Large Scale Distributed Machine Learning on Apache Spark

SSG, Big Data Technology

Yuhao.Yang@intel.com

Fraud Detection on Apache Spark

- Goal:
  - Given card transaction details, classify whether the transaction is fraud or normal.
- Data overview:
  - Billions of transaction records. (unbalanced)
  - Hundreds of raw features for each record (flag, categorical and numeric features).
  - Training dataset
  - Test Dataset
- Evaluation Matrices
  - Recall = predicted fraud / all real fraud transaction.
  - Precision = predicted fraud correctly / predicted fraud

Fraud can mean:

Buying with stolen credit cards

Abusing promotional programs

Account takeover

Spamming other users

Feature Engineering 1. understanding Data

- Statistical analysis for the raw features
  - categorical features => Distinct, group by count, etc.
  - for numeric features => Min, max, avg, variance, etc.
  - feature distribution
- Feature Generation
  - We need to derive more logical features
  - Reflect location change, abnormal amount, high risk region…
- Data verification and cleaning

Model Training 1. choosing algorithm

- Algorithm comparison:
  - Neural network vs. GBDT(Gradient Boosting Decision Tree)

- Bootstrap aggregating (bagging)
  - Single model reaches limit after certain rounds of tuning.
  - Reduce problems related to over-fitting
  - Allow better performance and flexibility during training

Some experiences and next step

- Experience:
  - Speed up the feedback cycle: The key in model tuning is to shorter the feedback cycle between training and evaluation. Automate the process wherever possible, by leveraging the pipeline, model persistence and grid search system.
  - Transform features according to its properties: Features should be carefully evaluated. Categorical features need to be handled differently according to their distribution and number of distinct values.
  - Automatic feature selection can be too expensive. Leverage Information value to narrow down the scope.

Topic Modeling

- Automatically infers the topics discussed in a collection of documents.
- These topics can be used to summarize and organize documents, or used for featurization and dimensionality reduction
  - What is document X discussing?
  - How similar are documents X and Y?
  - If I am interested in topic Z, which documents should I read first?
- Widely applied
  - Document, image clustering
  - feature deduction
  - Social network, advertising …

Intuition of LDA

Corpus

doc1. apple banana

doc2. apple orange

doc3. banana orange

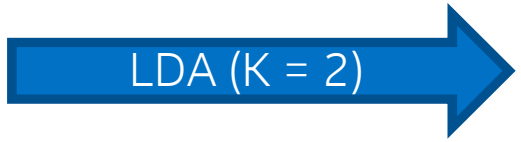
doc4. tiger cat

doc5. tiger dog

doc6. cat dog

doc7. Cat dog apple

Input (what we have)



	Topic 1	Topic2
Apple	33%	0%
Banana	33%	0%
Orange	33%	0%
Tiger	0%	33%
Cat	0%	33%
Dog	0%	33%

	Topic 1	Topic2
doc1	100%	0%
doc2	100%	0%
doc3	100%	0%
doc4	0%	100%
doc5	0%	100%
doc6	0%	100%
doc7	33%	66%

Output (what we want)

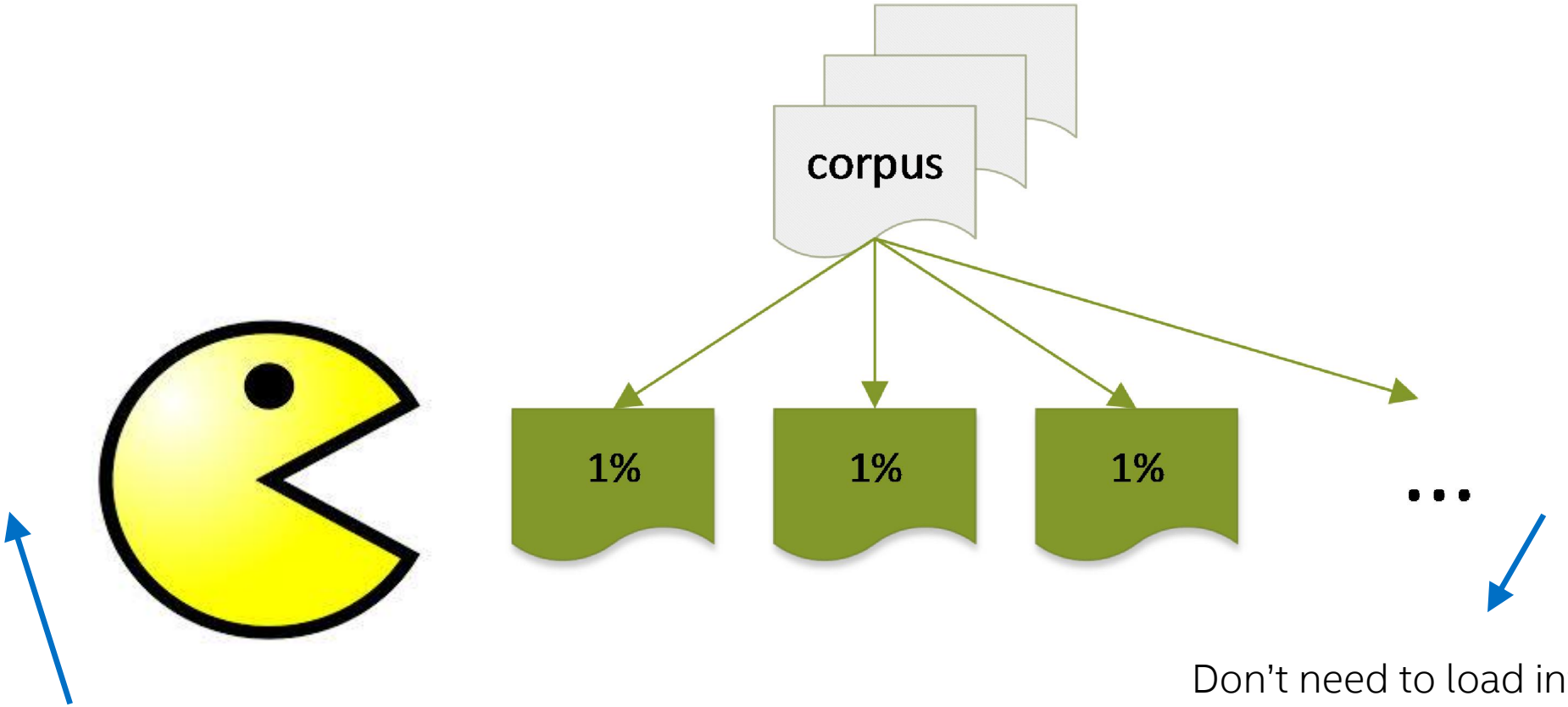
Existing Solution

- Variational EM (Expectation-maximization)
  - Numerical approximation using lower-bounds
  - Results in biased solutions
  - Convergence has numerical guarantees
- Gibss Sampling
  - Stochastic simulation
  - unbiased solutions
  - Stochastic convergence

OOM~

Both needs to load the entire corpus into memory and scan through the corpus in each iteration !!!

Make it online



Can be discarded after processing

Don’t need to load into memory in advance

Online LDA

Algorithm 2 Online variational Bayes for LDA

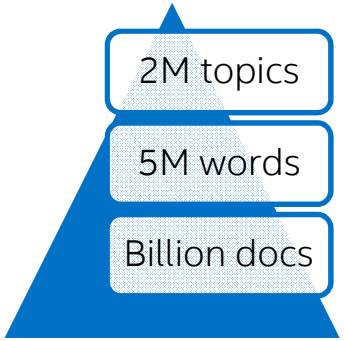
Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $t = 0$ to ∞ **do**
  E step:
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{twk} \propto \exp\{\mathbb{E}_q[\log \theta_{tk}] + \mathbb{E}_q[\log \beta_{kw}]\}$
    Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$
  until $\frac{1}{K} \sum_k |\text{change in} \gamma_{tk}| < 0.00001$
  M step:
  Compute $\tilde{\lambda}_{kw} = \eta + D n_{tw} \phi_{twk}$
  Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}}$.
end for

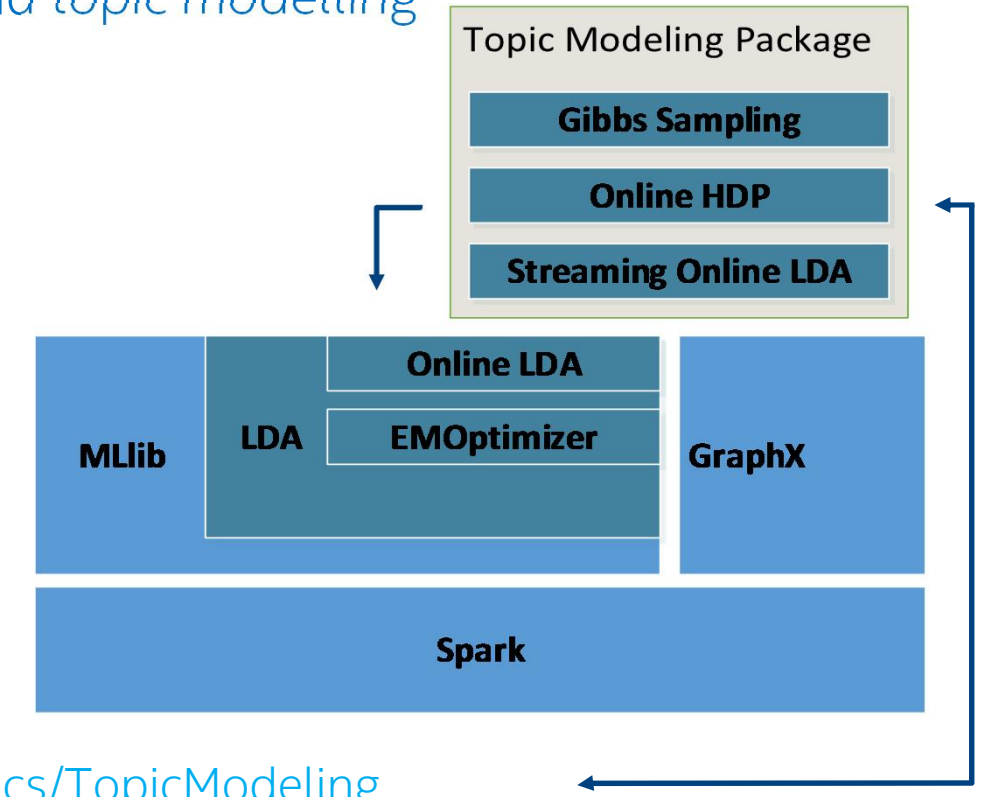
$$\rho_t \triangleq (\tau_0 + t)^{-\kappa}$$

Large-Scale Topic Modeling on Spark

- Adopted by many top Internet companies for text mining and topic modelling

Clustering	• Images • News, advertisements
Feature reduction	• Documents to Topics
Analysis	• Auto Tagging, • Emotional Analysis



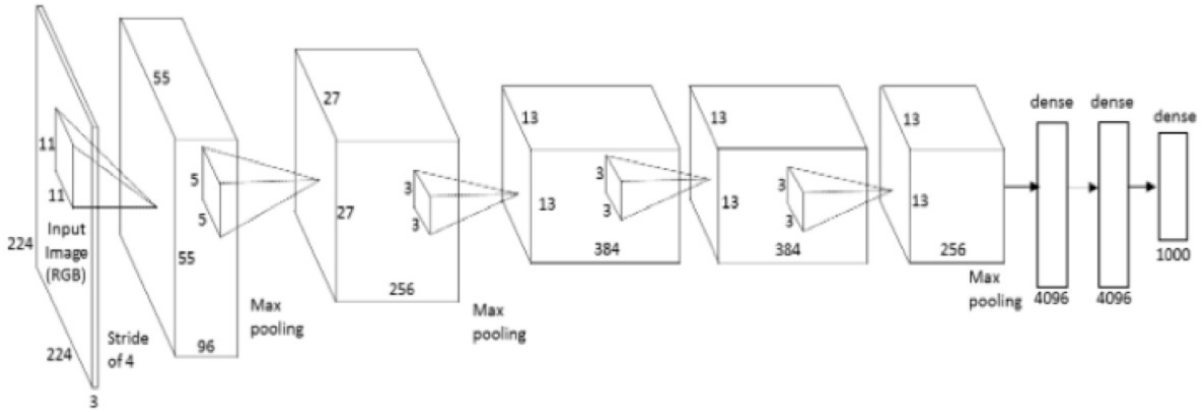


- Full function coverage
  - Large batch computation or online model
  - Training, Prediction, Analysis.
- Spark Package: Topic Modeling https://github.com/intel-analytics/TopicModeling
  - Under active development, welcome to join.
  - Intel is one of the primary contributors for LDA algorithms in Spark MLlib.

intel Software

Deep (Convolutional) neural network

Intuitive API with layer-based interface




```
val trainData = loadData()
val model = new Sequential(…)
model += new Convolution(…)
model += new maxPooling(…)
…
val criterion = new ClassNLLCriterion()
val optimizer = new ParallelOptimizer (model, new SGD)
optimizer.setCrossValidation(evaluator.accuracy)
optimizer.setPath("./model_save.obj")
optimizer.optimize(trainData)
```

Built on top of standard Big Data platforms

- Easily utilize your existing clusters

Engaging industry users and community early

- Evolving with feedback from practical cases
- Community version compatible with Spark MLP.

Targeting Full function coverage:

- Auto Encoder, Sparse Encoder
- Convolution with max and avg pooling
- RBM and DBN

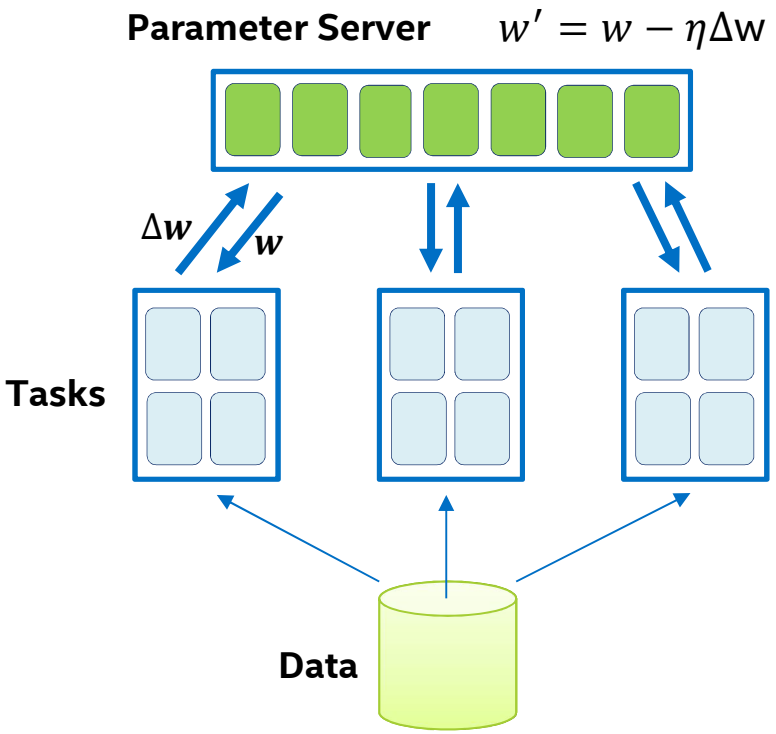
Benchmark with popular dataset / models

- Mnist
- AlexNet
- cifar 10

Easy MKL integration for IA acceleration

intel Software

Parameter Support for Distributed Machine Learning



Successful support for algorithms with linear algebra

- Gradient and weight in logistic regression
- Word-topic distribution/assignment in LDA
- Weight & bias in neural network
- More algorithms are under development and welcome to join.

Scale-out framework to support complex model & training

- Very large scale model/graph (billions of unique features)
- Asynchronous model for distributed training
- Parallel aggregation & update of parameters
- Coarse-grained fault tolerance scheme

Going Spark
Initial proposal/discussions in Spark-10041
Definition of the proper PS interface in Spark, Linear algebra objects, partitioning strategies, epoch support, parallel aggregation, etc.

Machine learning algorithms and Statistics

Need more statistics in Spark? We have a package

- https://github.com/intel-analytics/StatisticsOnSpark
- ANOVA, T sampling, Mann–Whitney U test and more.

Word2Vec Improvements

- Fundamental NLP algorithms, for similarity computation or preprocessing.
- Greatly enlarged the maximum dictionary size
- Reduce memory consumption and execution time.

KMeans support for sparse data:

- Clustering with web scale: Billions of records with dimension of 10M, 200 clusters
- Greatly reduce memory consumption for sparse data.

- Need support for other algorithms? Talk to us, we probably have done that !

intel Software

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.

“ Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.”

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

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.