BUILDING DEEP LEARNING POWERED BIG DATA ANALYTICS USING BIGDL

Radhika Rangarajan

Big Data Technology Team, Software and Service Group, Intel
BigDL is a distributed deep learning library for Apache Spark*
BRING DEEP LEARNING TO BIG DATA PLATFORM

- Open Sourced Deep Learning Library for Apache Spark*
- Make Deep learning more Accessible to Big data users and data scientists.
- Feature Parity with popular DL frameworks
- Easy Customer and Developer Experience
- High Performance
- Efficient Scale out
WHY?
MOTIVATION FOR BIGDL
DATA SCALE DRIVING DEEP LEARNING PROCESS

Andrew NG, Baidu. NIPS 2016 Tutorial
PRODUCTION ML/DL SYSTEM IS COMPLEX

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

“Hidden Technical Debt in Machine Learning Systems”,
Google, NIPS 2015 Paper
UNIFIED BIG DATA PLATFORM Driving Analytics and Data Science
How to Run **Deep Learning** Workloads Directly on **Big Data** Platform?

- Integrated with Big Data ecosystem
- Massively distributed, shared-nothing
- Scale-out
- Send compute to data
- Fault tolerance
- Elasticity
- Incremental scaling
- Dynamic resource sharing
- ...
BIGDL – IT IS A BIG DEAL!

Open sourced on Dec 30, 2016

1750+ stars, 330+ forks and growing strong...

https://github.com/intel-analytics/BigDL

https://software.intel.com/BigDL
WHERE?

USAGE AND EXAMPLES OF BIGDL
WHERE CAN I USE BIGDL?

- Analyze “big data” using deep learning on the same Hadoop/Spark cluster where the data are stored
- Add deep learning functionalities to the Big Data (Spark) programs and/or workflow
- Leverage existing Hadoop/Spark clusters to run deep learning applications
- Making deep learning more accessible for Big Data users and Data Scientists.
Fraud transaction detection is very important to finance companies. A good fraud detection solution can save a lot of money.

ML solution challenge
- Data cleaning
- Feature engineering
- Unbalanced data
- Hyper parameter
FRAUD TRANSACTION DETECTION

- History data is stored on Hive
- Data preprocess/cleaning with Spark-SQL
- Spark ML pipeline for complex feature engineering
- Under sample + Bagging solve unbalance problem
- Grid search for hyper parameter tuning

Powered by BigDL
PRODUCT DEFECT DETECTION AND CLASSIFICATION

Data source

- Cameras installed on manufacturing pipeline

Task

- Detect defect from the photos
- Classify the defect
PRODUCT DEFECT DETECTION AND CLASSIFICATION

Big Data management
- High resolution images
- Large volume of data

Proposal Extraction
- Extract proposals for each defect
- Parallel pipeline (KeyStone ML Pipeline)
- Running on Spark

Preprocessing
- Preprocess the proposals before model training or testing

Model training pipeline
- Train Convolutional Neural Networks on Spark
- Parameter tuning, optimize proposal extraction algorithms
LANGUAGE MODEL WITH RNN

- Sentence Tokenizer
- Dictionary Building
- Input Document Transformer

Generated sentences with regard to trigger words.
Learn from Shakespeare Poems

Output of RNN:

Long live the King. The King and Queen, and the Strange of the Veils of the rhapsodic. and grapple, and the entreatments of the pressure.

Upon her head, and in the world? ``Oh, the gods! O Jove! To whom the king: ``O friends!

Her hair, nor loose! If, my lord, and the groundlings of the skies. jocund and Tasso in the Staggering of the Mankind. and
MORE RNN SUPPORT: LSTM

The repeating module in an LSTM contains four interacting layers.

Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

BigDL also supports LSTM variants such as GRU and LSTM with peepholes.
FINE-TUNE CAFFE/TORCH MODEL ON SPARK

- Train on different dataset based on pre-trained model
- Predict image style instead of type
- Save training time and improve accuracy

Image source: https://www.flickr.com/photos/
VISUAL RECOGNITION AND OBJECT DETECTION

Faster-RCNN

SSD: Single Shot MultiBox Detector
OBJECT DETECTION ON PASCAL

(http://host.robots.ox.ac.uk/pascal/VOC/)
More BigDL Examples

BigDL provide examples to help developer play with bigdl and start with popular models.

https://github.com/intel-analytics/BigDL/wiki/Examples

Models (Train and Inference example code):

- LeNet, Inception, VGG, ResNet, RNN, Auto-encoder

Examples:

- Text Classification
- Image Classification
- Load Torch/Caffe model
HOW?

GET STARTED WITH BIGDL
BigDL On Github

https://github.com/intel-analytics/BigDL

Community

Mail List

bigdl-user-group+subscribe@googlegroups.com

Report bugs and feature request

https://github.com/intel-analytics/BigDL/issues
BigDL in Cloudera Data Science Workbench

Self-service data science for the enterprise

WHAT ELSE?
WHAT'S NEW WITH BIGDL?

0.2 coming out EOQ2

- 3D convolutions
- Tensorflow model read/write (initial)
- Recursive net support
- Spark 2.1 support
- Python 3.5 support etc
Unleash Your Potential with Intel’s Complete AI Portfolio
## Intel® AI Portfolio

### Experiences
- Intel® DL Training & Deployment
- Intel® Nervana™ DL Software & Cloud
- Intel® Computer Vision SDK
- Intel® GO™ Automotive SDK
- Movidius Fathom

### Toolkits
- Intel® DL, Spark ML, BeDL, TensorFlow
- MXNet, Theano, Torch, Caffe, Chainer

### Frameworks
- Intel® Nervana™ Graph*
- Intel® MKL, MKL-DNN, Intel® MLSL

### Libraries
- Python
- Intel® DAAL
- Intel Distribution

### Hardware
- Compute
- Memory/Storage
- Networking
- Computer Vision

---

**AI on Intel: Unleashing the Next Wave**
1. **Intel® Developer Zone for AI:** Online community to reach 1.5M developers, data scientists and students with frameworks, tools, trainings and support

2. **Training:** Partnership with Coursera to deliver series of online AI courses starting Q1 '17 plus 100 AI workshops, webinars, meet-ups world wide

3. **Competitions:** Intel has partnered with MobileODT, a company dedicated to enabling cervical cancer screening for every woman, everywhere, to host an AI Competition on Kaggle, starting February 2017

4. **Student Developers:** Intel Student Ambassadors leading AI Clubs and Workshops at universities world wide
PARTNER WITH INTEL

- Use BigDL and Share your Experience
- Use Intel Optimized Libraries
- Leverage Intel Developer Zone Resources

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A “Mission Critical Application” is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Intel, Quark, VTune, Xeon, Cilk, Atom, Look Inside and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Copyright ©2015 Intel Corporation.
Risk Factors

The above statements and any others in this document that refer to plans and expectations for the first quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company's expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; customer acceptance of Intel's and competitors' products; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel's response to such actions; and Intel's ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel's results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.