Inference at the edge: tuning compression parameters for performance

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

March 9, 2021

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: .....Sam Fay-Hunt.....

Date: .....10/12/2020.....

Abstract: *Abstract here*

Contents

1 Introduction

2 Background

- *Adapt from D1*
- *rewrite with more of a focus on the concrete channel and pruning methodology used*
- *Would be good to include wandb bayse hyperparam optimisation details*

3 Methodology

3.1 Conceptual Process

- *Sensitivity analysis - filter/channel selection*
- *Filter pruning implementation - Theory*
- *Channel pruning implementation - Theory*

3.2 Engineering steps

- *High level overview of physical system - justify need for multiple training agents*
- *Benchmarking setup - openvino + benchmark (getting latency/throughput)*
- *Pruning & retraining setup - Distiller (Pruning & training)*
- *Data processing - wandb + data visualisation steps*

4 Experiment Discussion

4.1 Filter and channel selection

Link back to selected model

- *Filter selection (visual representation of filters)*
- *Channel selection (visual representation of channels)*

5 Conclusion

5.1 Further work

- *Suggested improvements for methodology*
- *Next steps*

5.2 Discussion

- *Discuss results*
- *Criticism of methodology*

A Back matter

A.1 References