**Aleksander Lempinen**

# MLOps approach for application specific performance tuning for machine learning systems

Master's Thesis in Information Technology

January 10, 2023

University of Jyväskylä

Faculty of Information Technology

**Author:** Aleksander Lempinen

**Contact information:** aleksander.lempinen@gmail.com

**Supervisor:** TODO supervisor

**Title:** MLOps approach for application specific performance tuning for machine learning systems

**Työn nimi:** TODO samma på finska

**Project:** Master's Thesis

**Study line:** Educational Technology

**Page count:** 12+0

**Abstract:** TODO abstract

**Keywords:** LaTeX, gradu3, Master's Theses, Bachelor's Theses, user's guide

**Suomenkielinen tiivistelmä:** TODO tiivistelmä suomeksi

**Avainsanat:** MLOPS, TODO

# Glossary

| | |
|---|---|
| ML | Machine Learning |
| MLOps | Machine Learning Operations |
| TODO | TODO |

# Contents

# 1 Introduction

Machine learning (ML) systems are widely adopted and many organizations successfully have ML models running in production.

- Problem with MLOps/ML tools
  - traditional ML performance metrics such as accuracy
  - fancy features such as neural architecture search, AutoML, performance tuning
  - fancy techniques such as early stopping, grid search, bayesian optimization search etc.
  - little support for non-ML metrics: CPU util, memory used, latency, throughput (images/s etc.), hardware required (CPU, GPU, TPU etc.)
  - ML in production has many objectives besides accuracy for example: satellite image processing model A took 4-5h and model B took 5min. Model A is infeasible in production despite being more accurate.
  - "Better' depends on the specific application
- Hypothesis: Early stopping will speed up computing non-ML metrics (CPU util, Memory use, GPU/TPU requirement, latency, throughput)

TODO smaller and smaller devices, limited resources

Real-world ML systems in addition to ML performance metrics will have similar performance metrics as traditional software systems. The aim of the study is to tune performance metrics particularly relevant to real-world ML systems.

The main contribution of this master's thesis is using ML performance tuning techniques such as early stopping for tuning a wider range of real-world ML system performance metrics.

# 2 Machine Learning Operations

TODO intro

This chapter introduces the basic concepts of ML in section 2.1 and basic concepts of DevOps in section 2.2. Finally in section 2.3 DevOps and ML are combined to form a new concept of MLOps.

## 2.1 Introduction Machine Learning

TODO what is ML

TODO training vs serving

## 2.2 DevOps

TODO resource allocation/resource consumption, small memory software, benchmarking

## 2.3 MLOps

TODO what DevOps brings to ML

# 3 Methods

The scope of the study is limited to 5 performance metrics of 3 different ML models trained and tested on 3 different datasets.

This master's thesis asks the following research questions:

- *RQ1*: How can we measure the performance of real-world ML systems?
- *RQ2*: How can these performance metrics be effectively tuned?

TODO This is a methods chapter

TODO no tool comparison

TODO different models, different datasets

TODO different resources (memory, time, accuracy)

# 4 Results

TODO This is a results chapter

# 5 Discussion

## 5.1 Research Questions revisited

## 5.2 Limitations and Validity

## 5.3 Related Work

## 5.4 Future Work

TODO This is a discussion chapter

# 6 Related Work

To find relevant related work both reverse snowballing and forward snowballing is used on a set of MLOps papers previously known to the author.

Benchmarking ML systems (Cardoso Silva et al. December 2020)

# 7 Conclusions

TODO This is a conclusions chapter

# Bibliography

Cardoso Silva, Lucas, Fernando Rezende Zagatti, Bruno Silva Sette, Lucas Nildaimon dos Santos Silva, Daniel Lucrédio, Diego Furtado Silva, and Helena de Medeiros Caseli. December 2020. “Benchmarking Machine Learning Solutions in Production”. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA),* 626–633. 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/ICMLA51294.2020.00104.