ETH zürich

Torsten Hoefler

Deep500: An HPC Deep Learning Benchmark and Competition Birds of a Feather, SC18, Nov. 2018, Dallas, TX

0100

EuroMPI'19 September 11-13 2019 Zurich, Switzerland https://eurompi19.inf.ethz.ch Submit papers by April 15th!



Trends in deep learning: hardware and multi-node

The field is moving fast – trying everything imaginable – survey results from 240 papers in the area of parallel deep learning



Deep Learning is largely on distributed memory today!

Trends in distributed deep learning: node count and communication

The field is moving fast – trying everything imaginable – survey results from 240 papers in the area of parallel deep learning



Deep Learning research is converging to MPI!



Parallelism in Deep Learning

- Individual operators
- Network parallelism
- Optimization algorithm
- Distributed training



and the second real

Deep Learning is Supercomputing!





Challenges

Different:

- Communication schemes
- Model consistency requirements
- Software stacks and feature sets

Need to define:

- Open datasets from computational sciences
- Metrics robust to methods (or freeze methods)
- Standard benchmarking infrastructure





Training Agent Training Agent Training Agent Training Agent









So how to not do this

"Twelve ways to fool the masses when reporting performance of deep learning workloads" (my humorous guide to floptimize deep learning)



*** SPEL

HPC picking up!

1) Ignore accuracy when scaling up!

- Too obvious for this audience
 - Was very popular in 2015!



the second

***SPCL

2) Do not report test accuracy!

Training accuracy is sufficient isn't it?



All the second and



3) Do not report all training runs needed to tune hyperparameters!







So how to not do this

"Twelve ways to fool the masses when reporting performance of deep learning workloads" (my humorous guide to floptimize deep learning)



Modular Benchmarking Infrastructure for Reproducible DL

- Separates benchmarking into the 4 core components
- Metrics defined separately, shared across levels
- **Leverages ONNX for model definition**
- Contains reference implementations of operators, optimizers, and distributed schemes
- Supports custom C, C++, and CUDA implementations on all levels
 - No need to reimplement an optimizer to replace gradient compression!







Ben-Nun et al. "A Modular Benchmarking Infrastructure for High-Performance and Reproducible Deep Learning", soon on arXiv

A Modular Benchmarking Infrastructure for High-Performance and Reproducible Deep Learning

Tal Ben-Nun, Simon Huber, Maciej Besta, Alexandros Nikolaos Ziogas, Daniel Peter, Torsten Hoeffer Department of Computer Science, ETH Zurich

Advance—We include Deep806 the first constraints howeds. our different platforms, and executing caterian and a star-physical platforms and a star-platform and a star-platform and a star-platform and a platform of deep learning frameworks, algorithms, libraries, and lexibalitys. The lexip face betted free2006 kit is modules. Table 11 illustrates how various DL frameworks, libraries, and operators, network precessing, training, and distributed training. Table 11 illustrates and lexip face betted free2000 kit is modules forestanting interactions that Registrate is a contraciable (multiss). and benchmarking different deep learning codes) and *fair* (uses carefully selected metrics). Moreover, Deep500 is ast (incurs negligible overheads), *verifiable* (offers infrastructure screens, and *reproducible*. Finally, as the first producible software infrastructure to utilize the infrastructure to utilize the accuracy-haled aspects of DL (e.g., time required 09 provides software infrastructure to utilize the social specific test-set accuracy) and performance-related to ensure a specific test-set accuracy) and performance-means asset (e.g., communication volume). Of Performance means uted Deep Learning, High-Performance Deep Learning, Parallel Deep Learning, Benchmarking ducibility: https://aithub.com/doen500/deen500

I. INTRODUCTION

g (DL) has transformed the world and is no ubiquitous in areas such as speech recognition, image classi fication, or autonomous driving [3]. Its central concept is a Deep Neural Network (DNN), a structure modeled after the human brain. Thanks to rigorous training, DNNs are able to various problems, previously deemed unsolvable Recent years saw an unprecedented growth in the numbe of approaches, schemes, algorithms, applications, platforms

and frameworks for DL. First, DL computations can aim at inference or training. Second, hardware platforms can vary significantly, including CPUs, GPUs, or FPGAs. Third, oper-ators can be computed using different methods, e.g., im2col [5] or Winograd [26] in convolutions. Next. DL functionalities have been deployed in a variety of frameworks, such as TensorFlow [14] or Caffe [20]. These functionalities may in rate many parallel and distributed optimizations, such a data, model, and pipeline parallelism. Finally, DL workload are executed in wildly varying environments, such as mobile phones, multi-GPU clusters, or large-scale supercomputers. This richness of the DL domain raises a question we seen addressed so far: How can one ensure

leveled, fair ground for comparison, competition, and benchmarking in Deep Learning? The key issue here is that cent benchmarking approaches such as DAWNBench [9] or MLPerf [30] are merely lists of results that do not directly sider the rich nature of today's DL efforts.

To answer this question, we propose Deep500: a bench marking system that enables fair analysis and comparison of diverse DL efforts. Deep500 is based on the following five pillars:
 Customizability,
 Metrics,
 Performance, Validation, and @ Reproducibility. 0 "Customizability" in dicates that Deep500 enables benchmarking of arbitrary com binations of DL elements, such as various frameworks rur

is the only system that focuses on performance, accuracy, and convergence, while simultaneously offering a wide spectrum of metrics and criteria for benchmarking, enabling ability of design, and considering a diversity of workloads (L) cuDNN (L) MKL-DNP

embraces a complex nature of DL that, unlike benchmark such as Top500 [15], makes a single number such as FLOPS as

insufficient measure. To this end, we propose metrics that con

can be integrated with parallel and distributed DL codes.

"Validation" indicates that Deep500 provides infrastructure t ensure correctness of aspects such as convergence. Finally

Deep500 embraces @ "Reproducibility" as specified in recen

Table II compares Deep500 to other benchmarking infra

ructures with respect to the offered functionalities. Deep50

nitiatives [18] to help developing reproducible DL cod

that Deep500 is the first DL benchmarking inf



Fig. 11: Scaling Analysis of Level 3: Strong and weak scaling on Piz Daint.



Other Results

- SparCML: a sparse reduction protocol to implement faster reductions in parallel systems with sparse input vectors [arXiv'18, NIPS'18]
- Using deep learning to create learnable representations of code [NIPS'18]
 - State of the art in predicting fastest hardware mapping and algorithm classification
- Accelerating convolution operators using micro-batches [Cluster'18]
 - Key technique: Use ILP and Dynamic Programming
- Parallelism modeling of deep learning, from operator to distributed training on supercomputers [arXiv'18]

١.,	Neural Code Comprehension: A Learnable Representation of Code Semantics
I	Accelerating Deep Learning Frameworks with Micro-batches
A k isar he he lla equ see xe xe xe iiv olu	 A Marking Marking Market Market
Ц	are community and even academia. prov wo As datasets increase in size and DNNs in complexity, the computational intensity and memory demands of deep learning increase proportionally. Training a DNN to competitive accuracy today