



Sunna Torge Center for Information Services and High Performance Computing (ZIH)

Get Together – HPC, Big Data, and Machine Learning

ISC Workshop Machine Learning on HPC Systems 25th June 2020



Introduction

- **Necessity** to bring research fields together
- **Supported** by interdisciplinary
 - exchange (conferences, workshops,
 - summerschools, etc.)
 - institutional structures (competence centers, research institutions, ...)
 - education (lecture series, trainings, ...)











Introduction

Realisation in Dresden/Leipzig

- Competence Center ScaDS.Al
- Research projects within ScaDS.Al





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Overview

- The Big Data Competence Center ScaDS at Dresden/Leipzig
- The Center for Information Services and High Performance Computing (ZIH), TU Dresden
- Towards Artificial Intelligence and Machine Learning
- Selected Research Projetcs











Overview

- The Big Data Competence Center ScaDS at Dresden/Leipzig

- The Center for Information Services and High Performance Computing (ZIH), TU Dresden
- Towards Artificial Intelligence and Machine Learning
- Selected Research Projetcs





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Big Data Competence Center ScaDS Dresden/Leipzig





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Big Data Competence Center ScaDS Dresden/Leipzig

Scalable Data Services and Solutions (ScaDS)

Major goals

- Research and development for Big Data applications
- Transfer of methods and algorithms into new domains
- Support the full data life cycle

Phase 1: October 2014 – September 2018 Phase 2: October 2018 – September 2021 Scads.Al: November 2019 – December 2022









Structure of ScaDS Dresden/Leipzig





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Lessons learned

- Most important: bring experts together to investigate requirements of data-intensive applications and derive solution
- Service Center approach within the center:
 Connect experts and application domain scientists





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







ScaDS Service Center at ZIH



Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Overview

- The Big Data Competence Center ScaDS at Dresden/Leipzig
- The Center for Information Services and High Performance Computing (ZIH), TU Dresden
- Towards Artificial Intelligence and Machine Learning
- Selected Research Projetcs





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Center for Information Services and HPC (ZIH)

- Central scientific unit at TU Dresden
- Running computing and communication infrastructure for the university
- Development of algorithms and methods:
 Cooperation with users from all departments
- Infrastructure and higher-level service for scientists all over Saxony
- Dresden CUDA Center for Excellence
- Dresden Intel® Parallel Computing Center (IPCC)
- Competence center for "Parallel Computing and Software Tools"
- Competence center for Big Data ScaDS Dresden/Leipzig

Director Prof. Wolfgang E. Nagel

about 160 employees, 70 researchers











Areas of Expertise

- Research topics
- Scalable software tools to support the optimization of applications for HPC systems
- Data intensive computing and data life cycle
- Performance and energy efficiency analysis for innovative computer architectures
- Distributed computing and cloud computing
- Data analysis, methods and modelling in life sciences
- Parallel programming, algorithms and methods
- Pick up and preparation of new concepts, methods, and techniques
- Teaching and Education













Data Intensive Computing at ZIH

- HPC system design expertise
- Focussing on data intensive tasks for more than 15 years
- HRSK-I
- Two machines, one HPC, one Throughput, one Capability
- Lots of tape drives to move data in and out (SGI CXFS), almost 2 GB/s to tape in 2006
- HRSK-II
- Island concept with HPC and Throughput
- High I/O bandwidth
- HDD+SSD file system, 100 GB/s to disk and lots of IOPS
- Hardware installed (including DLR): 50 Mio. € in total
- Last investment: HPC-DA for machine learning and Big Data (hardware components: 10 Mio. €)

In 2020: two new systems for applications with high-memory requirements







Get Together – HPC, Big Data, and Machine Learning Sunna Torge











Get Together – HPC, Big Data, and Machine Learning Sunna Torge















Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Provision of Data Analytics@HPC

- Provisioning of required environments
 (Hadoop, Spark, Flink, ML-frameworks, ...)
- Big Data session created on demand
- Run directly as analytics service at HPC site
- Adoptable to other frameworks/applications







Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Overview

- The Big Data Competence Center ScaDS at Dresden/Leipzig
- The Center for Information Services and High Performance Computing (ZIH), TU Dresden
- Towards Artificial Intelligence and Machine Learning
- Selected Research Projetcs





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Long term continuation of ScaDS Dresden/Leipzig as one Germany wide center for data analytics and artificial intelligence

ScaDS.Al Dresden/Leipzig Data centric research (Big Data) Image: Control of the control



Get Together – HPC, Big Data, and Machine Learning Sunna Torge











Get Together – HPC, Big Data, and Machine Learning Sunna Torge











Get Together – HPC, Big Data, and Machine Learning Sunna Torge









ScaDS.Al started in November 2019 as Big Data / Al competence center!



Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Overview

- The Big Data Competence Center ScaDS at Dresden/Leipzig
- The Center for Information Services and High Performance Computing (ZIH), TU Dresden
- Towards Artificial Intelligence and Machine Learning
- Selected Research Projetcs





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Research Projects within ScaDS.AI

- Collaborative research
- Domain-specific text analysis
- Performance analysis of machine learning applications











Collaborative Research (Examples)

Settlement development using Computer vision



- Segmentation for structure detection in biological data using Crowd sourcing





typical user inputs

[Dmitrij Schlesinger, Peter Winkler TU Dresden, IÖR, MPI-CBG]



Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Machine Learning for Photogrammetry

- Content based Image Retrieval in digital image repositories within project HistStadt4D
- Goal: prototype for object recognition, esp. historical images
- Main problem: small training data

Using pretrained convolutional deep neural networks (VGG16) with TensorFlow on HPC

- Use case for hyperparameter optimization service



[https://www.urbanhistory4d.org/wordpress/]



Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Machine Learning for Photogrammetry - Results





Hofkirche





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Domain Specific Text Analysis

- Named Entity Recognition (NER) on domain specific entities
- Transfer Learning approaches for NER
- NER on domain-specific German text documents
- Common topics in German text documents









Text Analysis in the Financial Domain

Common project with industrial partner



Named Entity Regognition (NER) of domain specific entities

Thisprospectus org			upplement is not complete without, and may not be delivered orutilized except in connection with, the prospectus, including any amendmentsor supplements to it. There is currently no public market for our common											
stock.	GPE	We currer	ntlyintend to have our sh	ares quoted on the	TCQB PERSON	operated by	the OTC Markets org	although we may,	not be success	ful and our shares may never be quoted andowners c	of our Com	mon		
Stoo	Stock PERSON may not have a market in which to sell the shares. Also, no estimate may be given as to the time that this application processwill require. This prospectus supplement incorporates into our prospectus the													
Inform	ationcont	ained in o	Quarterly Report on	Form 10-Q for WORK	OF_ART the	quarterly DATE	period endedSeptembe	r 30, CARDINAL	2017 filed with	the Securities and Exchange Commission org	onNovem	ber 7,		
201	7 DATE	, and our	Current Report org	on Form 8-K filed with	the Securities	sand Exchange	Commission org on	November 22, 201	7 DATE , eac	ch attached hereto.Investing in our common stock invo	olves risks.	GPE		

Data

- ~ 10.000 text documents, labeled with different entity names
- entity "maturity date"
- training data (~ 1.000 documents) and test data (100 documents)









Approach and Insights

- Convolutional Neural Network (CNN) for NER
- Transfer Learning
 - Reuse of trained english models as starting point for new tasks or domains
 - Compare models trained with respect to other entities and blank models
- Use of NER standard tool (SpaCy)
- Recognition of entity ,maturity date' does not profit from pretraining with respect to other entities
- ,Blank' model yields best F1-score









Named Entity Recognition on Domain-Specific German Text Corpus

Questions:

Is transfer learning helpful for domain specific tasks?

Which size of annotated domain specific data is necessary?





Get Together – HPC, Big Data, and Machine Learning Sunna Torge







SmartData Corpus

- Published 2018 by DFKI
- Consists of news, RSS- and Tweet documents (2598)
- 16 different entity classes (~ 20.000 annotated entities)
- domain specific entities (transport and industry)

VerkehrNRW

twitter.com

A1 Bremen Richtung Münster zwischen Osnabrück-Nord und Kreuz Lotte/Osnabrück 5 km Stau (Zeitverlust: etwa eine Viertelstunde)



A38 Halle Richtung Leipzig

polizei.sachsen.de

A38 Halle Richtung Leipzig zwischen Leipzig-Neue Harth und Leipzig-Südost Baustelle, bis 30.06.2016



Wincor Nixdorf-Aktie plus 19 Prozent - Übernahme... finanznachrichten.de

Wincor Nixdorf-Aktie plus 19 Prozent - Übernahme durch Diebold glückt in letzter Minute Die Übernahme des Geldautomatenherstellers Wincor Nixdorf ist ...

[Schiersch2018]

[Schiersch2018] Schiersch, Martin, et al. "A german corpus for fine-grained named entity recognition and relation extraction of traffic and industry events." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). 2018.









Transfer Learning - Experiments

Deep Neural Network for NER

— BiLSTM-CRF [Huang2015, Reimers2017]

Train Model for 15 out of 16 entities with complete training data

— Source Model (Baseline)

Train Model for remaining entity (10%, 20%, ..., 70% of training data)

- Without use of source model
- Transfer of weights without output layer
- Transfer of weights with output layer

[Huang2015] Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991*(2015). [Reimers2017] Reimers, Nils, and Iryna Gurevych. "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks." *arXiv preprint* arXiv:1707.06799 (2017).









Transfer Learning (BiLSTM)



Entity, Duration

0.84 0.83 TF all

TF without_last_layer without

40%







50%

60%



70%



Similarity Analysis on Service Desk Tickets

Common project with Service Desk at ZIH, TU Dresden

Goal:

— Automated recognition of topic clusters in user requests



— Unlabeled german text corpus

Methods:

- Clustering algorithm on vector representation of text documents
- Transfer Learning approach, based on ,next sentence prediction' with Deep Learning











Topic Analysis of German Text Corpus

Common project with Institute of Lightweight Engineering and Polymer Technology and the Chair of Applied Linguistics, TU Dresden

Goal:

Analysis of german corpus of master and PhD thesis with respect to common topics

Methods:

- Linguistic methods
- Transfer Learning
- Unsupervised Machine Learning





Get Together – HPC, Big Data, and Machine Learning Sunna Torge Slide 41







Performance Analysis of Machine Learning Applications

- Benchmarking Natural Language Processing
- Development and Analysis of Distributed ML Benchmarks using tensorFlow









Benchmarking Natural Language Processing



Machine Learning Benchmark **MLPerf**

- Translation Task
- Deep neural network, transformer architecture
- Measure of wall clock time during training

ATTEN LIBERTIN	[Latin: related to p. Bioton 1.
famos or	dictatorial /diktar
) one of	like a dictator, 2 over
e played	orially adv. ILatin
at risks,	TATOR]
cut into	diction /'dikf(a)n/ n
	ciation in speaking manner
Sector State	dictio from dico dict or singin
	dictionary "dala cay]
risky.	book listing (ukjeneri/ n ()
	book insuing (usu. alphabeti
and di	explaining the words of a l
es) al-	giving corresponding word
efined	language, 2 reference hords in
ed to	the terms of pook e
	the terms of a particular

Goal:

Run reference implementation on ZIH infrastructure and compare performance results.

Insight:

Performance of reference implementation is comparable.



Get Together – HPC, Big Data, and Machine Learning Sunna Torge Slide 43







Distributed ML Benchmarks using TensorFlow

Distribution of training

Comparison of external module ("Horovod") and TF version 2.0 "strategies"

User requirements

- Analysis of raw performance and convergence
- Ease of use

Existing benchmarks

- Usually measure raw performance on synthetic data only
- Detailed analysis may not be transferable to real-world use cases

Need for representative example of DNN training

Compare raw and convergence performance (time to solution)

Analyze with in-house tools (Score-P, Vampir, ...) and provide optimization guidance









Distributed ML Benchmarks using TensorFlow

Comparison of Frameworks

- TensorFlow CNN Benchmark for image classification on NVIDIA V100 GPUs
- Horovod shines in the case of multi-node scalability whereas native TensorFlow performs well on a single node



TensorFlow CNN Benchmark ResNet-50 Training









Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Conclusion

- The Big Data Competence Center ScaDS at Dresden/Leipzig
- The Center for Information Services and High Performance Computing (ZIH), TU Dresden
- Towards Artificial Intelligence and Machine Learning
- Selected Research Projects within ScaDS.AI

Further Topics:

- Teaching
- Expansion of services and user defined environments
- Provide services also outside academia











Thank You!

Bernd Grüber Alexander Grund Waldemar Hahn Silvia Hentschel Christoph Lehmann Taras Lazariv Simon Meier-Vieracker

René Jäkel

Lena Jurkschat Norman Koch Wolfgang E. Nagel Andrei Politov Erhard Rahm Dmitrij Schlesinger Peter Winkler



Get Together – HPC, Big Data, and Machine Learning Sunna Torge







Thank You!

Contact

Sunna Torge

Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) Technische Universität Dresden 01062 Dresden

email: sunna.torge@tu-dresden.de



Get Together – HPC, Big Data, and Machine Learning Sunna Torge





