Aleksander Lempinen

MLOps approach for application specific performance tuning for machine learning systems

Master's Thesis in Information Technology

December 19, 2022

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: 10+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

- 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)

2 MLOps

TODO This is an MLOps chapter

3 Methods

TODO This is a methods chapter

4 Results

TODO This is a results chapter

5 Discussion

TODO This is a discussion chapter

6 Conclusions

TODO This is a conclusions chapter