Computing at the edge

Deliverable 1: Final year Dissertation

Bsc Computer Science: Artificial Intelligence

Sam Fay-Hunt — sf52@hw.ac.uk

Supervisor: Rob Stewart — R.Stewart@hw.ac.uk

October 13, 2020

DECLARATION

I, Sam Fay-Hunt confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: ....

Date: ....

Abstract: a short description of the project and the main work to be carried out – probably between one and two hundred words

Contents

1 Introduction

Summarising the context of the dissertation project, stating the aim and objectives of the project, identifying the problems to be solved to achieve the objectives, and sketching the organisation of the dissertation.

Edge devices have never been cheaper *citation*, stuff about how IoT devices are ubiquitous

Mention how there is an increasing trend to perform computing at the Edge - real time applications + privacy

These devices are often equipped with some form of AI application: Photo enhancment ect.

Online vs offline learning

Edge-side inference

These models can have a huge number of parameters so inference can sometimes be impractical. [1] - “see Table 1”

Issues with limited resource computation [2]

outline the document: We start with ..., then we cover x, y, and z ...

This dissertation is an investigation into the effect of pruning on inference in terms of latency and accuracy using hardware without specific optimisations for processing the resulting sparse matrices from pruning. (Reasoning for statement...).

1.1 Background

Discussing related work found in the technical literature and its relevance to your project.

This Section will be split into 4 subsections:

Hardware Memory architectures Section 1.1.1 - brief stuff about this section

Edge Computing Section 1.1.2 - stuff about edge comp

Deep Learning Section 1.1.3 - stuff

Compression Types Section 1.1.4 - ...

1.1.1 Hardware memory architectures

- Discuss VPU/TPU/APU/GPU/FPGA/ASIC memory arcitecture and how it handles matrix sparsity

- Show ineffectivity of pruning on hardware without optimisations for sparse matrices

The explosion of Deep Neural Network applications in recent years has prompted the production of a wave of specialised hardware architectures to improve the efficiency and compute of these kinds of workloads. The mainstay of this form of processing has been until recently been dominated by GPUs.

1.1.2 Computing at the edge

Some background on edge computing - maybe a detailed definition

- Challenges of resource bound deep learning

- Online vs offline learning

1.1.3 Neural networks & Deep learning

Types of deep learning & inference

Deep Neural Network (DNN)

- Layer structure (Input, Hidden, output) - Weight parameters updated using back-propagation

- Feed Forwards

- Feedback Nerual Network

- Self-organizing Neural Network

Convolutional Neural Network (CNN)

- A class of DNN

- CNN consist of: Convolutional Layers, Pooling layers & fully connected layers.

- Convolutional Layers contain sets of filters/kernels

Recurrent Neural Network (RNN)

Neural networks are a subfield within the category of Artifical Intelligence (AI) computing. Neural networkss are composed of layers of neurons that pass signals derived from weights through the network, this model of computing was inspired by our understanding of the human brain, see Fig. 1 for a simple example, the weights can be seen corresponding to the synapes and the output of the neruon is labeled as the axon. All neruons in a Neural network have wieghts corresponding to their inputs, these weights are are intended to mirror the behavour of our synapses value scaling effect by performing a weighted sum operation. The neuron then applies an non-linear activation function to the result, without which a Neural network would just be a linear algebra operation [2].

There are many popular deep learning network architectures, this document will focus primarily on the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures

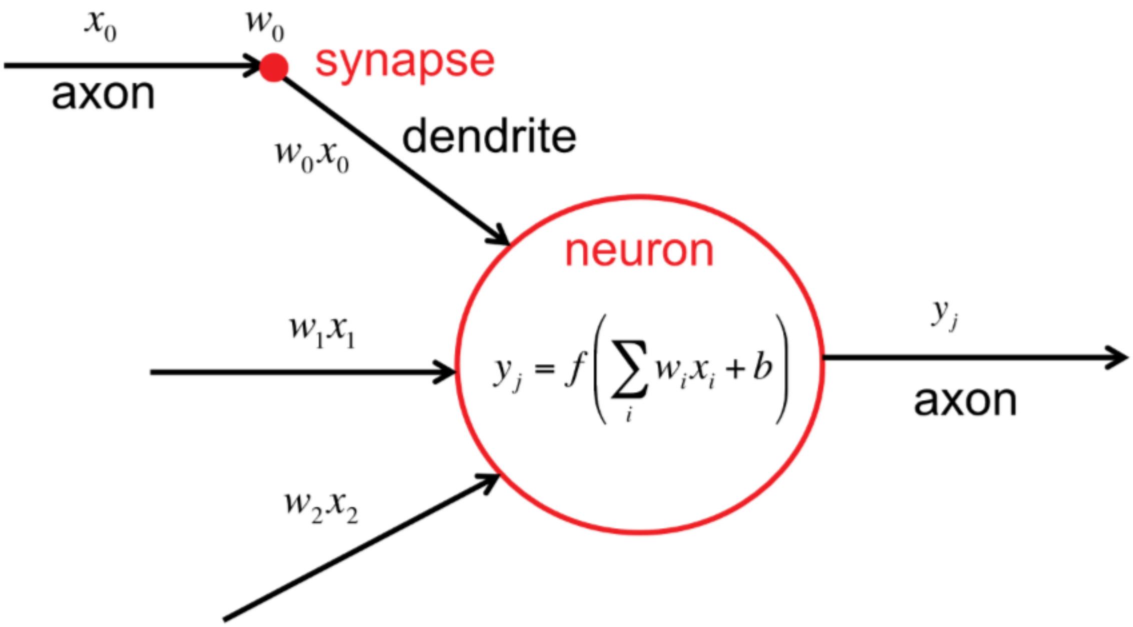

Figure 1: Neuron with corresponding biologically inspired labels.
(Adopted figure from [2])

1.1.4 Compression types

pruning

distillation

Quantization

Network design strategies

low-rank factorization

2 Research Methodology/ Requirements Analysis

2.1 Research Methodology

This is required for research projects and should be linked back to the project aim and objectives. It should describe the research methods that will be employed in the project and the research questions that will be investigated.

Find baselines/benchmarks

How to perform pruneing

systematic benchmark framework

Look at underlaying storage mechanism of parameters in Network

perform some engineering of refactoring/altering these mechanisms

rerun systematic benchmarking framework

2.2 Requirements Analysis

This is required for technical projects and should be linked back to the project aim and objectives. It should provide a detailed use case scenario and suitable use

3 Design

Initial software design/sketch of research Methodology

4 Evaluation Strategy

Details of the evaluation and analysis to be conducted

5 Project Management

5.1 Timetable

5.2 Risk Analysis

mention benchmarking NLP/NLG/Audio - text/text - audio models as a risk to the project

5.3 Professional, Legal & Ethical issues

A Back matter

A.1 References

References

[1] Y. Chen, B. Zheng, Z. Zhang, Q. Wang, C. Shen, and Q. Zhang, “Deep Learning on Mobile and Embedded Devices: State-of-the-art, Challenges, and Future Directions,” *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–37, Sep. 26, 2020, ISSN: 0360-0300, 1557-7341. DOI: `10.1145/3398209`. [Online]. Available: `https://dl.acm.org/doi/10.1145/3398209` (visited on 10/01/2020).

[2] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017, ISSN: 0018-9219, 1558-2256. DOI: `10.1109/JPROC.2017.2761740`. [Online]. Available: `http://ieeexplore.ieee.org/document/8114708/` (visited on 10/01/2020).

A.2 Appendices

to include additional material, consult with your supervisor.

Acronyms

AI Artifical Intelligence. 3

CNN Convolutional Neural Network. 3

RNN Recurrent Neural Network. 3