BOSS
SHOW YOUR DATA WHO'S BOSS

Shining the Light on Dark Data

Scaling Artificial intelligence with BOSS

Author

David Bauer Jr., Ph.D.

Reviewer

Joel Branch, Ph.D.

Author
David Bauer Jr., Ph.D.

Reviewer
Joel Branch, Ph.D.

Introduction

A cursory glance of the market landscape in big data, cyber security and Fintech might leave the uninitiated with the impression that most firms have machine learning and artificial intelligence mastered. Unfortunately, this is far from the case. Many firms have learned how to make incremental enhancements to their solutions with limited machine learning capabilities. However, few possess the skill, scale, or intellectual property to master artificial intelligence as a tool to solve real world problems, including enabling intelligent process automation and effective decision support using "dark" unstructured, semi-structured and structured data.

At BOSS, we are building BOSS AI-as-a-Platfom, an end-to-end artificial intelligence SaaS and licensable ente prise platform. BOSS is designed to run anywhere and at any scale, enabling dynamic deep learning for organizations of all sizes, such as small businesses, global enterprises, and hyper-scale service providers and government. Our initial focus areas are improving intelligence gathering and analysis capabilities for the intelligence community and tackling challenging scenarios in the financial services and healthcare industries.

"Replacing RPC with BOSS DOS™ for TensorFlow enables us to bring together high performance computing and all the functions of the most widely adopted library for deep learning."

The goal of this series of articles is to provide the reader with insight as to our thoughts and concepts and to raise awareness of our activities and aspirations among industry at large. Each article will focus on a specific topic: scaling artificial intelligence, enabling dynamic learning, and providing an end-to-end platform capability. We welcome discussion and questions over the materials and concepts contained in the white papers and desire to facilitate a robust conversation around these topics.

Scaling Artificial Intelligence

A growing approach to improving the accuracy and precision of deep learning models (e.g., C-NNs, R-NNs) often translates to employing wider and deeper layers of computation for larger and more complex data sets. Reaching this large scale of analysis is a non-trivial task and involves distributing the computation over multiple computing devices while managing the communications between them efficiently.

In this article, we outline our approach for scaling deep learning models using BOSS's Distributed Operating System, BOSS DOS™, a distributed compute engine. DOS has been tested on 2 million CPU cores on Lawrence Livermore National Lab's Sequoia Blue Gene/Q supercomputer with a parallel efficiency of 93% [2]. With BOSS DOS, we can achieve super-linear speedups which furthermore provide the efficiencies as much lower core GPU or CPU counts.

Author
David Bauer Jr., Ph.D.

Reviewer
Joel Branch, Ph.D.

Background / Problems

To minimize training time, the training of a deep neural network must be scaled beyond a single machine to as many machines as possible by distributing the optimization method used for training.

Training time on large datasets for deep neural networks is the principal workflow bottleneck in several important applications of deep learning, such as object classification and detection in automatic driver assistance systems. To minimize training time, the training of a deep neural network must be scaled beyond a single machine to as many machines as possible by distributing the optimization method used for training. It has also been observed that increasing the scale of deep learning, with respect to the number of training examples, the number of model parameters, or both, can drastically improve ultimate classification accuracy [1][2][3]. These results have led to a surge of interest in scaling up the training and inference algorithms used for these models [4] and in improving applicable optimization procedures [3] [5]. It has also been observed that increasing the scale of deep learning, with respect to the number of training examples, the number of model parameters, or both, can drastically improve ultimate classification accuracy [1][2][3]. These results have led to a surge of interest in scaling up the training and inference algorithms used for these models [4] and in improving applicable optimization procedures [3][5]

Scaling Artificial Intelligence

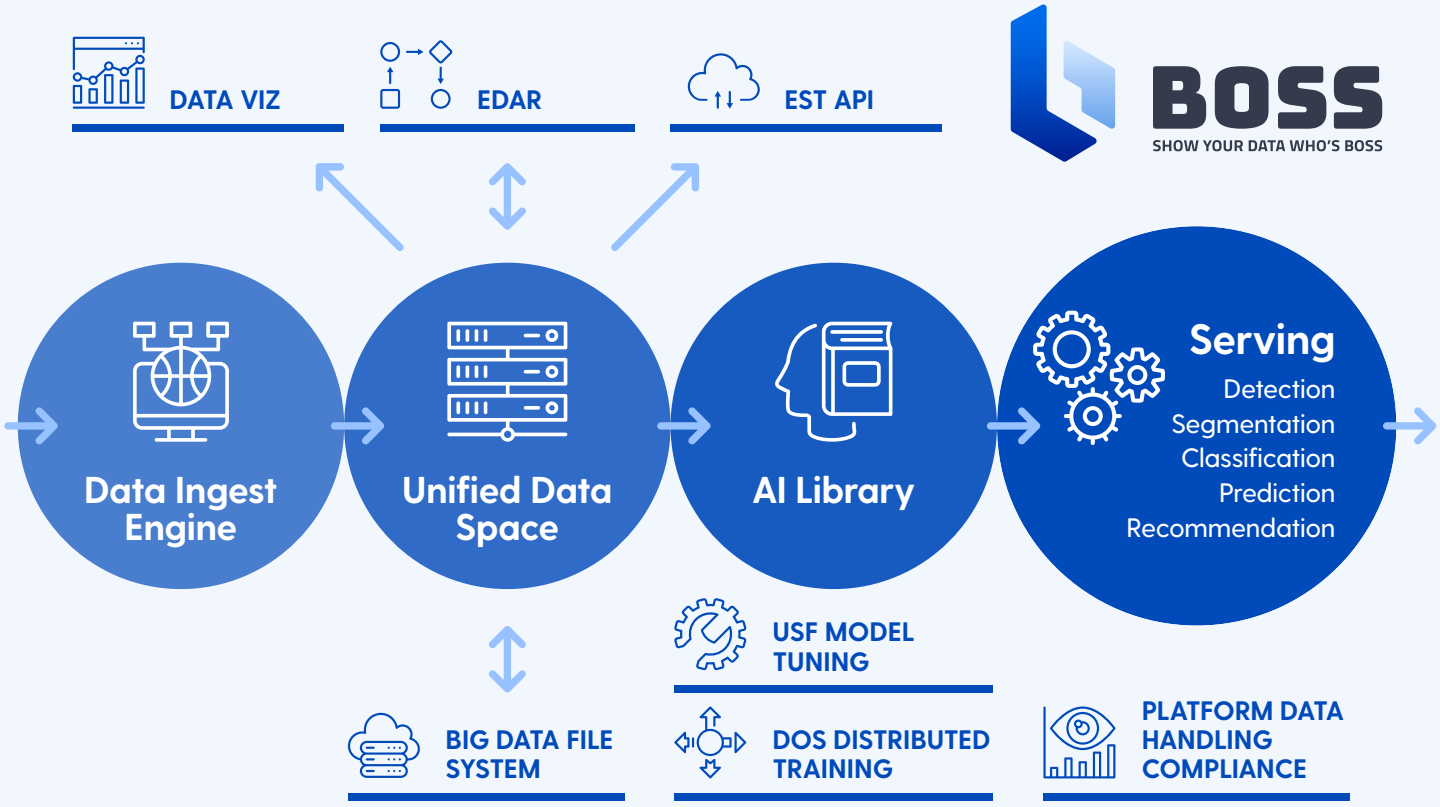
BOSS provides massive scaling across most high performance and supercomputing platforms available today, including all the available cloud hosting providers such as Amazon Web Services and Microsoft Azure. Developed over the past 17 years, BOSS has been tested on HPC clusters comprised of commodity HP, Dell, and Rackspace servers, as well as the IBM Blue Gene family and Cray XT series supercomputers. Our solution is developed as an ANSI C library for distributed computing and leverages either MPICH or TCP sockets for communications.

"Through the application of, BOSS DOS, our massively scalable distributed computing engine, we can provide the required scalability to generate higher quality models."

Author
David Bauer Jr., Ph.D.

Reviewer
Joel Branch, Ph.D.

One of our goals is to leverage existing deep learning models developed in toolkits such as Google's TensorFlow. TensorFlow was developed with a degree of modularity. It implements an event processing framework for distributed execution over Remote Procedure Calls (RPC). If one would like to scale beyond the limits of RPC, then one can seamlessly migrate code for large-scale execution into the Google Cloud. From a proprietary perspective, this represents a strong tie-in to Google's commercial offerings. We aim to disrupt this pipeline dependency by providing the ability to execute TensorFlow models on any public cloud, private cloud, or bare metal infrastructure platform of the user's liking.



BOSS DOS is provided as a replacement distributed computing engine for Google's TensorFlow, which enables efficient parallelization of both TensorFlow and Keras deep learning models. We are also currently developing native implementations of these libraries directly within Lucd to facilitate the power and speed of native execution. Replacing RPC with BOSS DOS for TensorFlow enables us to bring together high performance computing and all the functions of the most widely adopted library for deep learning. TensorFlow's distributed computation model is based on communicating events between servers at well-defined execution intervals to ensure serial compatibility. BOSS DOS is a fully distributed compute engine which is also based on event passing between what we define as logical processes or LPs. A logical process is the fundamental unit of decomposing the model state across multiple GPU or CPU cores. Execution in BOSS is then realized as the execution of multiple TensorFlow models communicating data for computation

Author
David Bauer Jr., Ph.D.

Reviewer
Joel Branch, Ph.D.

to create deep learning models of a nearly unlimited size in terms of both depth and width. Data decomposition occurs naturally through segmentation of the distributed test and training data sets, and not within individual samples. Our initial focus is on high throughput for large-scale data sets first, and then we will take the problem of decomposing large individual records for processing later. Therefore, the assumption is that all individual records are at least small enough to execute within the context of a single machine.

BOSS DOS enables TensorFlow distributed communications and event processing leveraging the discrete event paradigm. Then TensorFlow messages are sent between processor cores within a machine and between machines using our "optimistic execution" message passing paradigm.

Through the application of, BOSS DOS, our massively scalable distributed computing engine we can provide the required scalability to generate higher quality models. Our approach provides the additional benefit of algorithm and data portability across all computing platforms, rather than being tied to a single, proprietary system.

Using BOSS, businesses of all shapes and sizes can use their existing algorithms where their data is currently, and avoid becoming locked in to any one proprietary, vendor solution.

Conclusion

Through integration of, TensorFlow, the Deep Learning community's most prevalent library and BOSS DOS the world's most scalable distributing computing engine, BOSS's capability enables organizations to develop high quality deep learning models that are both wider and deeper than anything available in the commercial or open source communities. Developing high quality models is critical for solving real world problems, such has heart stent lifecycle management in the healthcare field or trade reconstruction in the Financial Services industry along with broad based management and operational use cases, and BOSS can ensure that Deep Learning becomes a valuable tool in your decision support and intelligent process automation toolbox.

Works Cited

[1] U. M. L. M. G. a. J. S. D. C. Ciresan, "Deep big simple neural nets excel on handwritten digit recognition.," in CoRR, 2010.
[2] H. L. a. A. Y. N. A. Coates, "An analysis of single-layer networks in unsupervised feature learning.," in AISTATS 14, 2011.
[3] J. N. A. C. A. L. B. P. a. A. N. Q.V. Le, "On optimization methods for deep learning.," 2011.
[4] A. M. a. A. Y. N. R. Raina, "Large-scale deep unsupervised learning using graphics processors.," ICML, 2009.
[5] J. Martens., "Deep learning via hessian-free optimization.," ICML, 2010.