be shared on how the data will be train into model than will be burn on microcontroller.

2.2 Deep learning on microcontroller

To run machine learning/deep learning model on microcontroller, the use of TensorFlow Lite is needed[10]. It is because, TensorFlow Lite is suitable for microcontroller or any other device with only few kilobytes of memory. This is because by using TensorFlow Lite it will compress model data in order to reduce edge device memory usage and computing time[11]. The core runtime just fits in 16 KB on an Arm Cortex M3 so that it can run many basic models. It doesn't require operating system support, any standard C or C++ libraries, or dynamic memory allocation.

According to [12], the journal discusses on implementation of deep learning concepts by using Arduino uno. In this journal, the writer modelling image classification based on a structure of a Convolutional Neural Network (CNN). The image classification system employs by CNN with 1800 trained data to categories two type of fruit image (orange and apple) successfully get accuracy up to 99.22% for image classified. However, the writer explain that the accuracy depends on the quality of the image it captures. If a poor-quality image is captured, then the accuracy is decreased resulting in a wrong classification.

Journal [13] shows that, the analysis of the implementation image classifier into TM4C123GH6PM by using OpenCV as the tool to train the SVM (Support Vector Machine). The implementation combines a GLCM (Gray Level Co-occurrence Matrix) and an SVM (Support Vector Machine), enabling the embedding of a computer vision application into a minimum resource platform. This research successfully gets accuracy 99.67% image classifier with 92% minimum use of RAM memory.

2.3 Convolution Neural Network Deep Learning model in image classification

Image classification domain can be improvised by using Convolutional neural networks (CNN). However, the large dataset is needed for training this Convolutional neural network model[14]. For image classification by using Convolutional neural network (CNN), image data will be the only input that will be learn based on itself characteristic and be classify into predefined output class [15]. This type of deep learning model is popular than other conventional algorithm is because it can apply filters to read 2d image so that the requirement for the pre-processing for this algorithm is very little[16].

In CNN, there will be three type of layer where is input layer, hidden layer and the output layer to help this algorithm to process and classify image data. However in the hidden layer it comprise other type of layer which is convolutional layers, ReLU layers, pooling layers, and fully connected layers, all of which play a crucial role[17].

In the convolution layer, the image will be determined either the image is black and white or color image. If the image is black and white, it will interpret as 2D layer with every pixel will be assigned a value between zero (0) and two hundred and fifty-five (255). Zero value means that the pixels are wholly black and two hundred and fifty-five if the pixel is completely white. However, if the image is color, it will become 3d array with a blue, green and red layer with each color value between 0 and 255[18].

After that, the reading of matrix will be started starting with smaller image called as filter or kernel. The depth between this filter and the input data depth is are never changed and this kernel will produce a convolution that move along the input image by 1 unit.

After that, this kernel matrix value will be multiplying with the original data (original image value). And all of this matrix will be combined together to generated single number. This process will be repeated along with other image (all input image) to obtained the final matrix where is smaller than original input image[19].

Final array in this CNN algorithm is called as the feature map of an activation map. The convolution of the image will produce other operation where is edge detection, blurring and sharpening of image. This operation can be done by applying different types of filters but the developer needs to define the specify aspect such as the size of filter or the number of kernel and the architecture of the network[20]. To increase the accuracy of the image, the non-linear layer (ReLU) will be activated. This is applied on the feature maps to increase the non-linearity of the image [21].

To make this network organize the object of the image, the image of the same object needs to be feed into the system. This is what be called as pooling layer where is to make the image become flexibility. This are referring to the image measurement where is height and width to reduced the size of the input image so the image can be identified[22].

At the end of the system, all layer will be connected together and artificial neural network will be added together to use this CNN model. The artificial network will combine the different features and functioning to improve the prediction of image with greater accuracy. Lastly the gradient of the error function is calculated concerning the neural network's weight.

2.4 Modelling image classification using teachable machine

Image classification can be defined as where machines is can look at an image and assign a (correct) label to it. Deep learning allows machines to identify and extract features from images. This means they can learn the features classification images by analyzing lots of pictures.

Teachable Machine (teachablemachine.withgoogle.com) is a web-based GUI tool for creating custom machine learning classification models without specialized technical expertise. Without coding and only use either webcam or image or sound, user can train their own machine learning model. Teachable Machine uses transfer learning to find patterns and trends within the images or sound samples and create a simple and easy classification model [23].

This is one of the platforms that used TensorFlow.js where is known as open source Javascript based library to developed machine learning or deep learning models. Teachable Machine are using transfer learning method to ensure the user can train their own machine learning or deep learning model with the dataset referring to the Google's Mobilenets architecture. Teachable Machine system are basically using convolutional neural network (CNN), where is known as a class of deep learning neural networks for analyzing visual imagery[24].

Teachable machine, offer three types of projects where is image project, audio project and pose project. In image classification project using teachable machine, need to upload image sample. The value of image sample with the variety of angle will determine the accuracy of the output model. After that, user can set up batch size and epoch before train the data. Batch size means a set of samples used in one iteration of training. For example: if the image is 80 and batch size is 16, that means data will split