Aleksander Lempinen

MLOps approach for application specific performance tuning for machine learning systems

Master’s Thesis in Information Technology

January 17, 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: 13+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

There is little consensus on the exact definition of DevOps, but especially collaboration between development and operation is emphasized (Mishra and Otaiwi November 1, 2020; Waller, Ehmke, and Hasselbring April 3, 2015). DevOps can be studied from different points of view such as culture, collaboration, automation, measurements and monitoring (Mishra and Otaiwi November 1, 2020; Waller, Ehmke, and Hasselbring April 3, 2015). This thesis is mostly focused on the automation, measurements and monitoring parts of DevOps.

Continuous integration, continuous deployment and continuous monitoring are well known practices in DevOps (Waller, Ehmke, and Hasselbring April 3, 2015) describing the automatic nature of integrating, deploying and monitoring code changes. Performance profiling and monitoring are similar activities and the main difference is whether it's done during the development process or during operations respectively (Waller, Ehmke, and Hasselbring April 3, 2015). DevOps bridges the gap between evaluating performance during the development process and during operations (Brunnert et al. August 18, 2015).

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

2.2.1 Performance Metrics

Performance metrics are fundamental to all activities involving performance evaluation such as profiling or monitoring (Brunnert et al. August 18, 2015). Common metrics involve measuring the CPU, but other metrics to collect include memory, network or I/O but might not be as well defined as a CPU metric (Brunnert et al. August 18, 2015).

- Task Completion time
- Throughput
- Latency
- CPU usage
- GPU usage
- RAM usage
- VRAM usage
- I/O usage
- Network traffic

2.2.2

Preprocessing

Training

Serving Latency

Resource demands might change depending on the inputs (Brunnert et al. August 18, 2015) making it important to systematically measure performance not only based on code changes but also on configuration changes or even data changes.

2.3 MLOps

Performance measuring software is not new, but ML brings additional challenges in the form of models and data which requires a modified approach (Breck et al. 2017). It is also important to note, that not every data scientist or ML engineer working on ML systems has a software engineering background (Finzer 2013) and might lack the necessary knowledge to apply software engineering best practices to ML systems.

TODO what DevOps brings to ML

TODO Continuous Training

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

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)

5.4 Future Work

TODO This is a discussion chapter

6 Conclusions

TODO This is a conclusions chapter

Bibliography

Breck, Eric, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. 2017. “The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction”. In *Proceedings of IEEE Big Data.*

Brunnert, Andreas, Andre van Hoorn, Felix Willnecker, Alexandru Danciu, Wilhelm Hasselbring, Christoph Heger, Nikolas Herbst, et al. August 18, 2015. *Performance-Oriented DevOps: A Research Agenda,* arXiv:1508.04752, August 18, 2015. Visited on January 17, 2023. https://doi.org/10.48550/arXiv.1508.04752. arXiv: 1508.04752 `[cs]`. http://arxiv.org/abs/1508.04752.

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.

Finzer, William. 2013. “The Data Science Education Dilemma”. *Technology Innovations in Statistics Education* 7 (2). Visited on January 17, 2023. https://doi.org/10.5070/T572013891. https://escholarship.org/uc/item/7gv0q9dc.

Mishra, Alok, and Ziadoon Otaiwi. November 1, 2020. “DevOps and Software Quality: A Systematic Mapping”. *Computer Science Review* 38 (November 1, 2020): 100308. ISSN: 1574-0137, visited on January 17, 2023. https://doi.org/10.1016/j.cosrev.2020.100308. https://www.sciencedirect.com/science/article/pii/S1574013720304081.

Waller, Jan, Nils C. Ehmke, and Wilhelm Hasselbring. April 3, 2015. “Including Performance Benchmarks into Continuous Integration to Enable DevOps”. *ACM SIGSOFT Software Engineering Notes* 40, number 2 (April 3, 2015): 1–4. ISSN: 0163-5948, visited on January 17, 2023. https://doi.org/10.1145/2735399.2735416. https://doi.org/10.1145/2735399.2735416.