NTNU
Norwegian University of Science and Technology (NTNU)
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

Master's Thesis

Candidate:	**Bukueva Elena**
Course:	**TTK4900 Master's Thesis**
Title (English):	**Compiler-Aware Model Optimization for Edge AI Accelerators**
Title (Norwegian):	**Kompilatorbevisst modelloptimalisering for Edge AI-akseleratorer**
Thesis description:	
The tasks will be:	1. Bla 2. Bla
Start date:	15. January, 2026
Due date:	08. June, 2026
Thesis performed at:	Department of Engineering Cybernetics
Supervisor:	Professor Geir Mathisen, Dept. of Eng. Cybernetics

Abstract

Sammendrag

Contents

Abbreviations

1 Introduction

1.1 Motivation

1.2 Limitations

1.3 Contributions

Within these limitations, the main contributions of this thesis are:

- :
- :
- :
- :
- :

1.4 Thesis Structure

The remainder of this report is organised as follows:

- **Chapter 2**
- **Chapter 3**
- **Chapter 4**
- **Chapter 5**
- **Chapter 6**
- **Chapter 7**
- **Chapter 8**
- **Chapter 9**

- **Chapter 10**
- **Appendix A**
- **Appendix B**

2 Literature Study

2.1 Introduction to Object Detection with Deep Learning Models

Artificial Intelligence (AI) is a broad term covering a set of algorithms and operations on matrices and vectors that enable computers to perform a big variety of tasks in language understanding and translation, object recognition, robotic performance etc. We will consider only models performing object detection tasks, mostly having CNN [1] and Transformer [2] architectures.

2.2 Introduction to Deep Learning Model Optimization

2.3 Hailo Compiler

3 Theory

4 Methodology

5 Proposed Solution / Design

5.1 Dataset and Annotation

6 Testing and Results

6.1 Experimental Setup

6.1.1 Performance overview

6.1.2 Prediction Examples

6.1.3 Hailo impact

7 Discussion

7.1 Discussion Points

7.1.1 Quantization and accelerator deployment

7.1.2 Unexplored hardware configurations and deployment possibilities

8 Conclusion

9 Future Work

10 Description of Use of Artificial Intelligence

In working on this specialization project, I have used artificial intelligence (AI) tools in the following limited ways. I used ChatGPT (OpenAI) to assist with idea development, clarifying concepts, and improving the structure and language of the text (for example, by suggesting alternative formulations, checking grammar, and helping to simplify explanations). AI was also used to get help with technical formatting issues in LaTeX and to debug small code examples by explaining error messages and proposing corrections.

All scientific content, including the problem formulation, choice of methods, implementation, experiments, analysis, and conclusions, has been developed, reviewed, and academically assessed by me. AI tools were not used to generate entire sections of the report, to perform literature searches independently, or to answer the assignment in its entirety. I have not uploaded sensitive or confidential information to the AI tools. I am solely responsible for the academic content of this assignment.

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

A Demonstration (Appendix A)