Department of Electrical and Electronic Engineering

Project Specification Form

STUDENT'S NAME: Elim Kwan		**PROGRAMME CODE:** H606
PROJECT TITLE: Real Time Object Recognition with FPGA-based Convolutional Neural Network		**DATE:** 28TH Oct, 2019
SUPERVISOR'S NAME: Dr Jose Nunez-Yanez	**ASSESSOR'S NAME** Dr Naim Dahnoun	

1. AIMS AND OBJECTIVES

In recent years, object recognition is seeing widespread adoption in diverse applications, especially in the field of autonomous car and robotics. One of the most common architecture in realizing computer vision technology is the use of convolutional neural network (CNN). CNN is a class of deep learning neural networks, a machine learning technique for projects heavy with imagery. However, since CNN is very computation and memory intensive, real time object recognition applications are often limited by hardware processing power and resources. CPU is not an optimal platform for resources-intensive applications. While GPU may be a better solution since it works well on applications with massive parallelism, it still falls short in terms of energy efficiency and complexity of the system. Therefore, this project aims to examine the state-of-the-art FPGA-based Binarized Neural Network (BNN) accelerator system [1] and improve it to be a promising solution for high performance, low power real time classification systems.

The main features of the system under investigation are a) adapting BNN instead of CNN as binary computation is more hardware-friendly, b) using FPGA can help alleviate Von Neumann bottleneck experienced by both CPU and GPU. This project aims to improve the performance of this novel FPGA-based BNN accelerator system such that it can be adopted in day-to-day applications. The current system has a few limitations: the use of sliding window classification filter resulted in redundant computation; region of interest is predetermined at the centre of the image only; and power usage statistics is estimated from stimulation only.

Thus, to improve the current system, the main objectives of the project are:

a) To recreate the BNN accelerator system on ZedBoard and improve the existing classification filter to reduce computation overlaps.

b) Implement adaptive region of interest, which analysis the image based on the locations of objects in the frame.

c) Allow the system to adapt to the surrounding environment. Enter energy saving mode when system is in idle state.

d) Investigate the power usage of the system. Migrate it to another board with extensive power measurement capabilities to carry out more thorough energy efficiency analysis.

It is hoped that at the end, a more efficient real time object recognition system implemented with FPGA-based BNN accelerator could be demonstrated.

[1] Asad, M. M. M. (2019). Energy Proportional Object Recognition with Convolutional Neural Networks. Msc Thesis. University of Bristol. Provided by Dr. Jose Nunez-Yanez.

2. SUMMARY OF RELATED WORK AND KEY REFERENCES

Similar System and Background Theory (Ref: [1], [2], [3])

Neural network (NN) is gaining exponential momentum every day due to its plethora of applications. However, the computation and memory intensive nature of convolutional neural network (CNN) restrict its real-time application. Therefore, to alleviate the problem, state-of-the-art FPGA-based binarized neural network accelerator system has been proposed in the paper "Energy Proportional Object Recognition with Convolutional Neural Networks" [1] for real-time object recognition, which is the starting point of this project. This project aims to recreate and improve the system suggested in [1]. A similar system "BinaryEye" [2] has also been implemented for image recognition on the MNIST dataset.

Binarized Neural Networks (BNNs) model is resulted from quantizing weights and activation values in CNN. BNN is much more hardware-friendly and can be mapped to FPGA more efficiently without impacting accuracy by using different optimisation techniques. According to "FINN: A Framework for Fast, Scalable Binarized Neural Network Inference" [3], a paper co-authored by Xilinx research labs, using popcount for accumulation and using an unsigned comparison to compute output activation instead of calculating batch normalized values helps reduce calculation cost and boosts system's efficiency.

The overall classification system consists of an image sensor with a specified image acquisition speed and a BNN classifier. To achieve real-time classification, it is essential for the classification rate to be larger than or equal to the camera frame rate.

These findings [1], [2], [3] are significant to the project. Understanding the underlying mathematics in the FINN framework [1] – a major building block in FPGA-based accelerator allows us to have more control over the system and hence, easier to optimise it. [2] is the starting point of the project and we aim to construct and improve the system described in this paper. Discussion on BinaryEye [3] is insightful as well as it is very similar to the system we aim to recreate. The main differences are BinaryEye uses a more advance camera, which is a custom-made 1 mega-pixel high speed image sensor and it is mainly used for hand-written digit recognition based on MNIST dataset, while our system aims to recognise objects such as car, deer etc. However, the paper described how to accomplish real-time on-board classification system for edge detection with BNN Image classifier, which also adopts the FINN framework discussed in [3]. Hence, it is a useful guide to this project.

Implementation and Set up (Ref: [4],[5],[6])

ZedBoard can be set up following the instructions on the set-up guides [4][5]. Pretrained binarized neural network could be mapped to ZedBoard by connecting it to a MicroSD card preloaded with a PYNQ image [6].

The references are vital in the set-up process and in recreating the system mentioned in [1] as they are official technical specifications [4], [5] from the manufacturer of ZedBoard. The power and latency estimation described in [4] will also be useful in analysing system's performance. In addition, since porting PYNQ to other ZYNQ board is not well-documented in PYNQ official website, [6] gives a useful guideline for porting PYNQ-linux on Zedboards.

Pre and Post Processing of Data (Ref: [7], [8], [9], [10])

To enhance system accuracy, streaming data are processed before and after passing through the BNN.

Post-processing is conducted with a classification filter, which averages out the classification results over a certain time frame to improve stability and accuracy of the system's outputs. Sliding window methodology has been used [1], yet it causes large computation overlaps. Hence, the project aims to improve the filter by introducing the concept of tumbling window and slicing data [7]. Pros and cons of various window aggregation techniques and methods for slicing stream data in general application is proposed in [7]. By reducing the sliding steps of the window, there will be less redundant computation, and hence can improve overall efficiency.

In terms of pre-processing, the project aims to implement an adaptive region of interest (ROI). The concept of ROI has been previously implemented in [1]. However, the ROI was predetermined at the centre of the frame, which may result in inaccuracy if the object is not located there. Thus, this project aims to employ an adaptive ROI via separating foreground from the background of the images. Traditionally, image segmentation can be realized with gaussian blur, threshold and canny edge detection functions. Alternatively, M. A. Mousse suggested the use of cobebook algorithms to detect fast moving object from dynamic background [8]. To minimise calculation cost, the algorithm exploits the colour space specification of CIE L*a*b* and uses the improved simple linear iterative clustering (SLIC) algorithm to model special dependencies between pixels. D. Sangeetha [9] also suggested an efficient hardware implementation of Canny Edge Detection Algorithm by utilising pipelining techniques. These measures allow the system to focus the attention on the important parts of the image and hence improve classification accuracy.

Another pre-processing technique is to allow the system to adapt to its surrounding environment. The clock rate can be slowed down when the system is idle. According to the article "A real-time object detecting and tracking system for outdoor night surveillance" [10], reliable object detection can be achieved by thresholding the luminance contrast of images sequence. High level of similarities among consecutive frames is a usual indication of the system is in idle, and it can be slowed down to save power.

[7] is highly relevant to the project since it provides insights to the performance of different window aggregation techniques, which can be adopted in the project. [8], [9], [10] are also crucial to the project. They sum up different image segmentation and object detection algorithms, which can be adopted in pre

and pro processing of data, and help improve system performance.

Performance Analysis (Ref: [11], [12], [13])

Power usage and latency of the system can be estimated from ZedBoard datasheet [4]. With reference to the schematics [11], by measuring the voltage across pins 1 and 2 of J21 (current sensing pins) and dividing it by 10milliohms [11], input current can be obtained. Multiplying it with the input voltage statistic gives the overall power usage of the board. Comparison can then be drawn between the predicted values [1][4] and the measured values. However, more detailed analysis on power consumption of programable logic on the board is deemed impossible, since ZedBoard does not provide access to internal power rails. Therefore, at later stage of the project, it is hoped that the system can be migrated to a different board, where power usage of individual rails can be measured. These references [12] [13] described how to monitor power usage on Xilinx ZC702, a hybrid CPU + FPGA evaluation board that includes extensive capabilities for power measurement.

The findings [12][13] are significant to the project since both ZC702 and ZedBoard share the same Zynq device (7z020), with power measurement features, ZC702 can be an alternative to ZedBoard. [4] [11] give information about the schematics and performance benchmarks of ZedBoard, which are of paramount importance to the project.

[1] Asad, M. M. M. (2019). *Energy Proportional Object Recognition with Convolutional Neural Networks*. Msc Thesis. University of Bristol. Provided by Dr. Jose Nunez-Yanez.

[2] P. Jokic, S. Emery, and L. Benini, “BinaryEye: A 20 kfps Streaming Camera System on FPGA with Real-Time On-Device Image Recognition Using Binary Neural Networks,” *2018 IEEE 13th Int. Symp. Ind. Embed. Syst. SIES 2018 - Proc.*, pp. 1–7, 2018.

[3] Y. Umuroglu *et al.*, “FINN: A Framework for Fast, Scalable Binarized Neural Network Inference” Dec. 2016.

[4] AVNET, “ZedBoard (Zynq^{TM} Evaluation and Development) Hardware User’s Guide,” January, 2014.

[5] R. Further, “ZedBoard ä Getting Started Guide,” August 2012 [Revised January, 2014].

[6] P. parikh, pynq linux on zedboard, https://superuser.blog/pynq-linux-on-zedboard/, 2017.(Accessed on 27th October, 2019)

[7] J. Traub *et al.*, “Efficient window aggregation with general stream slicing,” *Adv. Database Technol. - EDBT*, vol. 2019-March, pp. 97–108, 2019.

[8] M. A. Mousse, C. Motamed, and E. C. Ezin, “An adaptive algorithm for fast moving object detection

from dynamic background based on cobebook," 4th RSI Int. Conf. Robot. Mechatronics, ICRoM 2016, pp. 584–588, 2017.

[9] D. Sangeetha and P. Deepa, "An Efficient Hardware Implementation of Canny Edge Detection Algorithm," Proc. IEEE Int. Conf. VLSI Des., vol. 2016-March, pp. 457–462, 2016.

[10] K. Huang, L. Wang, T. Tan, and S. Maybank, "A real-time object detecting and tracking system for outdoor night surveillance," Pattern Recognit., vol. 41, no. 1, pp. 432–444, 2008.

[11] Digilent, Inc. "ZedBoard" ZedBoard schematics, March 2013

[12] M. Geier, D. Faller, M. Brandle, and S. Chakraborty, "Cost-effective energy monitoring of a zynq-based real-Time system including dual gigabit ethernet," *Proc. - 27th IEEE Int. Symposium Field-Programmable Custom Computing Machines FCCM 2019*, p. 327, 2019.

[13] A. F. Beldachi and J. L. Nunez-Yanez, "Accurate power control and monitoring in ZYNQ boards," *Conference Digest - 24th International Conference on Field Programmable Logic and Applications, FPL 2014*, 2014.

3. RESOURCE REQUIREMENTS

In terms of hardware, Avnet Zedboard and Logitech C160 webcam will be used for implementing the FPGA-based BNN accelerator system. If time permits, the system may migrate to a more advance board, namely Xilinx ZC702. Other equipment required are:

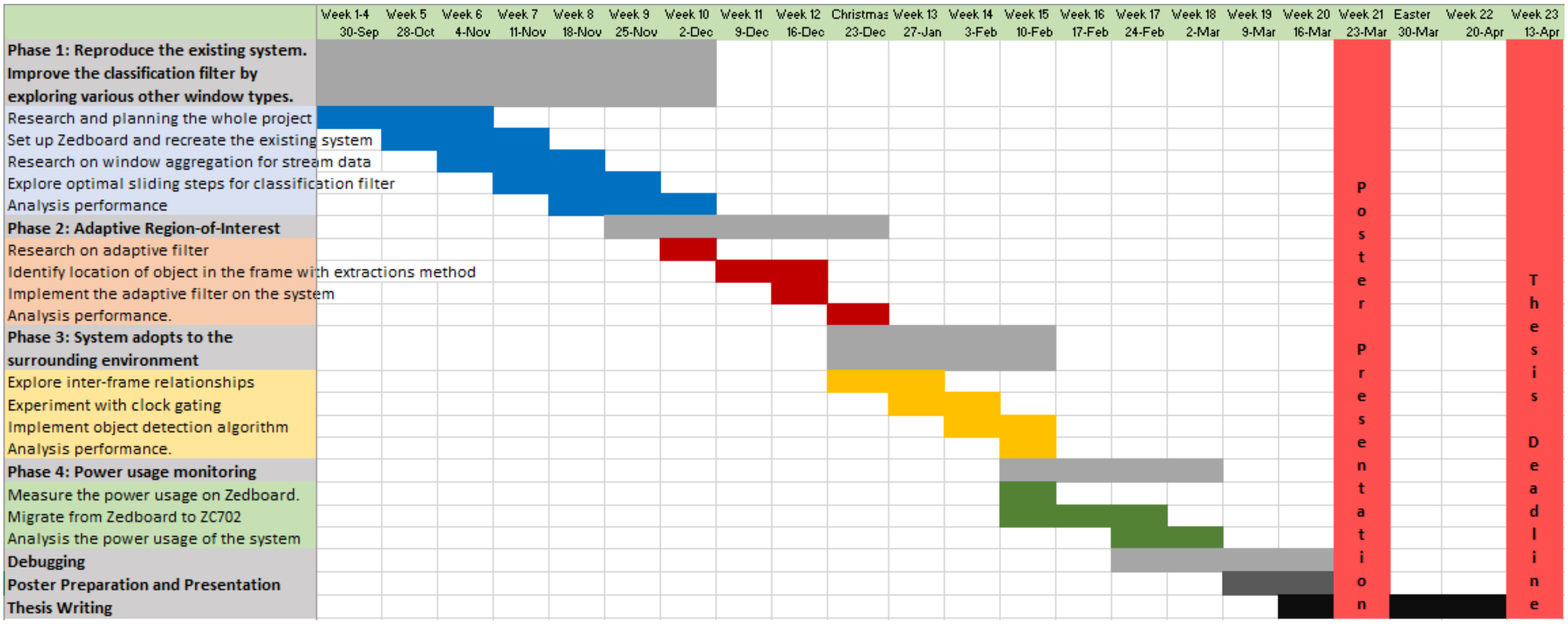
- SD Card for storing the trained neural network model on the board
- Laptop for communicating with the board using SSH protocol
- Standard Ethernet Cables

In terms of software, for basic set up, the following software will be used:

- WinSCP: to configure the IP address of the computer and the FPGA
- PuTTY: to communicate with the board using SSH protocol and allow file transfer
- Xming: to display the graphical applications of the FPGA on the computer screen.

Lastly, to test the functionality of the real-time object recognition system, in addition to testing it against image sets, figures of car, deer, etc will be placed in a constant lighting box as experiment set up. Therefore, the project also required a box, different figures and LED lights.

4. PROJECT WORK PLAN



	Week 1-4 30-Sep	Week 5 28-Oct	Week 6 4-Nov	Week 7 11-Nov	Week 8 18-Nov	Week 9 25-Nov	Week 10 2-Dec	Week 11 9-Dec	Week 12 16-Dec	Christmas 23-Dec	Week 13 27-Jan	Week 14 3-Feb	Week 15 10-Feb	Week 16 17-Feb	Week 17 24-Feb	Week 18 2-Mar	Week 19 9-Mar	Week 20 16-Mar	Week 21 23-Mar	Easter 30-Mar	Week 22 20-Apr	Week 23 13-Apr
Phase 1: Reproduce the existing system. Improve the classification filter by exploring various other window types.	■	■	■	■	■	■	■												Poster Presentation			Thesis Deadline
Research and planning the whole project	■	■	■																			
Set up Zedboard and recreate the existing system		■	■	■																		
Research on window aggregation for stream data			■	■	■																	
Explore optimal sliding steps for classification filter				■	■	■																
Analysis performance					■	■	■															
Phase 2: Adaptive Region-of-Interest						■	■	■	■	■												
Research on adaptive filter							■															
Identify location of object in the frame with extractions method								■	■													
Implement the adaptive filter on the system									■													
Analysis performance.										■												
Phase 3: System adopts to the surrounding environment										■	■	■	■									
Explore inter-frame relationships										■	■											
Experiment with clock gating											■	■										
Implement object detection algorithm												■	■									
Analysis performance.													■									
Phase 4: Power usage monitoring													■	■	■	■						
Measure the power usage on Zedboard.													■									
Migrate from Zedboard to ZC702													■	■	■							
Analysis the power usage of the system															■	■						
Debugging															■	■	■	■				
Poster Preparation and Presentation																	■	■				
Thesis Writing																		■		■	■	

The project will be divided into four phases:

1) Phase 1: Reproduce the BNN real time object recognition system on ZedBoard. Improve the efficiency of classification filter by adopting a different window aggregation technique. (e.g. a mix of sliding and tumbling window)
2) Phase 2: Implement adaptive region of interest, which analysis the image based on the location of object in the frame
3) Phase 3: Allow the system to adapt to the surrounding environment. Enter energy saving mode when system is in idle state.
4) Phase 4: Investigate the power usage of the system. Migrate it to another board with extensive power measurement capabilities to carry out more thorough energy efficiency analysis.

It is hoped that Phase1 and Phase 2 can be accomplished in TB1. Phase 3 can then be started during Christmas time. In TB2, more focus can be placed on Phase 4 and debugging any issue encountered.

5. PROJECT PROGRESS RISKS AND CRITICAL PATHS

The main progress risk may be time taken to set up unfamiliar hardware and software. Since the first phase of the project is to set up the FPGA and to update the classification filter, these concepts are relatively new, more time may be required. Therefore, more debugging and research time are planned for phase 1 of the project.

There are also progress risks regarding measuring the power usage of FPGA. To migrate the system to another board, more time should be allocated for ordering new board or borrowing it from respective professors. If it is to be ordered online, the shipping may cause delay to the schedule, hence components ordering should be conducted in advance. Furthermore, to measure power usage using the rails on the board will involve constructing own hardware circuitry, in which workings will be limited by laboratory opening hours and resources available. More time should be allocated for building and debugging.

In addition, since programming and computing is a major part of the project, the possibilities of losing files due to software failures exists. Therefore, constant back up is essential in preventing progress lost and version control software such as GitHub is proven to be useful in preserving previous work. Furthermore, some of the free software clearly stated that they do not have warranty. This imposes additional risk and hence should be used with caution.

Furthermore, hardware plays a main role in the project, especially for demonstrations and presentations purposes. The risk of the camera breaking down unexpectedly cannot be neglected. Also, it has a relatively low cost (less than GBP 10). Thus, it is worthwhile to purchase an identical webcam as a back up.

6. ETHICAL, LEGAL AND ECONOMIC ISSUES

In terms of ethical risk, violating copyright may be one of the potential risks. Since the project is based on previous work and we aim to refactor and enhance the code from related project, there will be copyright issues if citation is unclear.

In terms of legal aspect, the use of industrial software may impose legal risk. For instance, despite being free of charge, Xming is not licensed under General Public License. Therefore, one should carefully read through and respect the terms and conditions of various software applications.

In terms of economic aspect, the FPGA-based real-time object recognition system has great potential in the field of autonomous car and robotics. Furthermore, it can be used to reduce mundane work for humans, which help improve people's quality of life and brings social improvements. However, since the project is implemented with FPGA-based platform, which contains heavy metal and highly toxic industrial chemicals, e.g. polychlorinated biphenyls, it imposes environmental and health risk. Having said so, the system can save life in another domain. Real time object recognition has been proven useful in the scope of rescue robot or replacing humans in tasks deemed too dangerous, e.g. duelling with radioactive substances. This helps improve the general well-being of humanity and create a better place for the next generation, hence ultimately benefiting global sustainability.