Accelerating and Expanding End-to-End Data Science Workflows with DL/ML Interoperability Using RAPIDS

Bartley Richardson, PhD
NVIDIA
Santa Clara, CA, USA
brichardson@nvidia.com

Bradley Rees, PhD
NVIDIA
Santa Clara, CA, USA
brees@nvidia.com

Tom Drabas, PhD
Microsoft
Redmond, WA, USA
todrabas@microsoft.com

Even Oldridge
NVIDIA
Santa Clara, CA, USA
eoldridge@nvidia.com

David A. Bader, PhD
Institute for Data Science, NJIT
Newark, NJ, USA
david.bader@njit.edu

Rachel Allen, PhD
NVIDIA
Santa Clara, CA, USA
rachela@nvidia.com

ABSTRACT

The lines between data science (DS), machine learning (ML), deep learning (DL), and data mining continue to be blurred and removed. This is great as it ushers in vast amounts of capabilities, but it brings increased complexity and a vast number of tools/techniques. It's not uncommon for DL engineers to use one set of tools for data extraction/cleaning and then pivot to another library for training their models. After training and inference, it's common to then move data yet again by another set of tools for post-processing. The RAPIDS suite of open source libraries not only provides a method to execute and accelerate these tasks using GPUs with familiar APIs, but it also provides interoperability with the broader open source community and DL tools while removing unnecessary serializations that slow down workflows. GPUs provide massive parallelization that DL has leveraged for some time, and RAPIDS provides the missing pieces that extend this computing power to more traditional yet important DS and ML tasks (e.g., ETL, modeling). Complete pipelines can be built that encompass everything, including ETL, feature engineering, ML/DL modeling, inference, and visualization, all while removing typical serialization costs and affording seamless interoperability between libraries. All experiments using RAPIDS can effortlessly be scheduled, logged and reviewed using existing public cloud options. Join our engineers and data scientists as they walk through a collection of DS and ML/DL engineering problems that show how RAPIDS running on Azure ML can be used for end-to-end, entirely GPU pipelines. This tutorial includes specifics on how to use RAPIDS for feature engineering, interoperability with common ML/DL packages, and creating GPU native visualizations using cuxfilter. The use cases presented here give attendees a hands-on approach to using RAPIDS components as part of a larger workflow, seamlessly integrating with other libraries (e.g., TensorFlow) and visualization packages.

KDD'20, August 23-27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406702

CCS CONCEPTS

• Computing methodologies ~ Machine learning ~ Machine learning algorithms • Computer systems organization ~ Architectures ~ Other architectures ~ Neural networks • General and reference ~ Document types ~ Surveys and overviews. • General and reference ~ Cross-computing tools and techniques ~ Performance

KEYWORDS

Data science, GPU acceleration, machine learning, deep learning, PyData ecosystem

ACM Reference format:

Bartley Richardson, Bradley Rees, Tom Drabas, David Bader, Even Oldridge, and Rachel Allen. 2020. "Accelerating and Expanding End-to-End Data Science Workflows with DL/ML Interoperability Using RAPIDS." In *Proceedings of 2020 ACM Conference of Knowledge Discovery and Data Mining (KDD'20), August 23-27, Virtual Event.* ACM, New York, NY, USA.

1 TUTORIAL DESCRIPTION

1.1 Duration and Format

Three hours, hands-on.

1.2 Target Audience and Aims

The target audience for this tutorial are Data Scientists, ML Engineers, DL Engineers, Data Miners, Data Engineers, Data Researchers, and general programmers who are interested in exploring how accelerated interoperability between common data science/engineering platforms leads to faster, more efficient pipelines as well as faster research iterations. A general

knowledge of Python is required as is a general understanding of machine learning. An understanding of and familiarity with statistics, deep learning, PyTorch, and graph analysis is helpful, but it is not required.

1.3 Related Tutorial History

KDD 2019, PyData Miami 2019, KDD 2018

1.2 Outline

The tutorial will be presented as a wide collection of data science problems that intersect the various component libraries of RAPIDS as well as the larger PyData ecosystem. The various APIs and syntaxes will be discussed, but the focus is on how to construct accelerated data science workflows that utilize both RAPIDS as well as other common GPU-accelerated, open source libraries. All workflows below address real-world problems on large and domain-accurate datasets. Real data is used whenever possible, and authentic fake data is generated where PII concerns prevent or limit the use of real datasets.

1. Introduction (not hands-on)
2. Tutorial (hands-on)
    a. Deep Learning for Tabular Data (use case)
    b. Log Parsing using Neural Networks and a Language Based Model (use case)
    c. ML, Graph Analysis, and Visualization (use case)
3. Conclusions (not hands-on)

2 BRIEF BIOGRAPHIES OF TUTORS

Bartley Richardson, PhD is a Senior Cybersecurity Data Scientist and AI Infrastructure Manager (RAPIDS) at NVIDIA. He leads a team that researches and applies GPU-accelerated ML and DL to help solve today's information security and cybersecurity challenges. Previously, Bartley was a technical lead and performer on multiple DARPA research projects where he applied data science and ML/DL algorithms at scale to address large cybersecurity problems. His primary research areas are NLP and sequence-based methods for cyber network defense. Bartley holds a PhD in CS and CompE with a focus on unstructured logical query optimization. His BS is in Computer Engineering with a concentration in software design and AI.

Brad Rees, PhD is a Senior Manager at NVIDIA and lead of the RAPIDS cuGraph team. Brad has been designing, implementing, and supporting a variety of advanced software and hardware systems within the defense and research communities for over 30 years,specializing in complex analytic systems, primarily using graph analytic techniques for social and cyber network analysis. Brad has a Ph.D. in Computer Science with a focus on graph analytics.

Tom Drabas, PhD is a Senior Data Scientist at Microsoft in the Azure Machine Learning group. His research interests include parallel computing, deep learning, and ML algorithms and their applications. During his time at Microsoft, Tom has published multiple books and authored a video series on data science, machine learning, and distributed computing in Spark. He has over 17 years of international experience working in the airline, telecommunication and technology industries. Tom holds a Ph.D. in the Airline Operations Research field from the University of New South Wales.

David A. Bader, PhD is a Distinguished Professor in the Department of Computer Science at New Jersey Institute of Technology. Prior to this, he served as founding Professor and Chair of the School of Computational Science and Engineering, College of Computing, at Georgia Institute of Technology. He is a Fellow of the IEEE, AAAS, and SIAM, and advises the White House, most recently on the National Strategic Computing Initiative (NSCI). Dr. Bader is a leading expert in solving global grand challenges in science, engineering, computing, and data science. His interests are at the intersection of high-performance computing and real-world applications, including cybersecurity, massive-scale analytics, and computational genomics, and he has co-authored over 250 scholarly papers.

ACKNOWLEDGMENTS

Thanks to Even Oldridge (NVIDIA), Rachel Allen (NVIDIA), Corey Nolet (NVIDIA), Julio Perez (NVIDIA), and Alec Gunny (NVIDIA) for their work on the notebooks presented. Special thanks to Joshua Patterson (NVIDIA), Keith Kraus (NVIDIA), and the entire RAPIDS team, ecosystem, and contributors.

REFERENCES

[1] S. Rabhi, W. Sun, J. Perez, M. R. Kristensen, J. Liu, and E. Oldridge, 2019. "Accelerating recommender system training 15x with RAPIDS." In *Proceedings of the Workshop on ACM Recommender System Challenge (RecSys Challenge '19)*. ACM Press, New York, NY, 1-5. DOI: https://doi.org/10.1145/3359555.3359564

[2] W. McKinney, 2010. "Data structures for statistical computing in Python." In *Proceedings of the 9th Python In Science Conference (SCIPY 2010)*. Vol. 445, pp. 51-56

[3] L. McInnes, J. Healy, and J. Melville, 2018. "UMAP: Uniform manifold approximation and projection for dimension reduction." https://arxiv.org/abs/1802.03426

[4] V. Poulin and F. Théberge, 2018. "Ensemble clustering for graphs." In *International Conference on Complex Networks and their Applications*. pp. 231-243.

[5] A. Tripathy, F. Hohman, D. Chau, and O. Green, 2018. "Scalable K-core decomposition for static graphs using a dynamic graph data structure." In *IEEE Proceedings of the International Conference on Big Data (BIG DATA)*. pp. 1134-1141.

[6] RAPIDS, https://rapids.ai

[7] cyBERT, https://medium.com/rapids-ai/cybert-28b35a4c81c4

[8] DL recommender systems, https://medium.com/rapids-ai/accelerating-deep-learning-recommender-systems-by-15x-using-rapids-fastai-and-pytorch-b50b4d8568d1