SingaProposal

---- ⚠️ **FINAL** ⚠️ This proposal is now complete and has been submitted for a VOTE. ----

Singa Incubator Proposal

Abstract

SINGA is a distributed deep learning platform.

Proposal

SINGA is an efficient, scalable and easy-to-use distributed platform for training deep learning models, e.g., Deep Convolutional Neural Network and Deep Belief Network. It parallelizes the computation (i.e., training) onto a cluster of nodes by distributing the training data and model automatically to speed up the training. Built-in training algorithms like Back-Propagation and Contrastive Divergence are implemented based on common abstractions of deep learning models. Users can train their own deep learning models by simply customizing these abstractions like implementing the Mapper and Reducer in Hadoop.

Background

Deep learning refers to a set of feature (or representation) learning models that consist of multiple (non-linear) layers, where different layers learn different levels of abstractions (representations) of the raw input data. Larger (in terms of model parameters) and deeper (in terms of number of layers) models have shown better performance, e.g., lower image classification error in Large Scale Visual Recognition Challenge. However, a larger model requires more memory and larger training data to reduce over-fitting. Complex numeric operations make the training computation intensive. In practice, training large deep learning models takes weeks or months on a single node (even with GPU).

Rational

Deep learning has gained a lot of attraction in both academia and industry due to its success in a wide range of areas such as computer vision and speech recognition. However, training of such models is computationally expensive, especially for large and deep models (e.g., with billions of parameters and more than 10 layers). Both Google and Microsoft have developed distributed deep learning systems to make the training more efficient by distributing the computations within a cluster of nodes. However, these systems are closed source softwares. Our goal is to leverage the community of open source developers to make SINGA efficient, scalable and easy to use. SINGA is a full fledged distributed platform, that could benefit the community and also benefit from the community in their involvement in contributing to the further work in this area. We believe the nature of SINGA and our visions for the system fit naturally to Apache's philosophy and development framework.

Initial Goals

We have developed a system for SINGA running on a commodity computer cluster. The initial goals include,

- improving the system in terms of scalability and efficiency, e.g., using Infiniband for network communication and multi-threading for one node computation. We would consider extending SINGA to GPU clusters later.
- benchmarking with larger datasets (hundreds of millions of training instances) and models (billions of parameters).
- adding more built-in deep learning models. Users can train the built-in models on their datasets directly.

Current Status

Meritocracy

We would like to follow ASF meritocratic principles to encourage more developers to contribute in this project. We know that only active and excellent developers can make SINGA a successful project. The committer list and PMC will be updated based on developers' performance and commitment. We are also improving the documentation and code to help new developers get started quickly.

Community

SINGA is currently being developed in the Database System Research Lab at the National University of Singapore (NUS) in collaboration with Zhejiang University in China. Our lab has extensive experience in building database related systems, including distributed systems. Six PhD students and research assistants (Jinyang Gao, Kaiping Zheng, Sheng Wang, Wei Wang, Zhaojing Luo and Zhongle Xie) , a research fellow (Anh Dinh) and three professors (Beng Chin Ooi, Gang Chen, Kian Lee Tan) have been working for a year on this project. We are open to recruiting more developers from diverse backgrounds.

Core Developers

Beng Chin Ooi, Gang Chen and Kian Lee Tan are professors who have worked on distributed systems for more than 20 years. They have collaborated with the industry and have built various large scale systems. Anh Dinh's research is also on distributed systems, albeit with more focus on security aspects. Wei Wang's research is on deep learning problems including deep learning applications and large scale training. Sheng Wang and Jinyang are working on efficient indexing, querying of large scale data and machine learning. Kaiping, Zhaojing and Zhongle are new PhD students who jointed SINGA recently. They will work on this project for a longer time (next 4-5 years). While we share common research interests, each member also brings diverse expertise to the team.

Alignment

ASF is already the home of many distributed platforms, e.g., Hadoop, Spark and Mahout, each of which targets a different application domain. SINGA, being a distributed platform for large-scale deep learning, focuses on another important domain for which there still lacks a robust and scalable open-source platform. The recent success of deep learning models especially for vision and speech recognition tasks has generated interests in both applying existing deep learning models and in developing new ones. Thus, an open-source platform for deep learning will be able to attract a large community of users and developers. SINGA is a complex system needing many iterations of design, implementation and testing. Apache's collaboration framework which encourages active contribution from developers will inevitably help improve the quality of the system, as shown in the success of Hadoop, Spark, etc.. Equally important is the community of users which helps identify real-life applications of deep learning, and helps to evaluate the system's performance and ease-of-use. We hope to leverage ASF for coordinating and promoting both communities, and in return benefit the communities with another useful tool.

Known Risks

Orphaned products

Four core developers (Anh, Wei Wang, Jinyang and Sheng Wang) may leave the lab in two to four years time. It is possible that some of them may not have enough time to focus on this project after that. But, SINGA is part of our other bigger research projects on building an infrastructure for data intensive applications, which include health-care analytics and brain-inspired computing. Beng Chin and Kian Lee would continue working on it and getting more people involved. For example, three new developers (Kaiping, Zhaojing and Zhongle) joined us recently. Individual developers are welcome to make SINGA a diverse community that is robust and independent from any single developer.

Inexperience with Open Source

All the developers are active users and followers of open source projects. Our research lab has a strong commitment to open source, and has released the source code of several systems under open source license as a way of contributing back to the open source community. But we do not have much real experience in open source projects with large and well organized communities like those in Apache. This is one reason we choose Apache which is experienced in open source project incubation. We hope to get the help from Apache (e.g., champion and mentors) to establish a healthy path for SINGA.

Homogenous Developers

Although the current developers are researchers in the universities, they have different research interests and project experiences, as mentioned in the section that introduces the core developers. We know that a diverse community is helpful. Hence we are open to the idea of recruiting developers from other regions and organizations.

Reliance on Salaried Developers

As a research project in the university, SINGA's current developing community consists of professors, PhD students, research assistants and postdoctoral fellows. They are driven by their interests to work on this project and have contributed actively since the start of the project. The research assistants and fellows are expected to leave when their contracts expire. However, they are keen to continue to work on the project voluntarily. Moreover, as a long term research project, new research assistants and fellows are likely to join the project.

A Excessive Fascination with the Apache Brand

We choose Apache not for publicity. We have two purposes. First, we want to leverage Apache's reputation to recruit more developers to make a diverse community. Second, we hope that Apache can help us to establish a healthy path in developing SINGA. Beng Chin and Kian-Lee are established database and distributed system researchers, and together with the other contributors, they sincerely believe that there is a need for a widely accepted open source distributed deep learning platform. The field of deep learning is still at its infancy, and an open source platform will fuel the research in the area. Moreover, such a platform will enable researchers to develop new models and algorithms, rather than spending time implementing a deep learning system from scratch. Furthermore, the need for scalability for such a platform is obvious.

Relationship with Other Apache Products

Apache Mahout and Apache Spark's ML-LIB are general machine learning systems. Deep learning algorithm can thus be implemented on these two platforms as well. However, the there are differences in training efficiency, scalability and usability. Mahout and Spark ML-LIB follow models where their nodes run synchronously. This is the fundamental difference to Singa who follows the parameter server framework (like Google Brain and Microsoft Adam). Singa can run synchronously or asynchronously. The asynchronous mode is superior than the synchronous mode in terms of scalability. In addition, Singa has some optimizations towards deep learning models (e.g., model parallelism, data parallelism and hybrid-parallelism) which make Singa more efficient. We also provide ease of use programming model for deep learning algorithms.

There are also plans for integration with Apache Hadoop's HDFS as storage, to handle large training data. Specifically, we store the training data (e.g., images or raw features of images) in HDFS, then (pre-)fetch them online. We will also explore integration with Hadoop's Yarn and Apache Mesos to do resource management.

Documentation

The project is hosted at http://www.comp.nus.edu.sg/~dbsystem/project/singa.html. Documentations can be found at the Github Wiki Page: https://github.com/nusinga/singa/wiki. We continue to refine and improve the documentation.

Initial Source

We use Github to maintain our source code, https://github.com/nusinga/singa

Source and Intellectual Property Submission Plan

We plan to make our code base be under Apache License, Version 2.0.

External Dependencies

- required by the core code base: glog, gflags, google protobuf, open-blas, mpich, armci-mpi.
- required by data preparation and preprocessing: opencv, hdfs, python.

Cryptography

Not Applicable

Required Resources

Mailing Lists

Currently, we use google group for internal discussion. The mailing address is nusinga@googlegroup.com. We will migrate the content to the apache mailing lists in the future.

- singa-dev
- singa-user
- singa-commits
- singa-private (for private discussion within PCM)

Git Repository

We want to continue using git for version control. Hence, a git repo is required.

Issue Tracking

JIRA Singa (SINGA)

Initial Committers

- Beng Chin Ooi (ooibc @comp.nus.edu.sg)
- Kian Lee Tan (tankl @comp.nus.edu.sg)
- Gang Chen (cg @zju.edu.cn)
- Wei Wang (wangwei @comp.nus.edu.sg)
- Dinh Tien Tuan Anh (dinhtta @comp.nus.edu.sg)
- Jinyang Gao (jinyang.gao @comp.nus.edu.sg)
- Sheng Wang (wangsh @comp.nus.edu.sg)
- Kaiping Zheng (kaiping @comp.nus.edu.sg)
- Zhaojing Luo (zhaojing @comp.nus.edu.sg)
- Zhongle Xie (zhongle @comp.nus.edu.sg)

Affiliations

- Beng Chin Ooi, National University of Singapore
- Kian Lee Tan, National University of Singapore
- Gang Chen, Zhejiang University
- Wei Wang, National University of Singapore
- Dinh Tien Tuan Anh, National University of Singapore
- Jinyang Gao, National University of Singapore
- Sheng Wang, National University of Singapore
- Kaiping Zheng, National University of Singapore
- Zhaojing Luo, National University of Singapore
- Zhongle Xie, National University of Singapore

Sponsors

Champion

Thejas Nair (thejas at apache.org)

Nominated Mentors

- Thejas Nair (thejas at apache.org)
- Alan Gates (gates at apache dot org)
- Daniel Dai (daijy at apache dot org)
- Ted Dunning (tdunning at apache dot org)

Sponsoring Entity

We are requesting the Incubator to sponsor this project.