Aleksander Lempinen

MLOps approach for application specific performance tuning for machine learning systems

Master's Thesis in Information Technology

February 22, 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: 20+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. TODO Summary of thesis 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.

TODO Structure of the thesis

2 Machine Learning Operations

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

2.1.1 Overview

TODO what is ML

TODO supervised vs unsupervised vs reinforcement

Unsupervised learning has the advantage of not requiring labeled data which is an advantage for problems where labels are uncommon (Le et al. July 12, 2012). Unsupervised pretraining or feature learning on large datasets can yield to surprising results such as human facial recognition (Le et al. July 12, 2012). Unsupervised machine learning techniques can be used as part of a learning pipeline as a preprocessing or postprocessing step even if the rest of the system uses supervised machine learning.

TODO training vs serving

2.1.2 Machine learning algorithms

TODO Model evaluation

2.1.3 Hyperparameter optimization

Parameters given as part of a configuration to the machine learning model are called hyperparameters (Yang and Shami November 2020). Hyperparameter selection is a difficult and automatic tuning of hyperparameters can help achieve state-of-the-art performance (Maclaurin, Duvenaud, and Adams April 2, 2015). Hyperparameter optimization techniques include grid search, random search, gradient based optimization and Bayesian optimization and they have different benefits and limitations (Yang and Shami November 2020).

Neural Architecture optimization and Meta modeling are similar to hyperparameter optimization where model structure or modeling algorithm is treated as a tunable parameter (Baker et al. November 8, 2017). Traditional hyperparameter tuning methods such as Bayesian optimization are unfeasible for more than 10-20 hyperparameters (Maclaurin, Duvenaud, and Adams April 2, 2015).

Performance prediction is an important step to reduce the amount of computation required for neural architecture search and hyperparameter optimization (Baker et al. November 8, 2017). TODO early stopping reference.

2.2 DevOps

2.2.1 Overview

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.

TODO: picture about 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

2.2.2 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 such as memory usage, network traffic or I/O usage are not 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

2.2.3 Performance tuning

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

2.3.1 Overview

Requirements for a machine learning system are different depending on the task. For example speech and object recognition might have no particular performance requirements during training but strict latency and computational resource restrictions when deployed to serve large amounts users (Hinton, Vinyals, and Dean March 9, 2015). One of the key areas of MLOps is using machine learning in production systems in addition to data processing and machine learning model training.

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 machine learning engineer working on machine learning systems has a software engineering background (Finzer 2013) and might lack the necessary knowledge to apply software engineering best practices to machine learning systems.

TODO what DevOps brings to ML

TODO Continuous Training

2.3.2 AutoML

Machine learning systems in addition to machine learning performance metrics and system performance metrics will have their performance metrics tied to product or organization metrics such as user churn rate or click-through rate (Shankar et al. September 16, 2022). Choosing the right metrics to evaluate a machine learning system is important and the metrics will be different for different machine learning systems (Shankar et al. September 16, 2022).

Automated Machine Learning (AutoML) aims to minimize human intervention in completing data analytics tasks using machine learning algorithms (Yang and Shami November 2022).

2.3.3 Performance prediction and early stopping

3 Methods

3.1 Research setup

3.1.1 Scope

The scope of the study is limited to 5 performance metrics of 3 different ML models trained and tested on 3 different datasets using a distributed computing framework Ray Tune (Liaw et al. July 13, 2018).

TODO different models, different datasets

TODO Vertailukriteeristö: tapana ohjelmistopuolella + tapana koneoppimispuolella

TODO different resources (memory, time, accuracy)

3.1.2 Research Questions

This master's thesis asks the following research questions:

- *RQ1*: How early stopping affects performance metrics during hyperparameter optimization?
- *RQ2*: How early stopping affects performance metrics during neural architecture optimization?

3.1.3 Methodology

Methodology used is expanded from an existing methodology for machine learning experiment design (Fernandez-Lozano et al. December 1, 2016) to include AutoML and

3.2 Experiments

3.2.1 Datasets

MNIST (Deng November 2012)

3.2.2 Machine Learning algorithms

3.2.3 Metrics

3.2.4 Training and validation

3.2.5 Inference

4 Results

TODO This is a results chapter

5 Discussion

5.1 Research Questions revisited

5.1.1 Research question RQ1

5.1.2 Research question RQ2

5.1.3 Research question RQ3

5.2 Interpretation

5.2.1 Implications for research

5.2.2 Implications for practice

5.3 Limitations

5.3.1 Datasets

5.3.2 Machine Learning algorithms

5.3.3 Metrics

5.3.4 Training and validation

5.3.5 Inference

5.4 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)

5.5 Future Work

TODO This is a discussion chapter

6 Conclusions

Summary

TODO This is a conclusions chapter

Bibliography

Baker, Bowen, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. November 8, 2017. *Accelerating Neural Architecture Search Using Performance Prediction,* arXiv:1705.10823, November 8, 2017. Visited on January 25, 2023. https://doi.org/10.48550/arXiv.1705.10823. arXiv: 1705.10823 [cs]. http://arxiv.org/abs/1705.10823.

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.

Deng, Li. November 2012. "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]". *IEEE Signal Processing Magazine* 29, number 6 (): 141–142. ISSN: 1558-0792. https://doi.org/10.1109/MSP.2012.2211477.

Fernandez-Lozano, Carlos, Marcos Gestal, Cristian R. Munteanu, Julian Dorado, and Alejandro Pazos. December 1, 2016. "A Methodology for the Design of Experiments in Computational Intelligence with Multiple Regression Models". *PeerJ* 4 (December 1, 2016): e2721. ISSN: 2167-8359, visited on February 15, 2023. https://doi.org/10.7717/peerj.2721. pmid: 27920952. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5136129/.

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.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. March 9, 2015. *Distilling the Knowledge in a Neural Network,* arXiv:1503.02531, March 9, 2015. Visited on February 3, 2023. https://doi.org/10.48550/arXiv.1503.02531. arXiv: 1503.02531 `[cs, stat]`. http://arxiv.org/abs/1503.02531.

Le, Quoc V., Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. July 12, 2012. *Building High-Level Features Using Large Scale Unsupervised Learning,* arXiv:1112.6209, July 12, 2012. Visited on February 3, 2023. https://doi.org/10.48550/arXiv.1112.6209. arXiv: 1112.6209 `[cs]`. http://arxiv.org/abs/1112.6209.

Liaw, Richard, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. July 13, 2018. *Tune: A Research Platform for Distributed Model Selection and Training,* arXiv:1807.05118, July 13, 2018. Visited on February 22, 2023. https://doi.org/10.48550/arXiv.1807.05118. arXiv: 1807.05118 `[cs, stat]`. http://arxiv.org/abs/1807.05118.

Maclaurin, Dougal, David Duvenaud, and Ryan P. Adams. April 2, 2015. *Gradient-Based Hyperparameter Optimization through Reversible Learning,* arXiv:1502.03492, April 2, 2015. Visited on February 3, 2023. https://doi.org/10.48550/arXiv.1502.03492. arXiv: 1502.03492 `[cs, stat]`. http://arxiv.org/abs/1502.03492.

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.

Olson, Randal S., William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore. December 11, 2017. "PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison". *BioData Mining* 10, number 1 (December 11, 2017): 36. ISSN: 1756-0381, visited on February 22, 2023. https://doi.org/10.1186/s13040-017-0154-4. https://doi.org/10.1186/s13040-017-0154-4.

Shankar, Shreya, Rolando Garcia, Joseph M. Hellerstein, and Aditya G. Parameswaran. September 16, 2022. *Operationalizing Machine Learning: An Interview Study,* arXiv:2209.09125, September 16, 2022. Visited on December 7, 2022. https://doi.org/10.48550/arXiv.2209.09125. arXiv: 2209.09125 [cs]. http://arxiv.org/abs/2209.09125.

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.

Yang, Li, and Abdallah Shami. November 2020. "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice". *Neurocomputing* 415 (): 295–316. ISSN: 09252312, visited on January 25, 2023. https://doi.org/10.1016/j.neucom.2020.07.061. arXiv: 2007.15745 [cs, stat]. http://arxiv.org/abs/2007.15745.

———. November 2022. "IoT Data Analytics in Dynamic Environments: From An Automated Machine Learning Perspective". *Engineering Applications of Artificial Intelligence* 116 (): 105366. ISSN: 09521976, visited on January 25, 2023. https://doi.org/10.1016/j.engappai.2022.105366. arXiv: 2209.08018 [cs, eess]. http://arxiv.org/abs/2209.08018.