Hardware for Machine Learning

Subtitle if any

MS/PhD Thesis Proposal

Submitted in partial fulfillment of the requirements for the award of the degree of

Master/DOCTOR OF PHILOSOPHY

IN

ELECTRICAL ENGINEERING

By

Hazoor Ahmad

PhDEE17004

Supervisor Name

Dr. Rehan Hafiz

Co-Supervisor Name

Dr. Mohsin Ali

Department of Electrical Engineering, Faculty of Engineering

Information Technology University

Lahore, Punjab, Pakistan

August 2019

Abstract

This section consists of the topic/research problem, theoretical approach, research methodology and significance of the study. ddasdad sggfgs fgfdsgsg. ddasdad sggfgs fgfdsgsg. ddasdad sggfgs fgfdsgsg.

1 Introduction

This brief paragraph states the objective and/or goals of the PhD research. The information provided in this area gives the reader a quick overview of what is included in the proposal. reseacha adaadad asdasdasd [Krizhevsky et al., 2012] fsdafh dsafas fasf safsa fasfas fas f. This research ais shown in Fig. 1. Our focus is design strategies.

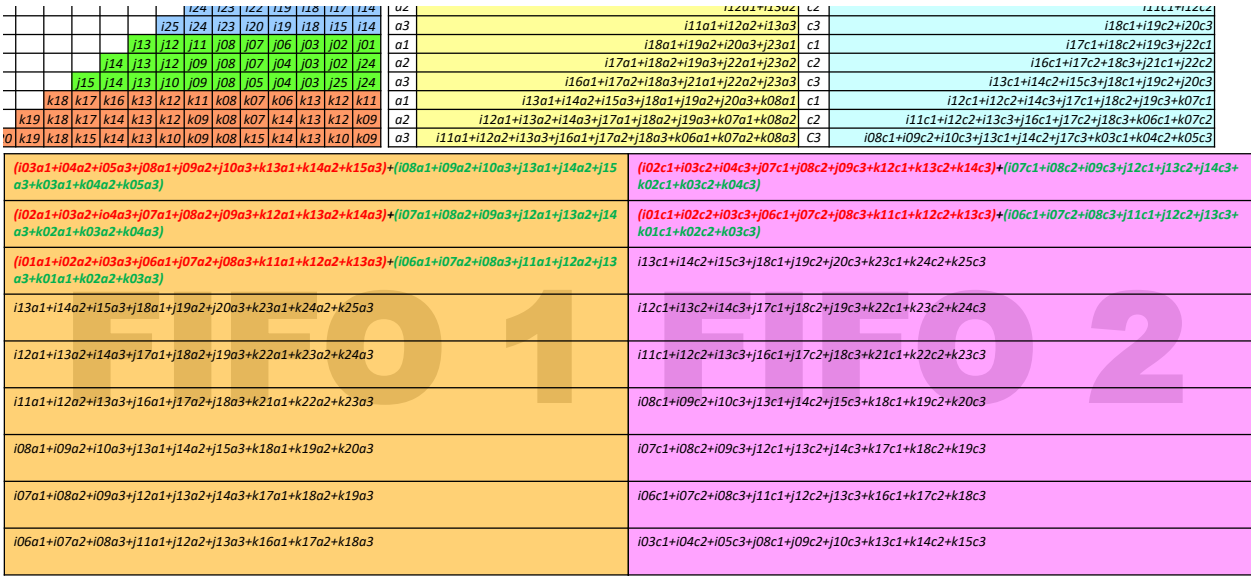


Figure 1: Overview of Hardware for Machine Learning

Organized as follows: section 2 contains a comprehensive literature survey on deep learning, section 3 limitations of state-of-the-art , section 4 includes problem statement, its scope, novelty and justification, section 5 explains significance and impact of proposed approach, section 6 provides a detailed design, modelling, analysis, simulation and evaluation of preliminary research results, section 7 enlists a clear set tasks or major milestones and evaluation strategies with their time-line in the form of Gantt chart.

2 Literature Survey

A detailed review of literature establishes your command in your area of research. This chapter should provide a complete and critical review of the state of the art work and not just summaries of books/articles.

This survey is focused on NeuFlow [Farabet et al., 2011]. All these implementations include Zhang [Zhang et al., 2015], Ayat [Ayat et al., 2018], and Caffeine [Zhang et al., 2018].

gsdfgsdfgg gfsgsdf fdsgdgfsfdgsdf . SysArrAccel [Wei et al., 2017] pointed out exploration. SmartShuttle [Li et al., 2018] provides configuration. Systimator [Hazoor et al., 2018] have tested networks .
Fused-Layer [Alwani et al., 2016] performs t optize fo a ngle FPGA.
Escher [Shen et al., 2017] utputs. .

3 Limitations of the State of the Art

In this section you need to clearly report the current Limitations of the State of the Art. The purpose is to clearly identify the limitations and weakness of the state of the art so that your chosen problem statement is justified.
State-of-the-art techniques lack [Wei et al., 2017,Li et al., 2018] or do not utilize [Shen et al., 2017,Alwani et al., 2016,Kwon et al., 2018]. In addition to above, 3D systolic array has not yet been fully deployed for DNNs [Huan et al., 2018].

4 Problem Statement & Proposed Research

In this section you need to report your identified Problem Statement. Furthermore, also provide a clear scope of your proposed research. The novelty of the problem being addressed and the particular approach being explored should also be commented upon/justified.
This research is focused on . We will . The sessgdfg fgsdfgdf fdgsdg gsdg . The sessgdfg fgsdfgdf fdgsdg gsdg . The sessgdfg fgsdfgdf fdgsdg gsdg .

5 Significance of the Study

This section is to provide the significance and impact of your proposed research. It should answer the following questions: Why you believe the study is significant? What implications your findings may have? Who will benefit from it? What will it contribute to the existing body of knowledge?
Systolic arrays are most useful .

6 Preliminary Research and Results

A written summary of some or all of the research performed is presented in a coherent manner. This section would include the approach taken and some preliminary results.

In this section we will demonstrate .

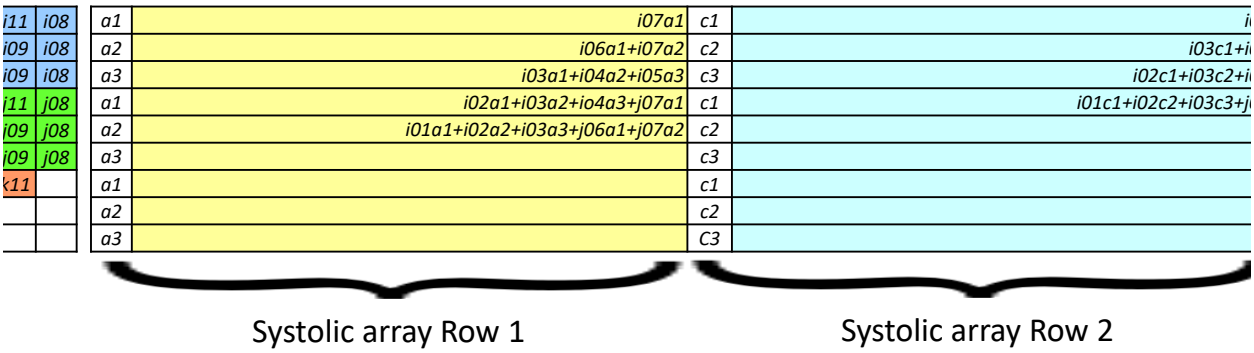


Figure 2: Dataflow engine.

6.1 E-Syzer

In this section we propose E-Syzer: An Efficient Systolic array sizer for DNN . Our proposed Fig. 2. We consider three different types of data traversals:

1. Input:
2. Weight:
3. Output:

systolic array in Fig. 2.

6.1.1 Transr Enges (WTE, ITE, OTE)

Weiht, and oput tansfr gines.

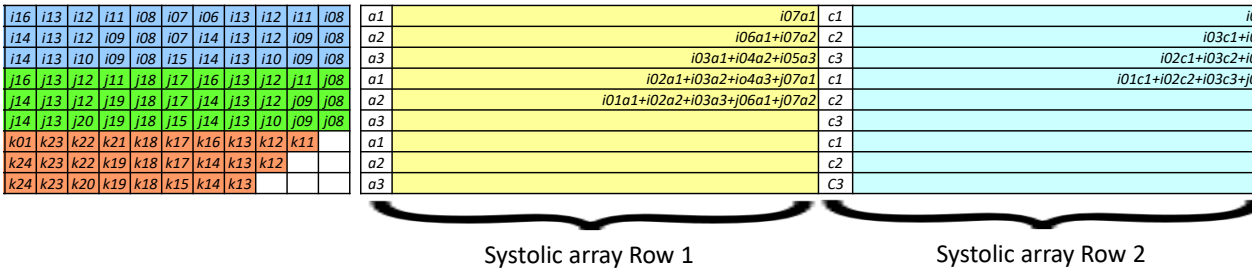
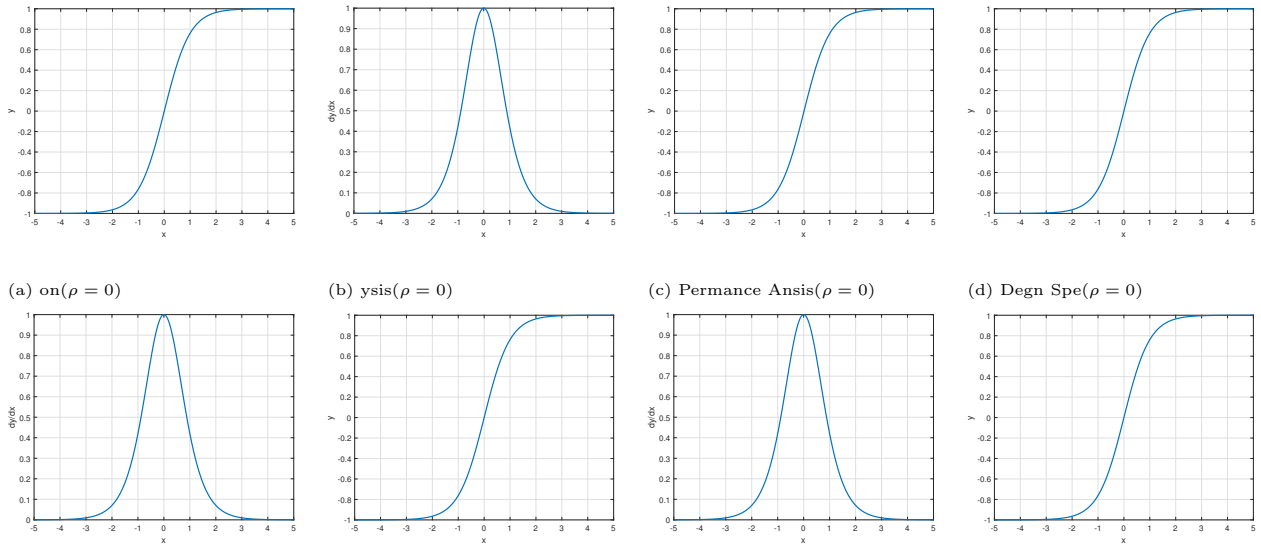


Figure 3: Systolic array



(a) on($\rho = 0$)

(b) ysis($\rho = 0$)

(c) Permance Ansis($\rho = 0$)

(d) Degn Spe($\rho = 0$)

(e) Layer wise resource estimation($\rho = 1$)

(f) Resource Analysis($\rho = 1$)

(g) Performance Analysis($\rho = 1$)

(h) Design Space($\rho = 1$)

Figure 4: Layer rders.

6.1.2 Systlic Aay (SA)

Onevey clok cyle in Fig. 2.

Table 1: Syzer Parameters

CNN Parameters			
Parameter Name	**Symbol**	**Tiling**	**indexing**
output	M	T_m	m
input	N	T_m	n
rows	R	T_m	r
columns	C	T_m	c
FPGA/Hardware Resources			
No. of available DSP blocks	$\mathbb{D}$		
No. of available slice LUTs	$\mathbb{G}$		

Before we move on, $\mathcal{O}$

$$T_m = min(\lceil \frac{T_m}{T_m} \rceil, R) \tag{1}$$

For simplification we take all columns of a layer to be the tiling factor for ofm columns.

$$T_c = min(\lceil \frac{C_{TH}}{\mathcal{O}} \rceil, C) \tag{2}$$

To minimize optimization problem

$$\underset{<\mathcal{A},\mathcal{B}>}{\text{minimize}} \qquad \sum_{i=1}^{\mathcal{C}} \mathfrak{D}_i$$

subject to

$$T_m \leq R$$

$$T_m \leq C$$

$$\mathfrak{d} \leq \mathbb{D}$$

6.1.3 Evaluation

7 Proposed Research Plan

This objective of this section is to provide a clear set of tasks and intended approaches that shall be executed and evaluated to complete the PhD research work. All major milestones should be clearly mentioned. Evaluation plan can also be provided as to how the results shall be evaluated. A Gantt chart should be provided highlighting the outline of the work needed to complete the PhD research, and the time required for completion

There are following milestones to study

1. First Goal
2. Second Goal
3. Third Goal
4. Froth GOal

MATLAB® will be usd fr al simultions. Keras®, TensorFlow® or Caffe® fr taining. Harware implemntatn will be ing Vivado Desgn Sute - HLx Ediions by Xilinx®.
Table 2 provides a summary of major tasks along with their expected time of completion.

Time / Activity	F 18	S 19	F 19	S 19	F 20	S 20
Literature Review and Study	✓					
Task 1		✓				
Task 2			✓			
Task 3				✓		
Task 4					✓	
Final Write-up & Thesis Submission						✓

Table 2: List of tasks

Other Considerations

References should be presented in a consistent manner throughout the proposal. Use APA Referencing style for formatting the bibliographic material. It is important that figures, tables, and references in the proposal are presented in a manner consistent with professional publication standards. When placing a figure or table and its identifying description in the proposal, it is important to consider ease of access for the document reader. In most cases the text which introduces a figure or table will precede the placement of the figure or table in the proposal

References

[Alwani et al., 2016] Alwani, M., Chen, H., Ferdman, M., and Milder, P. (2016). Fused-layer cnn accelerators. In *The 49th Annual IEEE/ACM International Symposium on Microarchitecture*, page 22. IEEE Press.

[Ayat et al., 2018] Ayat, S. O., Khalil-Hani, M., and Ab Rahman, A. A.-H. (2018). Optimizing fpga-based cnn accelerator for energy efficiency with an extended roofline model. *Turkish Journal of Electrical Engineering and Computer Science*, 26(2):919–935.

[Farabet et al., 2011] Farabet, C., Martini, B., Corda, B., Akselrod, P., Culurciello, E., and LeCun, Y. (2011). Neuflow: A runtime reconfigurable dataflow processor for vision. In *CVPR Workshops*, pages 109–116.

[Hazoor et al., 2018] Hazoor, A., Tanvir, M., Abdullah, M., Javed, M. U., Hafiz, R., and Shafique, M. (2018). Systimator: A design space exploration methodology for systolic array based cnns acceleration on the fpga-based edge nodes. *arXiv preprint arXiv:1901.04986*.

[Huan et al., 2018] Huan, Y., Xu, J., Zheng, L., Tenhunen, H., and Zou, Z. (2018). A 3d tiled low power accelerator for convolutional neural network. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[Kwon et al., 2018] Kwon, H., Samajdar, A., and Krishna, T. (2018). Maeri: Enabling flexible dataflow mapping over dnn accelerators via reconfigurable interconnects. *ACM SIGPLAN Notices*, 53(2):461–475.

[Li et al., 2018] Li, J., Yan, G., Lu, W., Jiang, S., Gong, S., Wu, J., and Li, X. (2018). Smartshuttle: Optimizing off-chip memory accesses for deep learning accelerators. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 343–348. IEEE.

[Shen et al., 2017] Shen, Y., Ferdman, M., and Milder, P. (2017). Escher: A cnn accelerator with flexible buffering to minimize off-chip transfer. In *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pages 93–100. IEEE.

[Wei et al., 2017] Wei, X., Yu, C. H., Zhang, P., Chen, Y., Wang, Y., Hu, H., Liang, Y., and Cong, J. (2017). Automated systolic array architecture synthesis for high throughput cnn inference on fpgas. In *Proceedings of the 54th Annual Design Automation Conference 2017*, page 29. ACM.

[Zhang et al., 2015] Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., and Cong, J. (2015). Optimizing fpga-based accelerator design for deep convolutional neural networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 161–170. ACM.

[Zhang et al., 2018] Zhang, C., Sun, G., Fang, Z., Zhou, P., Pan, P., and Cong, J. (2018). Caffeine: Towards uniformed representation and acceleration for deep convolutional neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.