Parallelization of Neural Network Training

Final presentation | High-Performance Big Data and Artificial Intelligence Systems (高效能巨量資料與人工智慧系)

Lukas Voss | 5 June 2023

Agenda

Parallelization of Neural Network Training

1	**Motivation**	• Growing impact of Artificial Intelligence (AI) makes efficient resource use crucial for advancements • GPU provides attractive parallelization capabilities for matrix-operations like in Neural Networks
2	**Implementation + Results**	• CPU vs. GPU in like-for-like analysis for image classification with MNIST data set • Varying hyperparameters to investigate impact on performance and accuracy KPIs
3	**Summary and Outlook**	• Key findings and highlights • Discussion and Q&A

Agenda

Parallelization of Neural Network Training

1 **Motivation**

- Growing impact of Artificial Intelligence (AI) makes efficient resource use crucial for advancements
- GPU provides attractive parallelization capabilities for matrix-operations like in Neural Networks

2 Implementation + Results

- CPU vs. GPU in like-for-like analysis for image classification with MNIST data set
- Varying hyperparameters to investigate impact on performance and accuracy KPIs

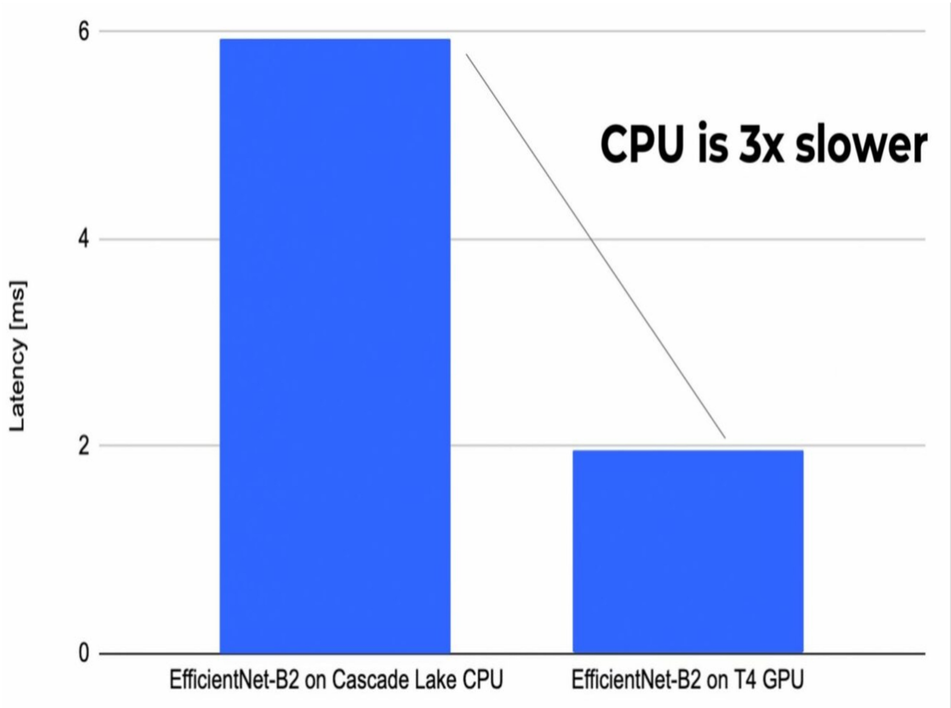
3 Summary and Outlook

- Key findings and highlights
- Discussion and Q&A

Training of Neural Networks

CPUs are not well-suited to handle workload and lead to prohibitively long training times

The current performance gap [1]



Deep Learning's popularity >10x over the last decade [2]

- Research on deep learning models has accelerated due to increased data availability, better algorithms and hardware
- Especially in focus for AI-related tasks

- Text, object and speed recognition for NLP tasks
- Robotics, including autonomous vehicles and robotic control systems

- Widely available CPUs cannot efficiently process the amount of used data for training large models
- Time and resource constraints become important factors

Deep Dive: Motivation

AI has the potential to lead to major efficiency improvements across industries

“

It's very clear that AI is going to impact every industry. I think that every nation needs to make sure that AI is a part of their national strategy. Every country will be impacted.

- Jensen Huang[1]

Healthcare

- Enhance diagnostics, e.g., medical imaging analysis
- Predict disease outcomes
- Enable personalized medicine, and facilitate drug discovery

Education

- Intelligent educational content generation
- Support personalized learning
- Adaptive tutoring systems, automated grading

1) Founder and CEO of NVIDIA; Quote taken from an interview in 2018

Agenda

Parallelization of Neural Network Training

1 **Motivation**

- Growing impact of Artificial Intelligence (AI) makes efficient ressource use crucial for advancements
- GPU provide attractive parallelization capabilites for matrix-operations like in Neural Networks

2 **Implementation + Results**

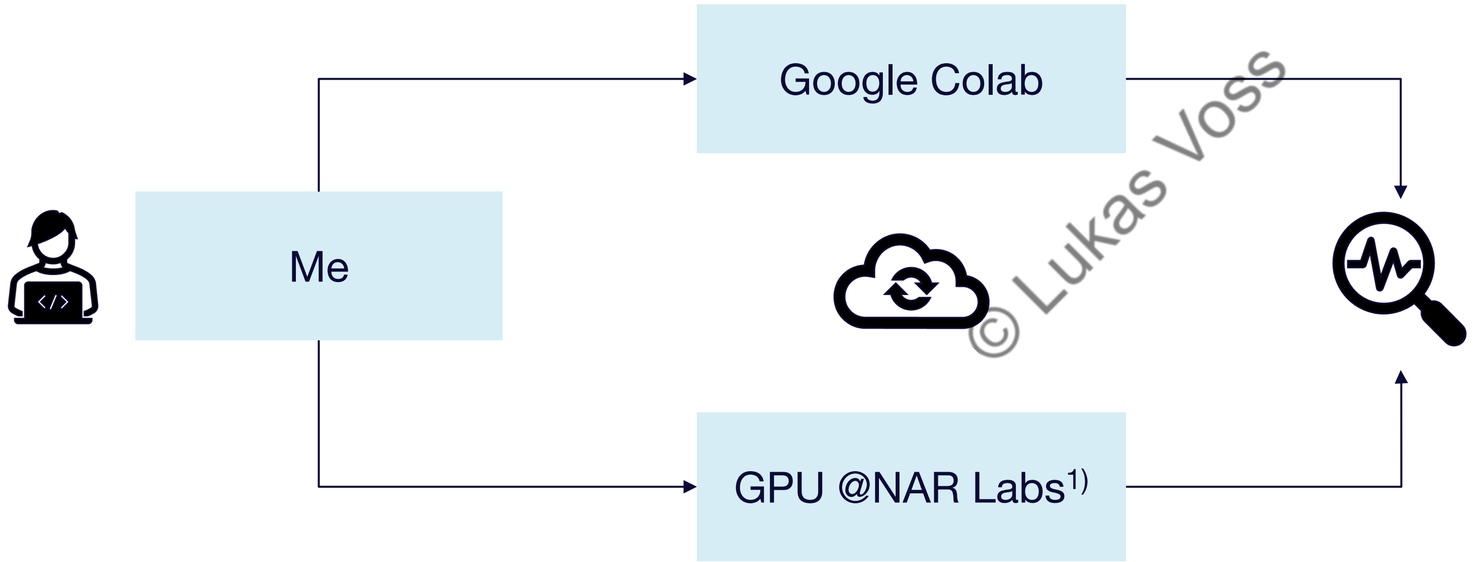
- CPU vs. GPU in like-for-like analysis for image classification with MNIST data set
- Varying hyperparameters to investigate impact on performance and accuracy KPIs

3 **Summary and Outlook**

- Key findings and highlights
- Discussion and Q&A

System Architecture

Access to GPU was granted by Taiwan High-Performance Computing Cluster NAR Labs



System Design

- An account at NAR Labs enabled us to use their NVIDIA RTX 3070 GPU as a cloud-based service
- Convenient access on-demand for improved flexibility

1) National Center for High-Performance Computing

Neural Network Training

A comprehensive approach to test performance potential lifted by GPU

Dataset

- Image recognition task with 28x28 pixel on a grayscale
- The MNIST dataset has handwritten digits with 60,000 training and 10,000 test images

Method

- Convolutional Neural Network with ReLu activation and pooling layer
- Softmax output layer

Evaluation and benchmarking again CPU

- Conduction like-for-like comparisons[1)] between Google Colab CPU and GPU
- GPU will be tested with and without explicit data loading parallelization

Note MLP: multi-layer perceptron **1)** Evaluation on the same task

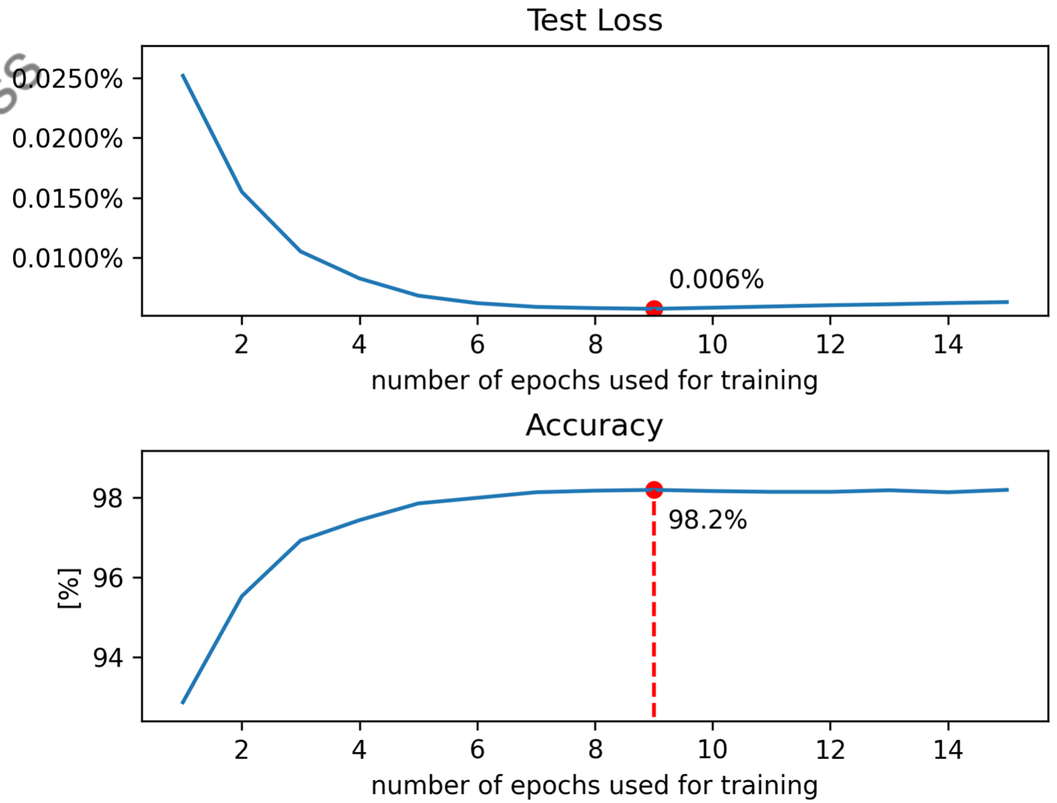
Google Colab CPU: Training epochs

After a sweet spot of #epochs has been surpassed, more epochs do more harm than good

Parameters

Execution Time	48 sec
Optimizer	SGD[1]
Learning Rate	0.1

Performance



1) Stochastic Gradient Descent

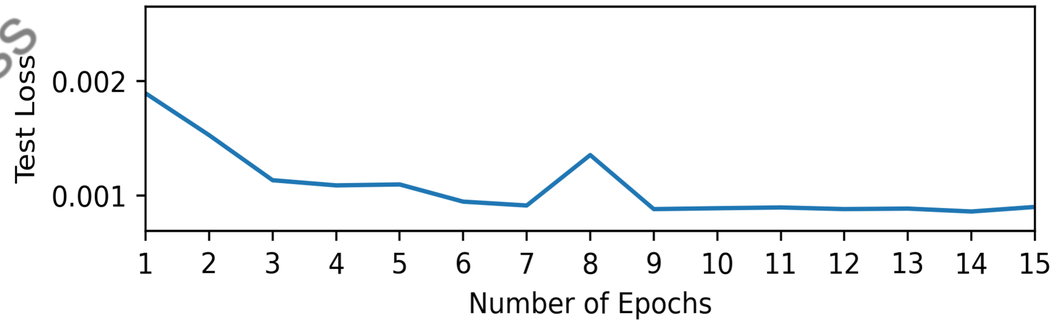
NVIDIA RTX 3070: Training epochs

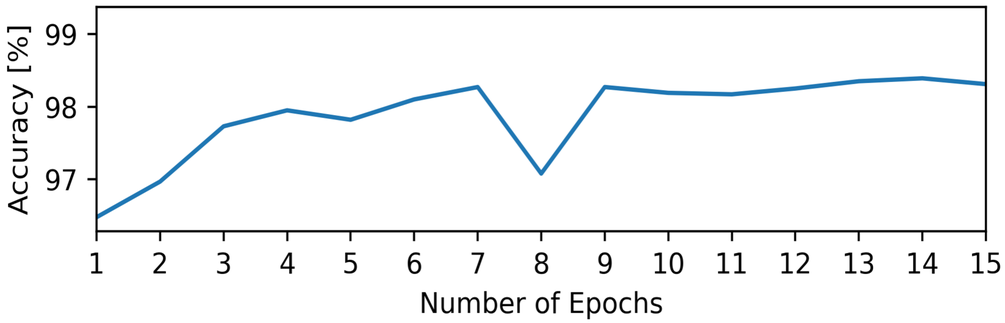
Data loading operations parallelized with four worker processes leads to performance gain

Parameters

Execution Time	16 sec
Optimizer	SGD[1]
Learning Rate	0.1

Performance



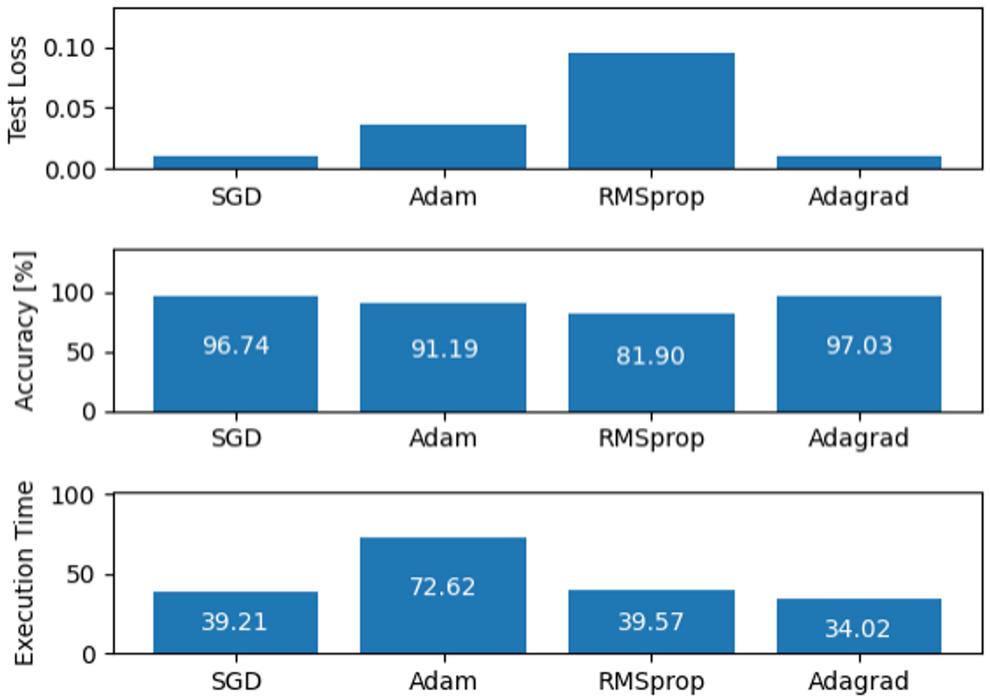


1) Stochastic Gradient Descent

Deep Dive: Variation of Optimizers

Leveraging the capabilities of the GPU, we achieve a significant speed up in training time

Google Colab CPU: Intel(R) Xeon(R) CPU @ 2.20GHz

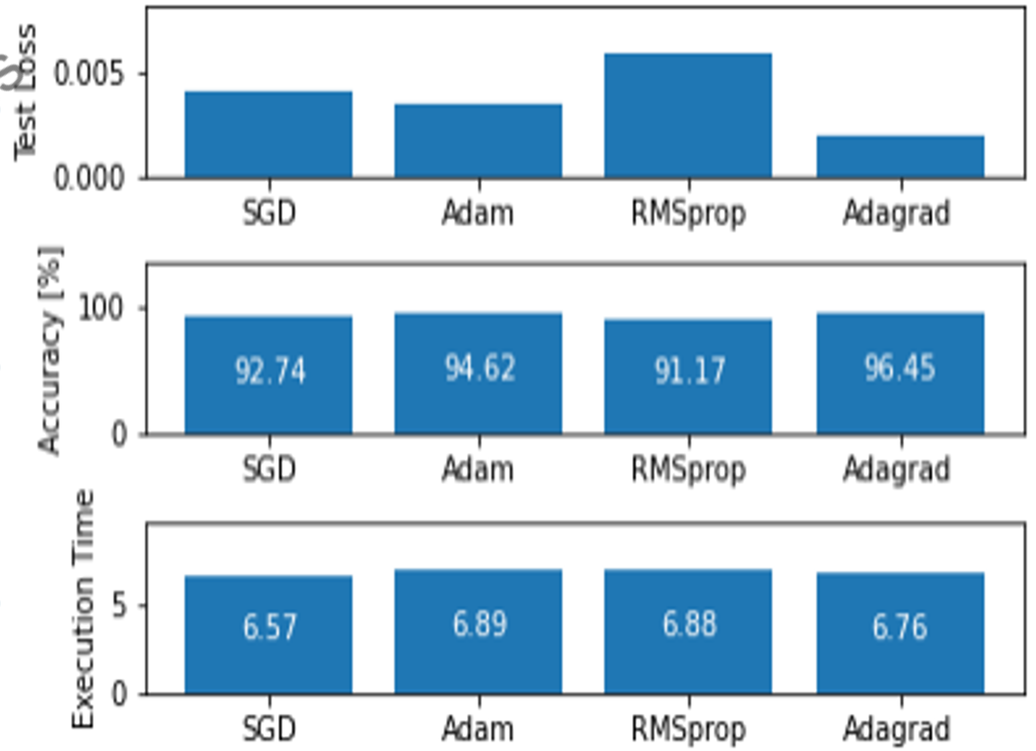


10x

similar

>6x

GPU: NVIDIA RTX 3070

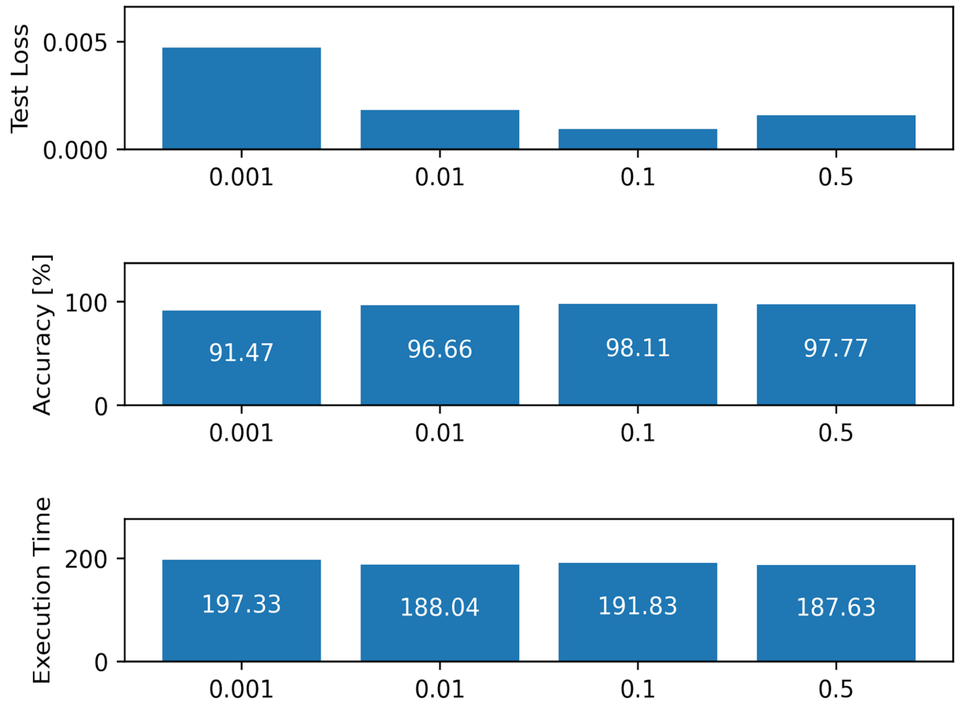


Improvement by GPU

Deep Dive: Learning Rates

Leveraging the capabilities of the GPU, we achieve a significant speed up in training time

Google Colab CPU: Intel(R) Xeon(R) CPU @ 2.20GHz

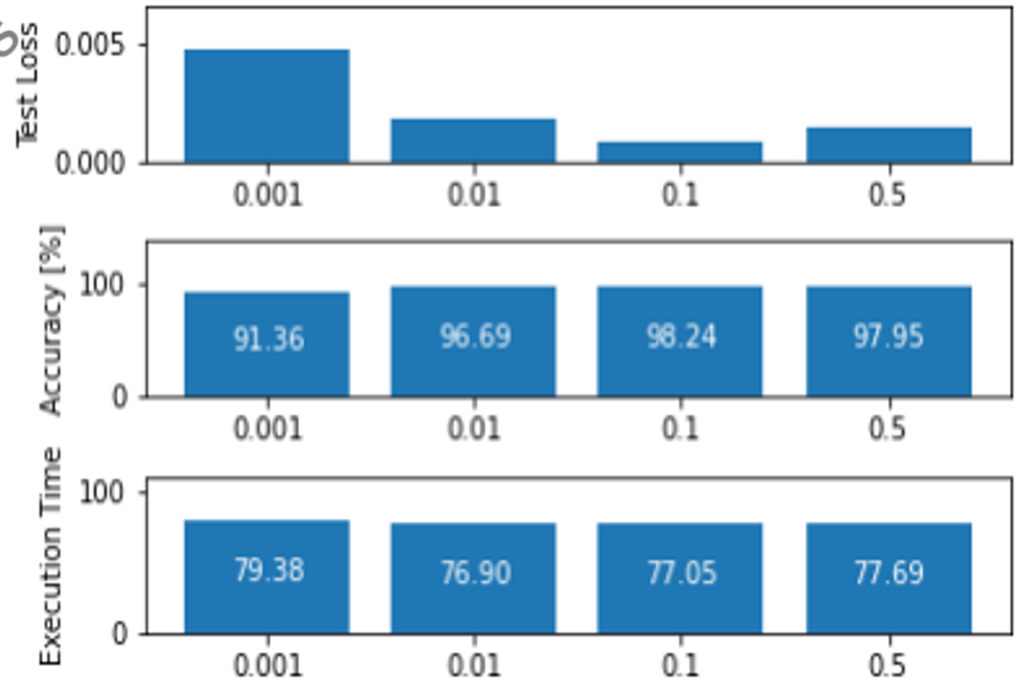


same

same

>2x

GPU: NVIDIA RTX 3070

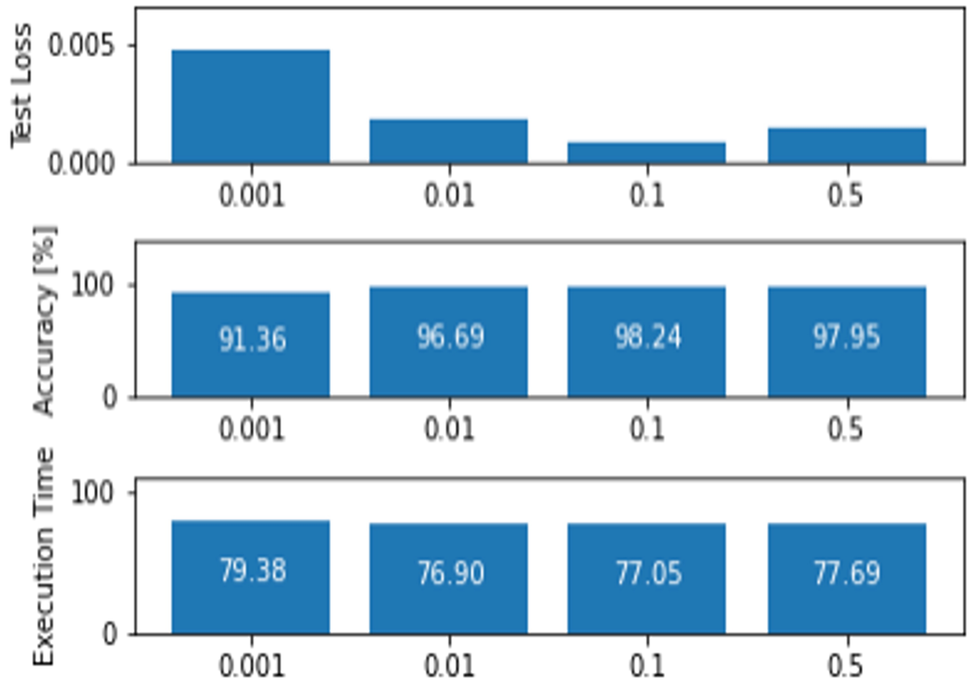


Improvement by GPU

Data Parallelization on GPU: Learning Rates

Data loading operations **parallelized with four worker processes** leads to performance gain

Without Data Parallelization: Learning Rates

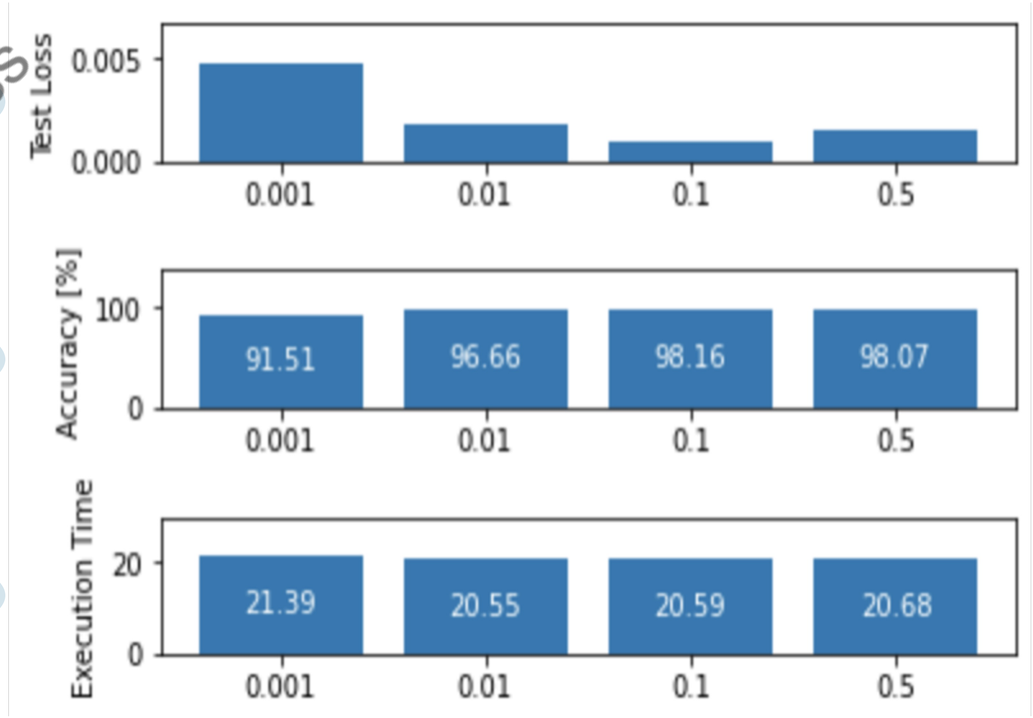


same

similar

ca. 4x

With Data Parallelization: Learning Rates

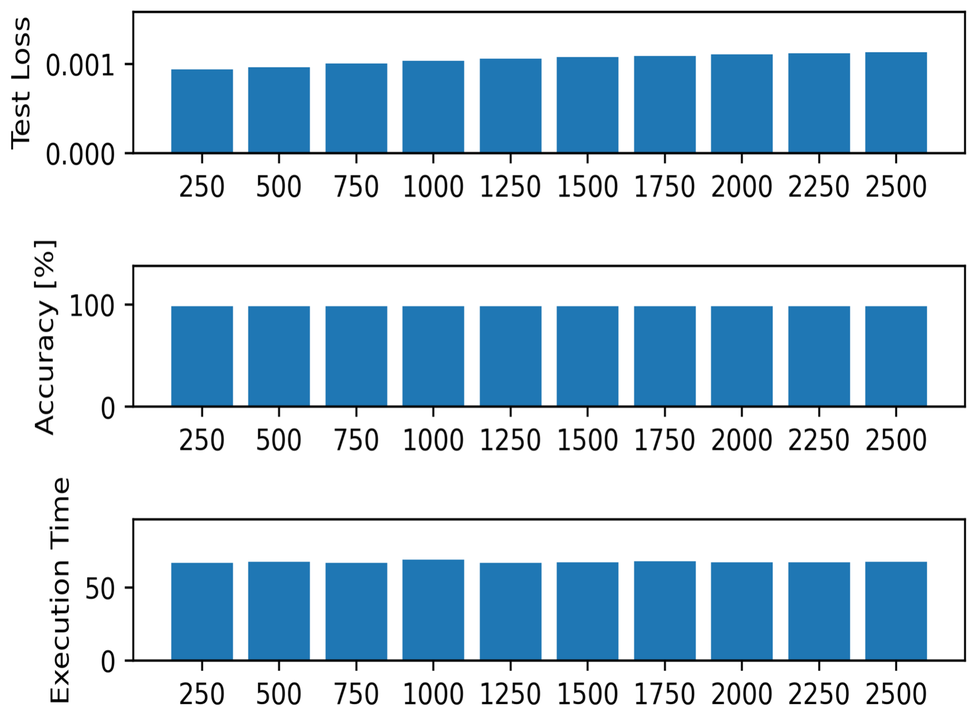


Improvement by GPU

Data Parallelization on GPU: Batch Size

Data loading operations **parallelized with four worker processes** leads to performance gain

Without Data Parallelization: Batch Size

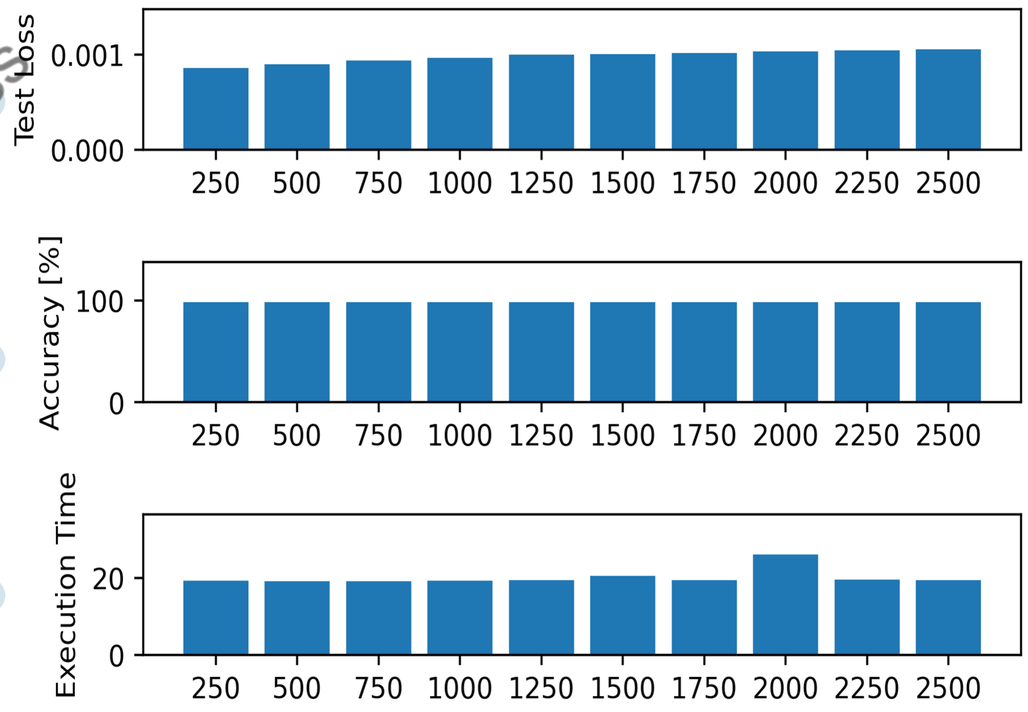


similar

similar

ca. 3x

With Data Parallelization: Batch Size



Improvement by GPU

Agenda

Parallelization of Neural Network Training

1 **Motivation**

- Growing impact of Artificial Intelligence (AI) makes efficient ressource use crucial for advancements
- GPU provide attractive parallelization capabilites for matrix-operations like in Neural Networks

2 **Implementation + Results**

- CPU vs. GPU in like-for-like analysis for image classification with MNIST data set
- Varying hyperparameters to investigate impact on performance and accuracy KPIs

3 **Summary and Outlook**

- Key findings and highlights
- Discussion and Q&A

Executive Summary

Customizing the GPU‘s inbuilt capacity to parallelize further improves performance

Performance

- Significant **improvements** in **execution time up to 8x faster** compared to CPU
- Model has different sensitivity to different parameters

Distributed Architecture

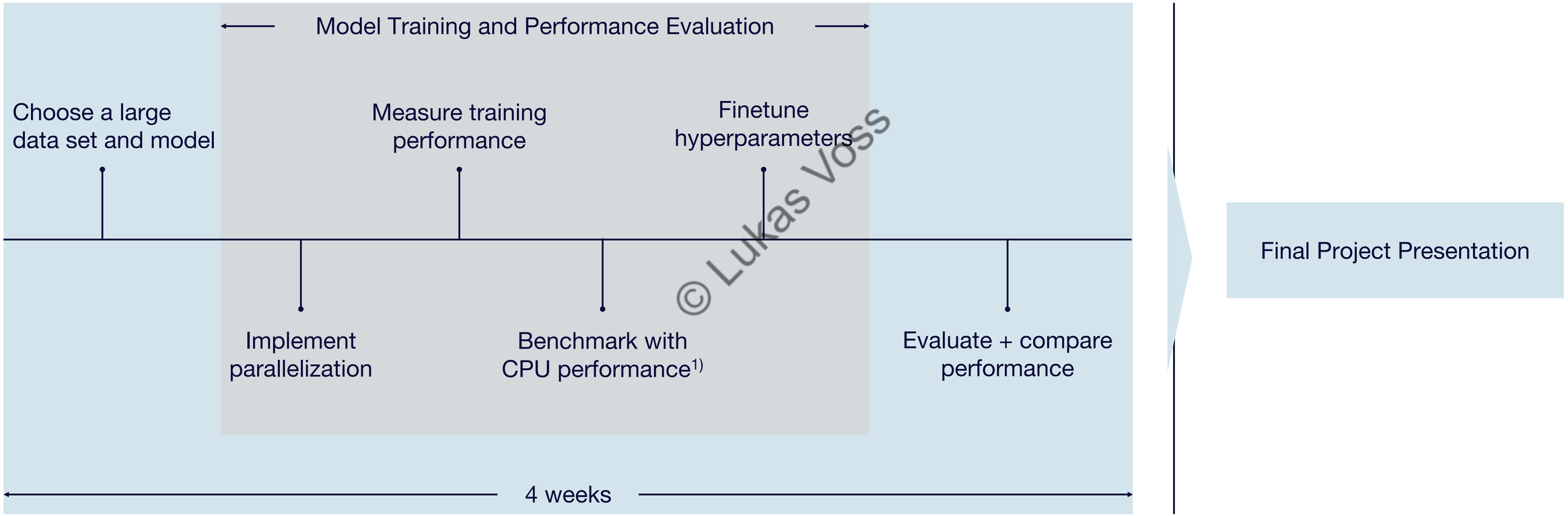
- Efficiently load and preprocess data in mini-batches during the training and testing phases

Outlook

- Fine-tuning the implementation on the architecture
- Explore model parallelism
- Consult with industry experts and literature

Appendix: Timeline of Implementation

Top-level view of the proposed research idea: Parallelization of Deep Learning Training



1) Google Colab CPU: Intel(R) Xeon(R) CPU @ 2.20GHz

Appendix: Literature

Overview of used research articles and other sources

- [1]: Deci AI: Can You Close the Performance Gap Between GPU and CPU for Deep Learning Models?
- [2]: Md Jaber Al Nahian et al. „Camera Model Identification using Deep CNN and Transfer Learning Approach“ (2019)
- [3] M. Abadi et al. "Deep Learning on GPUs versus CPUs: A Performance Comparison" (2016)
- [4] Y. Liu et al. "Deep Learning on CPU and GPU: Performance and Energy Considerations" (2017)
- [5] D. Ciresan et al. "Accelerating Deep Convolutional Neural Networks using Specialized Hardware" (2018)
- [6] C. Zhang et al. "Comparison of Deep Learning Training on CPUs, GPUs, and FPGAs" (2019)
- [7] J. Li et al. "Accelerating Deep Learning via Speed-of-Light Neural Networks on GPUs and CPUs" (2020)