BigDL

Distributed, High-Performance

Deep Learning Framework

for Apache Spark

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

ANALYTICS ZOO

Unified Analytics + AI Platform

Distributed TensorFlow, Keras, PyTorch and BigDL on Apache Spark

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

Accelerating Data Analytics + AI Solutions At Scale

What's on-going in Spark + AI Community

- views from a contributor & practitioner

Missing feature handling

Shengsheng Huang

Intel AnalyticsZoo team

Agenda

- Efforts for building unified data analytics + AI in production
- Efforts to support emerging AI applications

Agenda

- Efforts for building unified data analytics + AI in production
- Efforts to support emerging AI applications

What's new in spark + ai community

Spark 3. 0

- Optimizations on SQL execution (adaptive query execution, dynamic partition pruning)
- DataSourceV2
- Project Hydrogen (Barrier execution mode, Accelerator-aware scheduling, optimized data exchange)
- Spark Graph
- Spark on Kubernetes
- …

All for Productivity

MLFlow – ML lifecycle management

- Tracking – log code, data, config, results of experiments, and compare & query
- Projects – code packaging format for reproducible runs on any platform
- Models – model packaging format for sending models to diverse deployment tools.

Koalas – pandas API on Spark

Delta Lake – ACID layer upon data lakes

Rationale behind the efforts in community

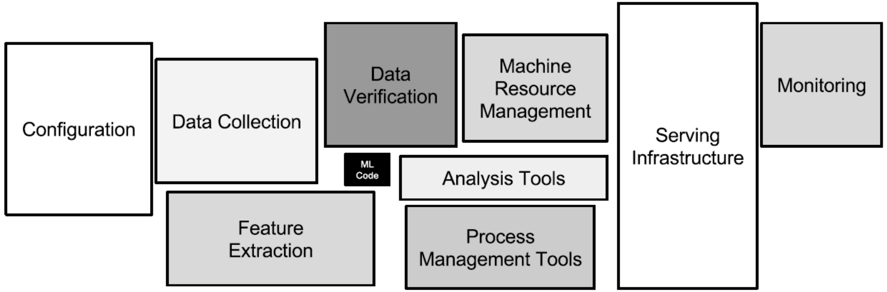


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

“Hidden Technical Debt in Machine Learning Systems”, Sculley et al., Google, NIPS 2015 Paper

- Integration/Injection of heterogenous data models/sources, computation models, software/hardware components, … (e.g. DataSourceV2, Project Hydrogen, Spark Graph)
- E2E Workflow, ML Lifecycle, Serving, Deployment, Orchestration, … (e.g. MLFlow, KubeFlow, Seldon, TFX)
- Efficiency & Reliability (e.g. SQL-related optimizations, Delta Lake)
- Friendly APIs (e.g. Koalas)

AI ON BIG DATA

BigDL

High-Performance
Deep Learning Framework
for Apache Spark

software.intel.com/bigdl

ANALYTICS ZOO

Unified Analytics + AI Platform
Distributed TensorFlow*, PyTorch*, Keras* and BigDL on Apache Spark

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

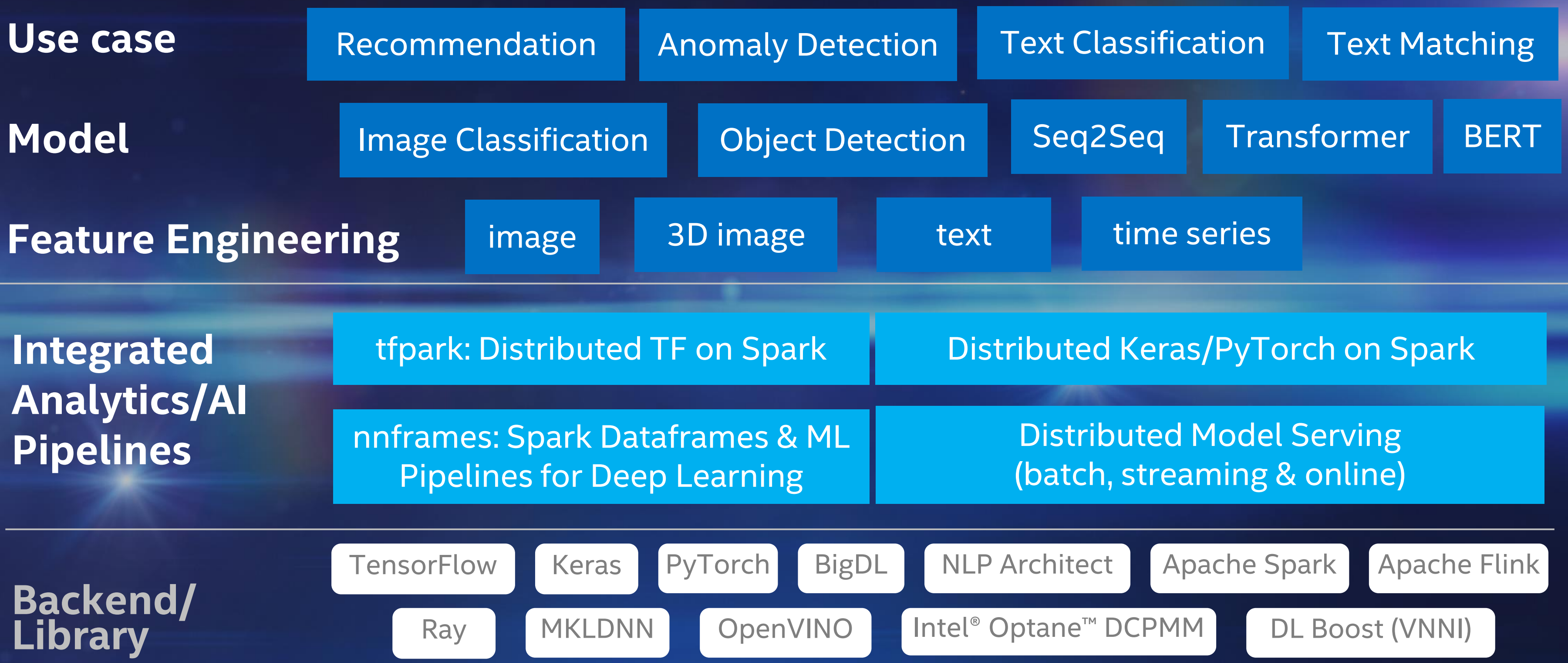
ACCELERATING DATA ANALYTICS + AI SOLUTIONS DEPLOYMENT AT SCALE

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

intel AI

Analytics Zoo

Unified End-to-End Data Analytics + AI Platform



Use case: Recommendation | Anomaly Detection | Text Classification | Text Matching

Model: Image Classification | Object Detection | Seq2Seq | Transformer | BERT

Feature Engineering: image | 3D image | text | time series

Integrated Analytics/AI Pipelines:
- tfpark: Distributed TF on Spark
- Distributed Keras/PyTorch on Spark
- nnframes: Spark Dataframes & ML Pipelines for Deep Learning
- Distributed Model Serving (batch, streaming & online)

Backend/ Library: TensorFlow | Keras | PyTorch | BigDL | NLP Architect | Apache Spark | Apache Flink | Ray | MKLDNN | OpenVINO | Intel® Optane™ DCPMM | DL Boost (VNNI)

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

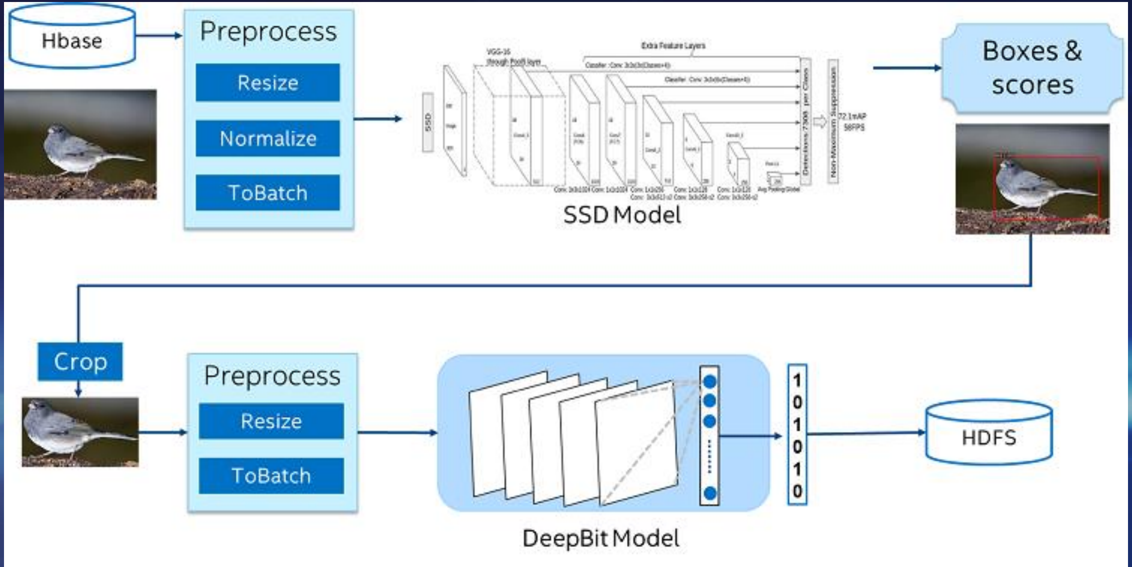
Distributed TensorFlow on Spark

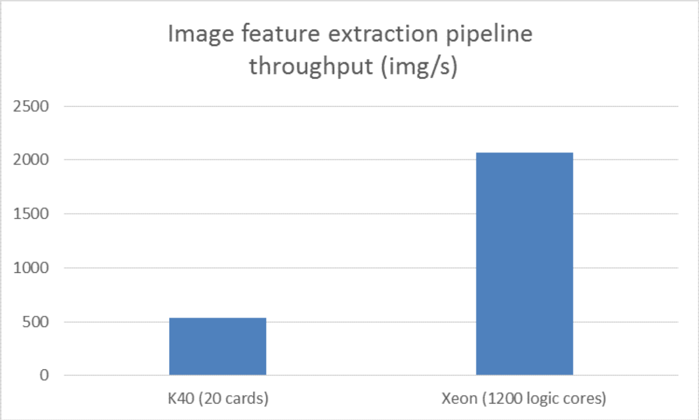
- Data wrangling and analysis using PySpark
- Deep learning model development using TensorFlow or Keras
- Distributed training / inference on Spark

```
#pyspark code
train_rdd = spark.hadoopFile(…).map(…)
dataset = TFDataset.from_rdd(train_rdd,…)

#tensorflow code
import tensorflow as tf
slim = tf.contrib.slim
images, labels = dataset.tensors
with slim.arg_scope(lenet.lenet_arg_scope()):
   logits, end_points = lenet.lenet(images, …)
loss = tf.reduce_mean( \
   tf.losses.sparse_softmax_cross_entropy( \
   logits=logits, labels=labels))

#distributed training on Spark
optimizer = TFOptimizer.from_loss(loss, Adam(…))
optimizer.optimize(end_trigger=MaxEpoch(5))
```

Write TensorFlow code inline in PySpark program

Object Detection and Image Feature Extraction at JD.com*



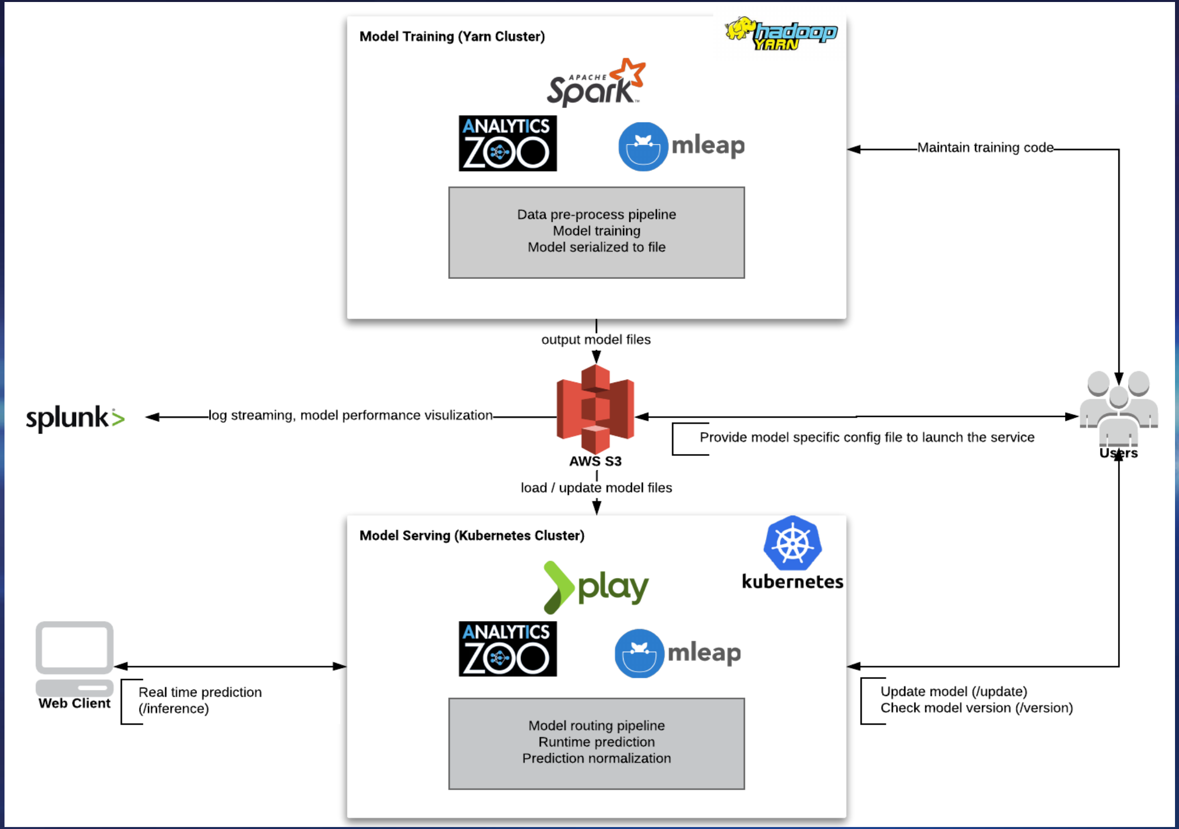


- Reuse existing Hadoop/Spark clusters for deep learning with no changes (image search, IP protection, etc.)
- Efficiently scale out on Spark with superior performance (***3.83x*** speed-up vs. GPU severs) as benchmarked by JD

http://mp.weixin.qq.com/s/xUCkzbHK4K06-v5qUsaNQQ
https://software.intel.com/en-us/articles/building-large-scale-image-feature-extraction-with-bigdl-at-jdcom

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

Product Recommendations in Office Depot*



Model Training (Yarn Cluster)

Data pre-process pipeline
Model training
Model serialized to file

Maintain training code

output model files

log streaming, model performance visulization

Provide model specific config file to launch the service

load / update model files

Model Serving (Kubernetes Cluster)

Model routing pipeline
Runtime prediction
Prediction normalization

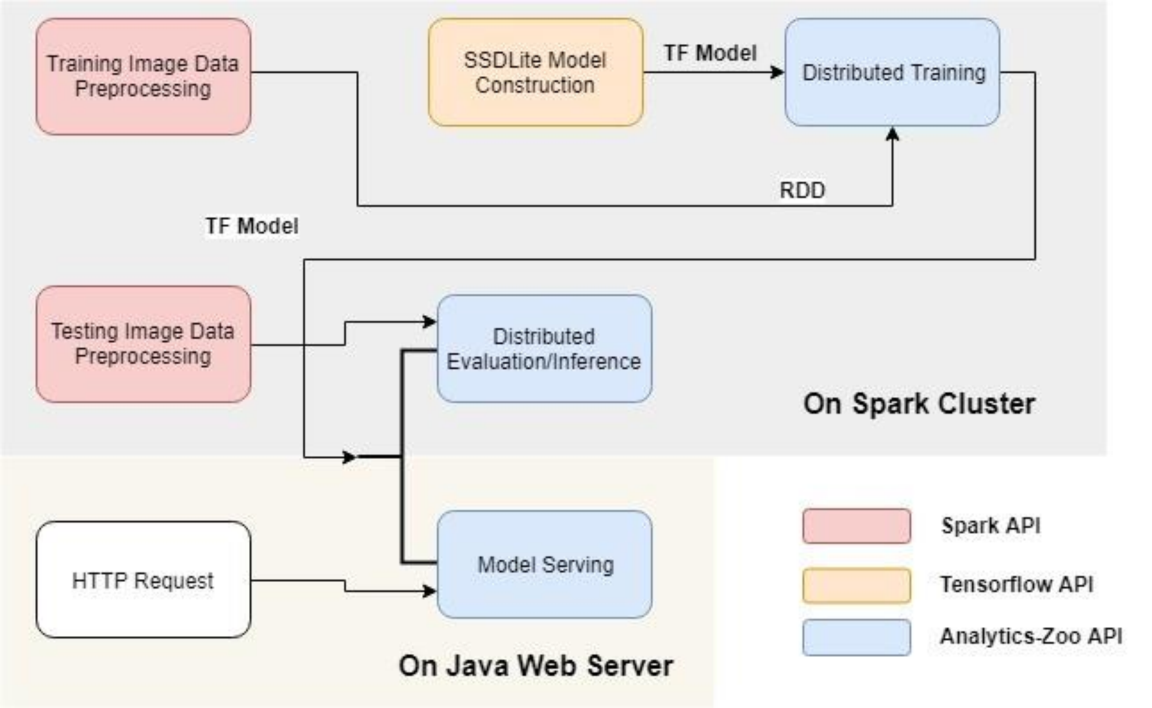
Real time prediction
(/inference)

Update model (/update)
Check model version (/version)

https://software.intel.com/en-us/articles/real-time-product-recommendations-for-office-depot-using-apache-spark-and-analytics-zoo-on

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

Computer Vision Based Product Defect Detection in Midea*



Training Image Data Preprocessing

SSDLite Model Construction

TF Model

Distributed Training

RDD

TF Model

Testing Image Data Preprocessing

Distributed Evaluation/Inference

On Spark Cluster

HTTP Request

Model Serving

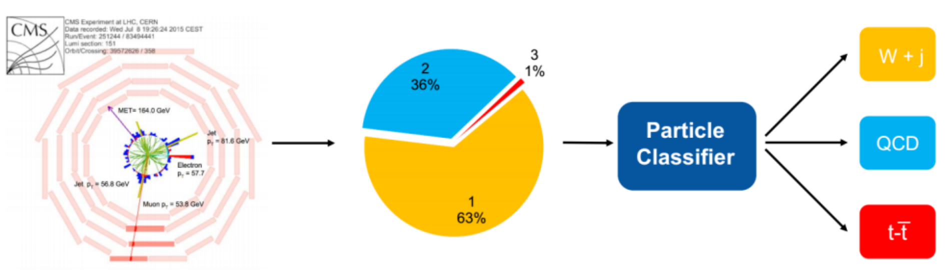
On Java Web Server

🟥	Spark API
🟧	Tensorflow API
🟦	Analytics-Zoo API

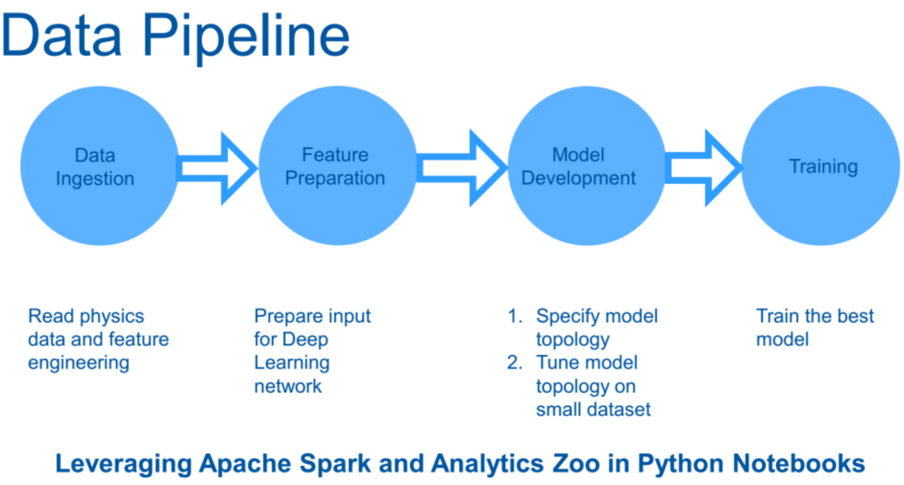
https://software.intel.com/en-us/articles/industrial-inspection-platform-in-midea-and-kuka-using-distributed-tensorflow-on-analytics

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

Particle Classifier for High Energy Physics in CERN*



Deep learning pipeline for physics data



Data Pipeline

Read physics data and feature engineering

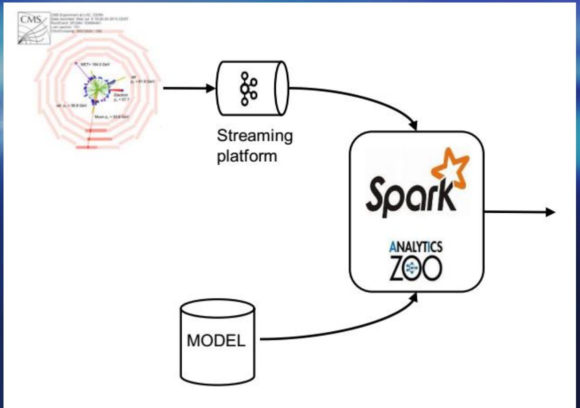
Prepare input for Deep Learning network

1. Specify model topology
2. Tune model topology on small dataset

Train the best model

Leveraging Apache Spark and Analytics Zoo in Python Notebooks

Model serving using Apache Kafka and Spark



https://db-blog.web.cern.ch/blog/luca-canali/machine-learning-pipelines-high-energy-physics-using-apache-spark-bigdl

https://databricks.com/session/deep-learning-on-apache-spark-at-cerns-large-hadron-collider-with-intel-technologies

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

Wrap Up

Community is making efforts to make Spark a unified Analytics + AI platform

Analytics Zoo is also working towards similar goal, by

- Seamless integration various components, e.g. Tensorflow, PyTorch, BigDL, etc.
- Providing full-stack optimizations involving hardware/software (VNNI, MKL-DNN, OpenVINO, etc.)
- Providing ease of use, end-to-end, from laptop to production platform

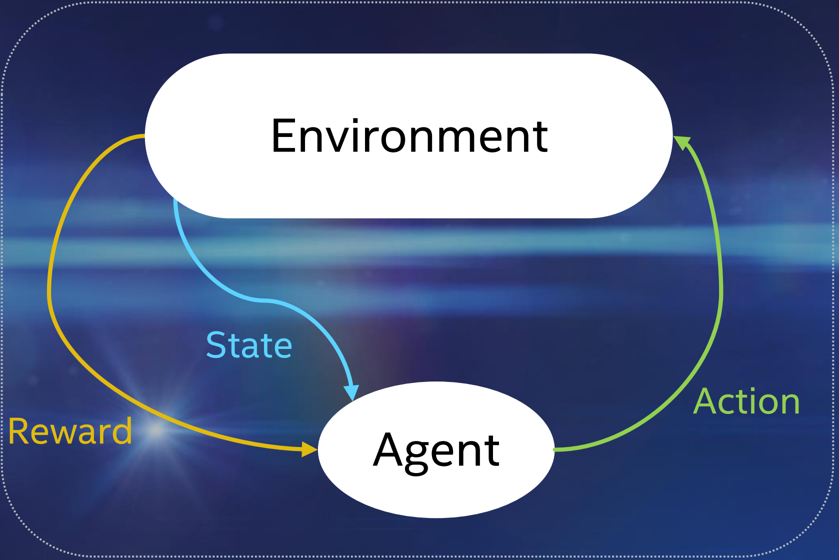
We are both contributors and practitioners. We use, learn, and contribute.

Agenda

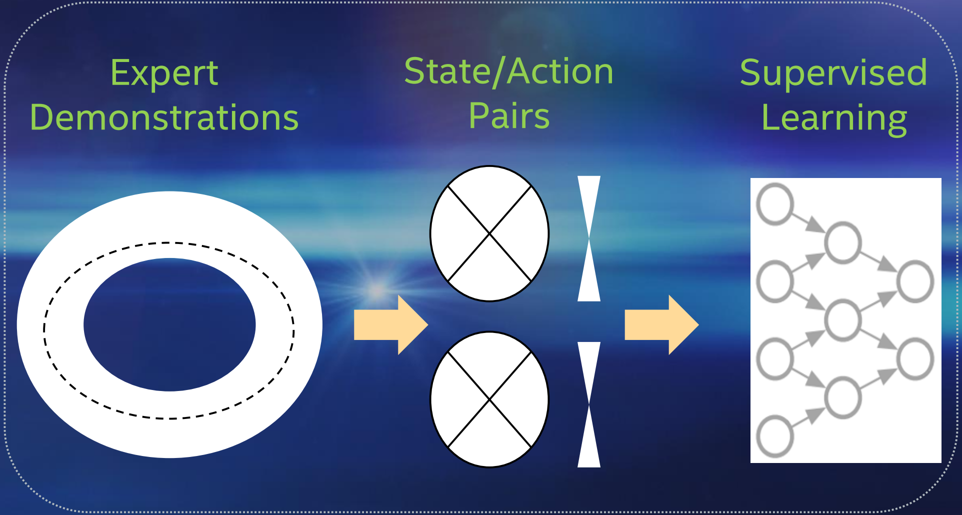
- Efforts for building unified data analytics + AI in production
- Efforts to support emerging AI applications

Towards General AI

Strategies to build AI for game playing, robots, autonomous driving, etc.

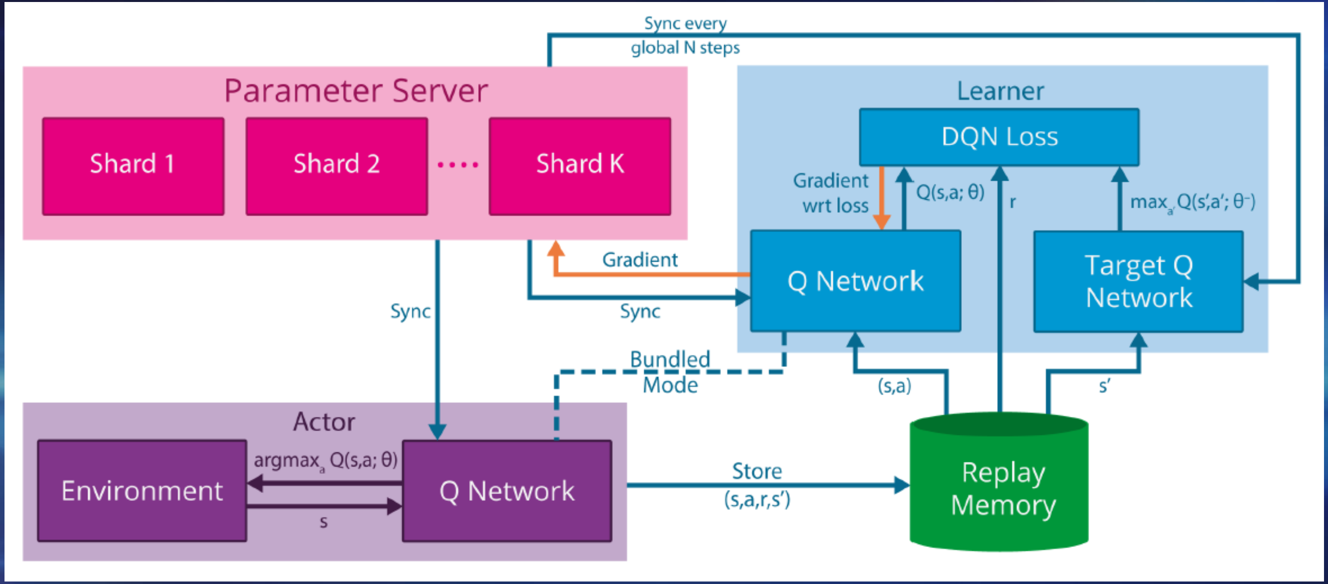


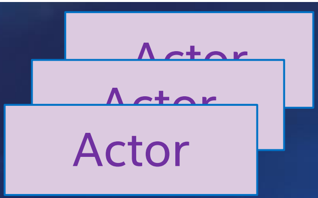
Deep Reinforcement Learning (DRL)



Imitation Learning

Parallel Architecture for Deep RL





Massively Parallel Methods for Deep Reinforcement Learning
https://arxiv.org/abs/1507.04296

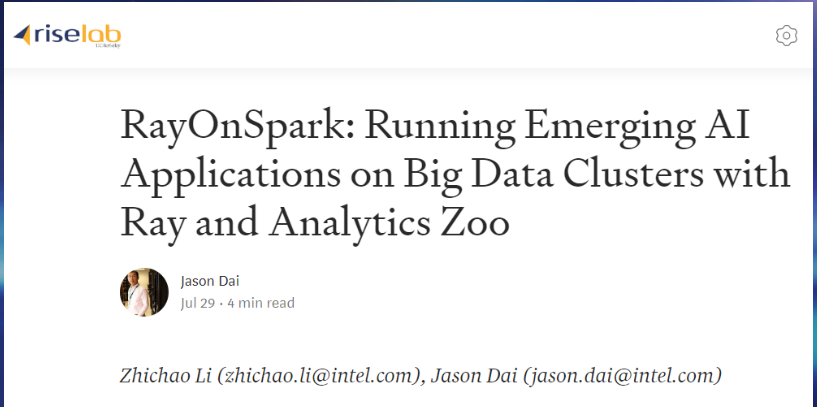
Ray On Spark

Ray

- https://github.com/ray-project/ray
- a distributed framework for emerging AI applications open-sourced by UC Berkeley RISELab

RayOnSpark

- a feature recently added to Analytic Zoo
- allows users to directly run Ray programs on Apache Hadoop*/YARN
- Ray applications can be seamlessly integrated into Spark pipeline and operate directly on Spark RDDs or DataFrames.



RayOnSpark: Running Emerging AI Applications on Big Data Clusters with Ray and Analytics Zoo

Jason Dai
Jul 29 · 4 min read

Zhichao Li (zhichao.li@intel.com), Jason Dai (jason.dai@intel.com)

https://medium.com/riselab/rayonspark-running-emerging-ai-applications-on-big-data-clusters-with-ray-and-analytics-zoo-923e0136ed6a

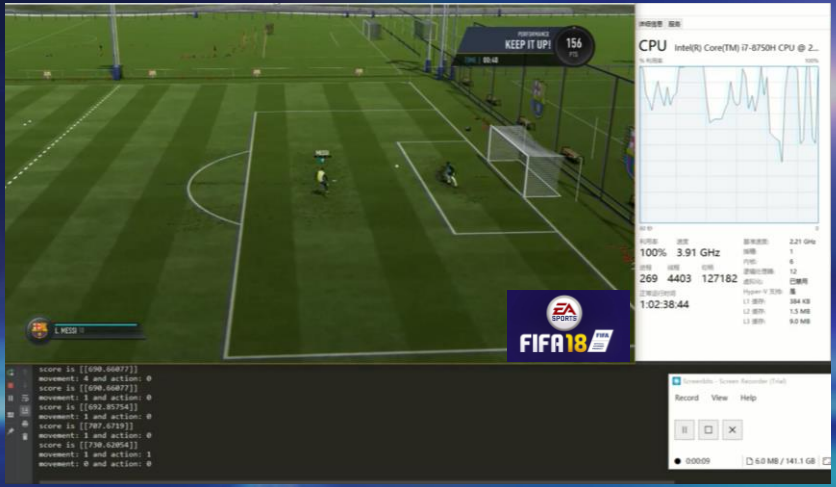
*Other names and brands may be claimed as the property of others.

Building AI to Play FIFA

FIFA18* – A real-time 3D soccer simulation video game by Electronic Arts*

Our Experiment Platform (collaborations w/ SJTU)

- runs alongside FIFA game in a non-intrusive way
- provides abstraction of game environment (observations, actions, rewards, scores, semantics, etc.)
- Implemented agents: RL, IL, Hybrid (IL + RL)



```
score is [[690.66077]]
movement: 4 and action: 0
score is [[690.66077]]
movement: 1 and action: 0
score is [[692.85754]]
movement: 1 and action: 0
score is [[707.6719]]
movement: 1 and action: 0
score is [[730.62054]]
movement: 1 and action: 1
movement: 0 and action: 0
```

Future Work:

- Transfer between Google Research Football and FIFA?
- Train agents in massive scale w/ Ray & RayOnSpark
- Additional models/scenarios, etc.

https://www.slideshare.net/jason-dai/building-ai-to-play-the-fifa-video-game-using-distributed-tensorflow-on-analytics-zoo

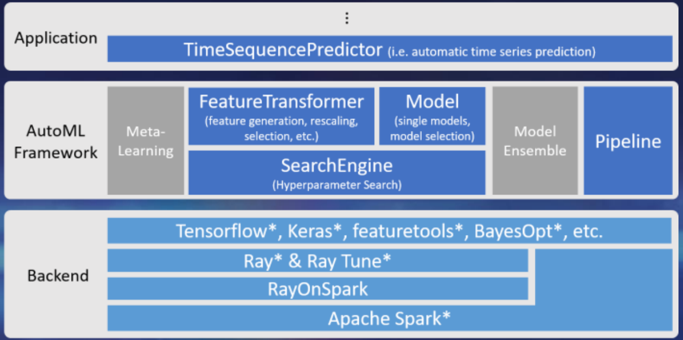
Results on Shooting Bronze Scenario

		Score	Goal Ratio
Human	master	10112.78	92%
	demonstrator	7284.98	84.96%
Agent	IL	10345.18	92.54%
	RL (Policy Gradient)	5606.31	40.25%
	Hybrid (RL+IL)	10514.43	95.59%

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

Scalable AutoML for Time Series Analysis

AutoML Framework



Application

⋮

TimeSequencePredictor (i.e. automatic time series prediction)

AutoML Framework

Meta-Learning

FeatureTransformer (feature generation, rescaling, selection, etc.)

Model (single models, model selection)

SearchEngine (Hyperparameter Search)

Model Ensemble

Pipeline

Backend

Tensorflow*, Keras*, featuretools*, BayesOpt*, etc.

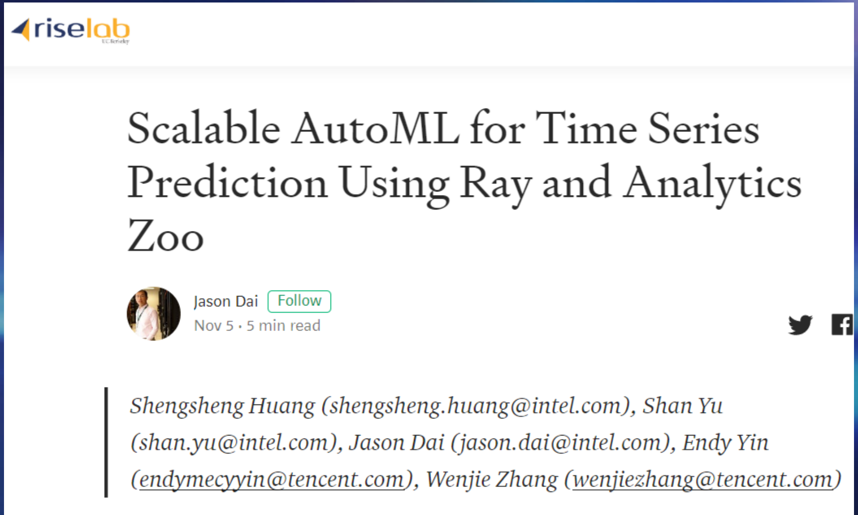
Ray* & Ray Tune*

RayOnSpark

Apache Spark*

Time Series Forecasting w/ AutoML

- Data processing and feature engineering
- Neural network based (hybrid) models
- Automated feature selection, model selection, hyper parameter tuning



riselab

Scalable AutoML for Time Series Prediction Using Ray and Analytics Zoo

Jason Dai Follow

Nov 5 · 5 min read

> *Shengsheng Huang (shengsheng.huang@intel.com), Shan Yu (shan.yu@intel.com), Jason Dai (jason.dai@intel.com), Endy Yin (endymecyyin@tencent.com), Wenjie Zhang (wenjiezhang@tencent.com)*

https://medium.com/riselab/scalable-automl-for-time-series-prediction-using-ray-and-analytics-zoo-b79a6fd08139

Wrap Up

We’re extending the Spark stack to support emerging AI applications

- RayOnSpark

We’re building emerging AI applications

- Building AI to play FIFA
- Scalable AutoML for Time Series Analysis

More Information on Analytics Zoo

- **Project website**
  - *https://github.com/intel-analytics/analytics-zoo*
- **Tutorials**
  - CVPR 2018: *https://jason-dai.github.io/cvpr2018/*
  - AAAI 2019: *https://jason-dai.github.io/aaai2019/*
- **“BigDL: A Distributed Deep Learning Framework for Big Data”**
  - *In proceedings of ACM Symposium on Cloud Computing 2019 (SOCC’19)*
- **Use cases**
  - *Azure, CERN, MasterCard, Office Depot, Tencent, Midea, etc.*
  - *https://analytics-zoo.github.io/master/#powered-by/*

intel® AI

LEGAL NOTICES AND DISCLAIMERS

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/performance.

Intel does not control or audit the design or implementation of third-party benchmark data or websites referenced in this document. Intel encourages all of its customers to visit the referenced websites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

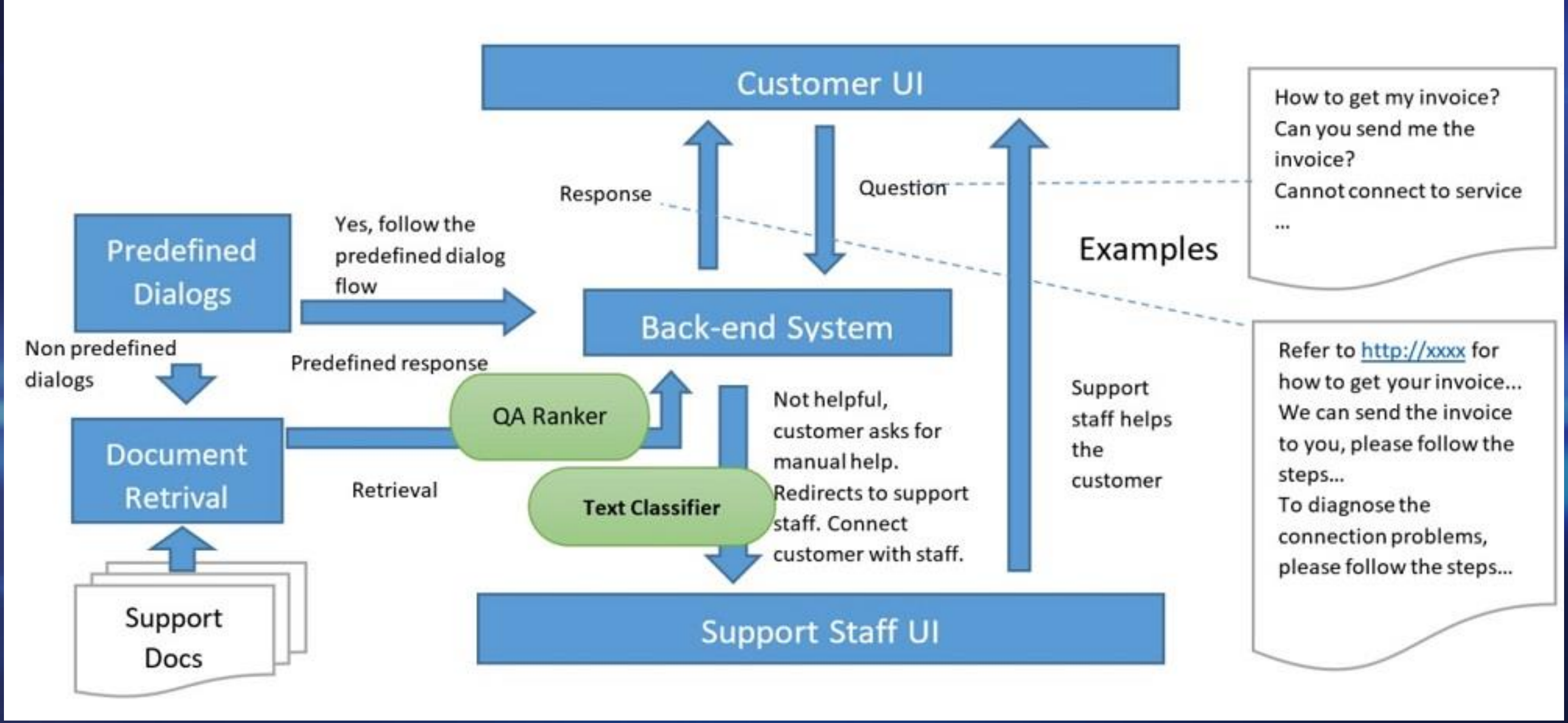
Optimization notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com/benchmarks.

Intel, the Intel logo, Intel Inside, the Intel Inside logo, Intel Atom, Intel Core, Iris, Movidius, Myriad, Intel Nervana, OpenVINO, Intel Optane, Stratix, and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

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

NLP Based Customer Service Chatbot for Microsoft Azure


Customer UI
Response
Question
Predefined Dialogs
Yes, follow the predefined dialog flow
Back-end System
Examples
Non predefined dialogs
Predefined response
QA Ranker
Document Retrival
Retrieval
Text Classifier
Not helpful, customer asks for manual help. Redirects to support staff. Connect customer with staff.
Support staff helps the customer
Support Docs
Support Staff UI
How to get my invoice?
Can you send me the invoice?
Cannot connect to service
...
Refer to http://xxxx for how to get your invoice...
We can send the invoice to you, please follow the steps...
To diagnose the connection problems, please follow the steps...

https://software.intel.com/en-us/articles/use-analytics-zoo-to-inject-ai-into-customer-service-platforms-on-microsoft-azure-part-1
https://www.infoq.com/articles/analytics-zoo-qa-module/

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

BigDL

Distributed, High-Performance

Deep Learning Framework

for Apache Spark

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

ANALYTICS ZOO

Unified Analytics + AI Platform

Distributed TensorFlow, Keras, PyTorch and BigDL on Apache Spark

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

Accelerating Data Analytics + AI Solutions At Scale

intel AI