Intel AI

使用分布式自动机器学习进行时间序列分析

Shengsheng Huang, Shan Yu, Jason Dai

Intel *Analytics Zoo* Team

ANALYTICS ZOO

Nov 11, 2019

Agenda

A Unified Analytics + AI platform – Analytics Zoo

Background about Time Series Forecasting

- Time Series Forecasting and its applications
- Pain points & how we address them

Time Series Forecasting with AutoML in Analytics Zoo

- Architecture & Training Workflow
- Features & Usage

Nov 11, 2019

A Unified Analytics + AI Platform

Nov 11, 2019

What is Analytics Zoo

BigDL

Distributed, High-Performance

Deep Learning Framework

for Apache Spark

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

ANALYTICS ZOO

Unified Analytics + AI Platform

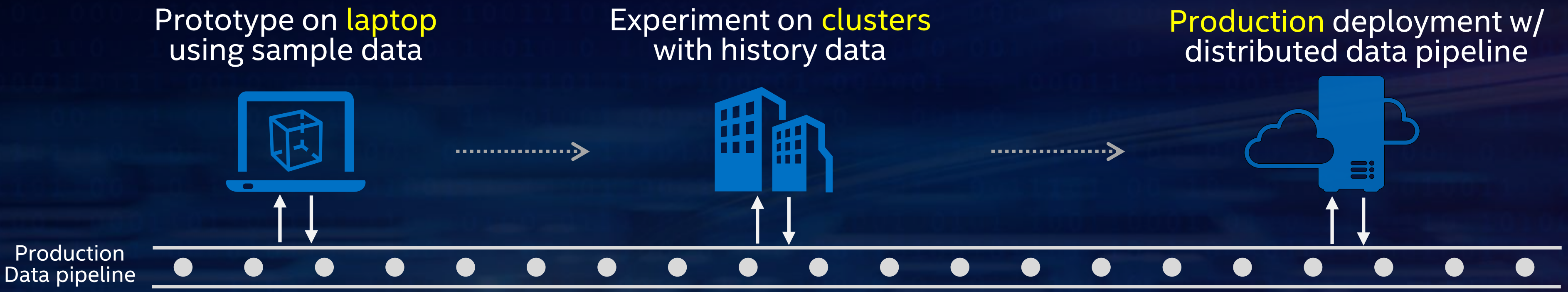
Distributed TensorFlow, Keras, PyTorch and BigDL on Apache Spark

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

Accelerating Data Analytics + AI Solutions At Scale

Unified Big Data Analytics and AI Platform

Seamless Scaling from Laptop to Production



Prototype on laptop using sample data

Experiment on clusters with history data

Production deployment w/ distributed data pipeline

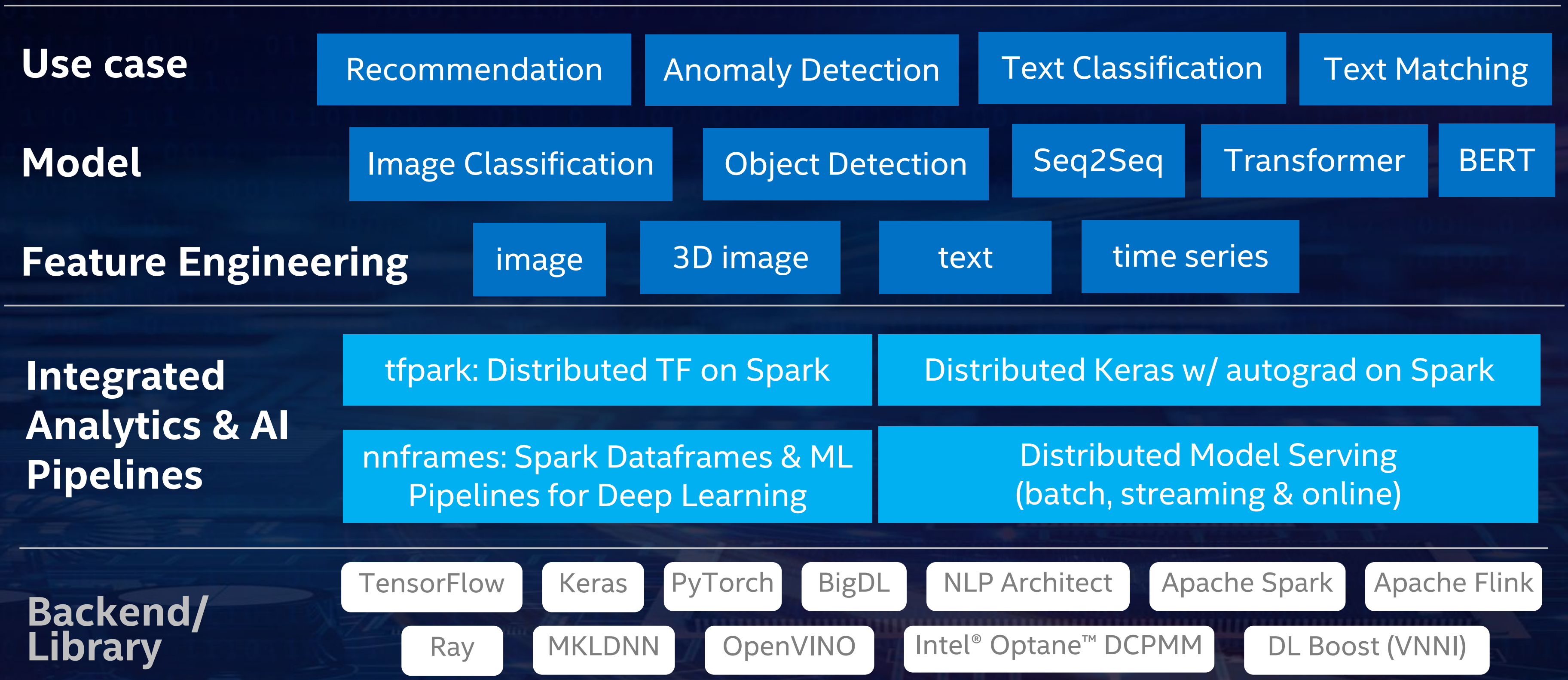
Production Data pipeline

- Easily prototype the **integrated data analytics & AI solution**
- **"Zero" code change** from laptop to distributed cluster
- **Directly access production data** (Hadoop/Hive/HBase) without data copy
- Seamlessly deployed on **production big data clusters**

Nov 11, 2019

Analytics Zoo

Unified Big Data Analytics and 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 w/ autograd 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

Nov 11, 2019

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/

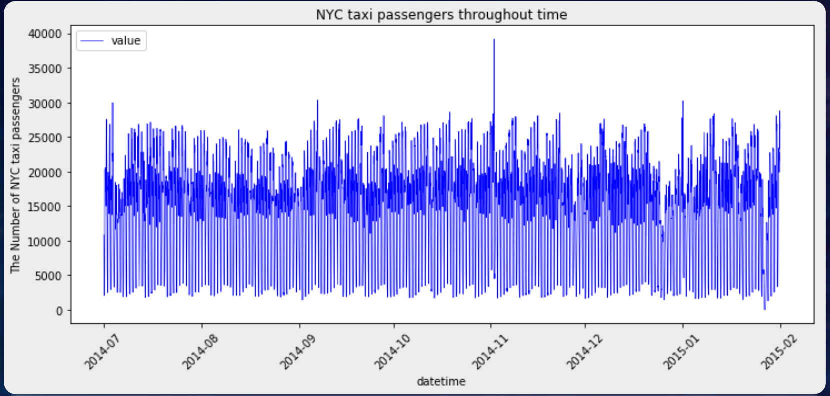
Background about Time Series Forecasting

Nov 11, 2019

Time Series Data

- **What is Time Series**
  - A time series is a series of data points indexed/listed in **time order**.
  - Usually **numerical**
    - scalar (univariant)
    - vector (multivariant)
  - Unstructured data (video, songs, etc.)

- **Examples**
  - Stock prices, sales volume, IoT sensor readings, CPU/IO monitoring data, etc.

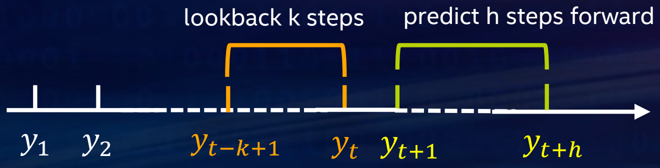


Total volume of taxi passengers in NYC from 2014/07-2015/02 (source : https://github.com/intel-analytics/analytics-zoo/blob/master/apps/anomaly-detection/anomaly-detection-nyc-taxi.ipynb)

Nov 11, 2019

Time Series Forecasting

- **What is Time Series Forecasting**
  - Given all history observations $y_1, \ldots, y_t$, Predict values of next **h** steps, $y_{t+1}, \ldots, y_{t+h}$
  - Usually only lookback **k** steps, $y_{t-k+1}, \ldots, y_t$



- **Applications**
  - Sales volume/demand prediction, etc.
  - As the 1st step for Anomaly Detection
  - AIOps (anomaly detection, root case analysis, resource planning, etc.)

Nov 11, 2019

Pain points and how we address them

- Pain Points of Traditional Methods
  - Widely-used statistics based models (AR, MA, ES, ARIMA, etc.)
    - Hard to capture complex non-linear, cross-series patterns in (multivariant) data
    - Make (unreasonable) assumptions about underlying distribution
  - Some methods are computational costly (e.g. Gaussian Process based methods)
  - Hard to integrate & scale with production solutions/pipelines

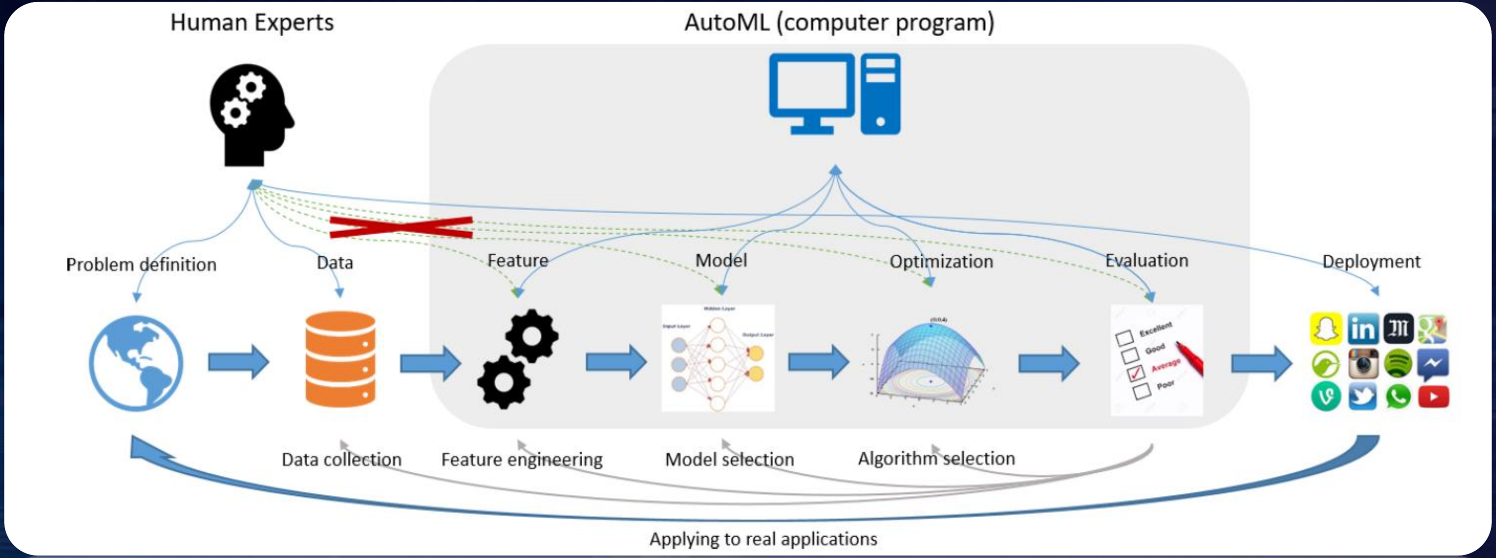
- What's in Analytics Zoo
  - Neural networks based (hybrid) models – more flexible and expressive
  - additional data processing, features, and metrics for time series
  - AutoML for hyper-parameter tuning, model selection, feature selection, etc.
  - Scalability and E2E Pipelines

Nov 11, 2019

Time Series Forecasting w/ AutoML in *Analytics Zoo*

Nov 11, 2019

AutoML Overview

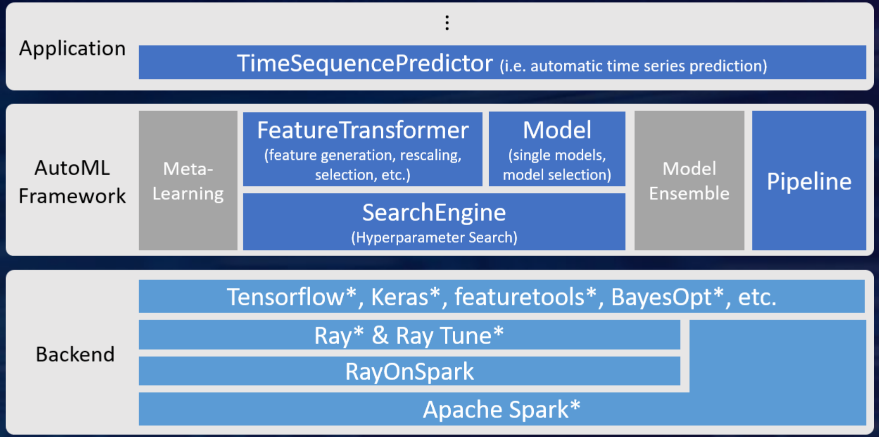


Source: Taking the Human out of Learning Applications : A Survey on Automated Machine Learning. Yao, Q., Wang, et. al

Nov 11, 2019

AutoML + Time Series Prediction In *Analytics Zoo*

- **AutoML Framework**
  - FeatureTransformer
  - Model
  - SearchEngine
  - Pipeline
- **Time Series Prediction w/ AutoML**
  - TimeSequencePredictor
  - TimeSequencePipeline



⋮

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*

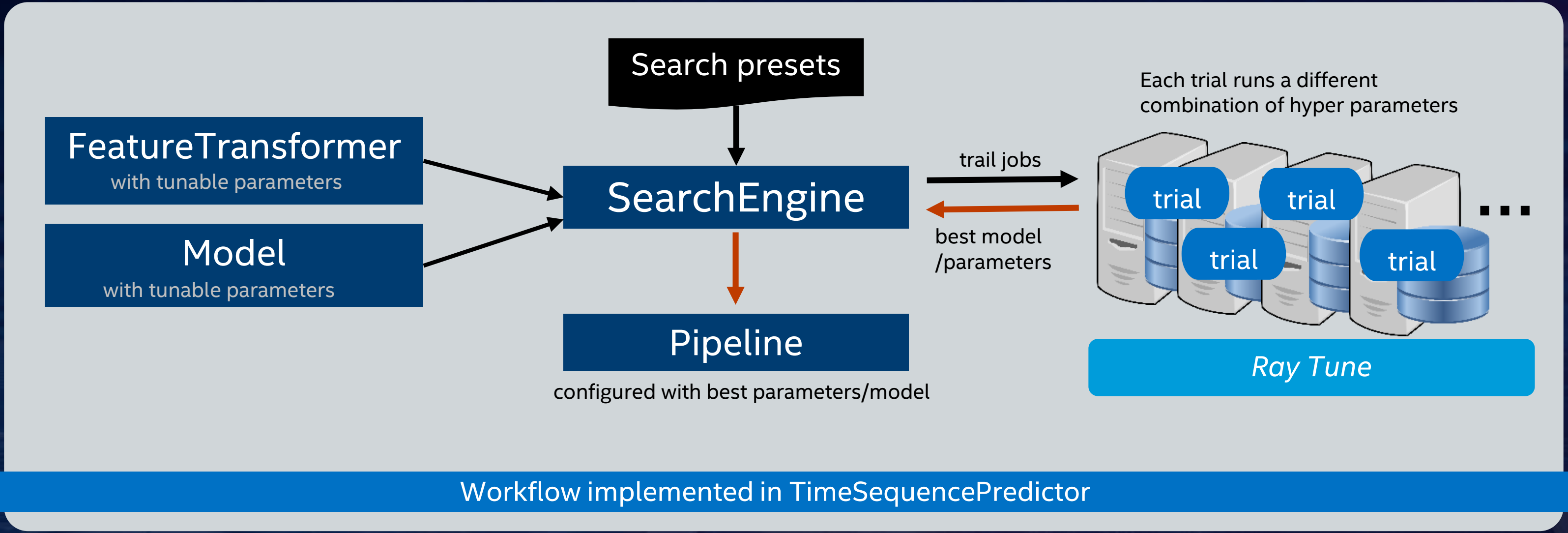
RayOnSpark

Apache Spark*

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

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

Typical Workflow of Training w/ AutoML



Search presets

FeatureTransformer
with tunable parameters

Model
with tunable parameters

SearchEngine

Pipeline
configured with best parameters/model

trail jobs

best model
/parameters

Each trial runs a different
combination of hyper parameters

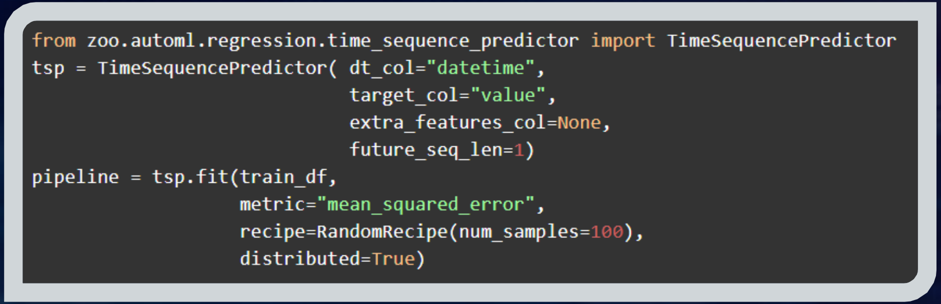
Ray Tune

Workflow implemented in TimeSequencePredictor

Nov 11, 2019

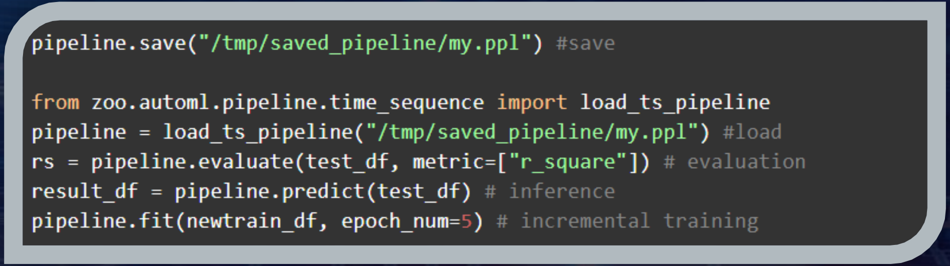
General API Usage

- Training a **Predictor**
  - **fit** (w/ automl)
  - recipe
  - distributed



```python
from zoo.automl.regression.time_sequence_predictor import TimeSequencePredictor
tsp = TimeSequencePredictor( dt_col="datetime",
                             target_col="value",
                             extra_features_col=None,
                             future_seq_len=1)
pipeline = tsp.fit(train_df,
                   metric="mean_squared_error",
                   recipe=RandomRecipe(num_samples=100),
                   distributed=True)
```

- Using a **Pipeline**
  - save/load
  - evaluate/predict
  - fit (incremental)



```python
pipeline.save("/tmp/saved_pipeline/my.ppl") #save

from zoo.automl.pipeline.time_sequence import load_ts_pipeline
pipeline = load_ts_pipeline("/tmp/saved_pipeline/my.ppl") #load
rs = pipeline.evaluate(test_df, metric=["r_square"]) # evaluation
result_df = pipeline.predict(test_df) # inference
pipeline.fit(newtrain_df, epoch_num=5) # incremental training
```

State-of-Art Neural Networks for Time Series Forecasting

- Non-linear(NN) + Linear (AR)
- NN handles time series as a **sequence modeling problem** (strategies usually seen in NLP are used, e.g. LSTM/GRU, encoder-decoder, attention, memory networks, transformer, etc.)

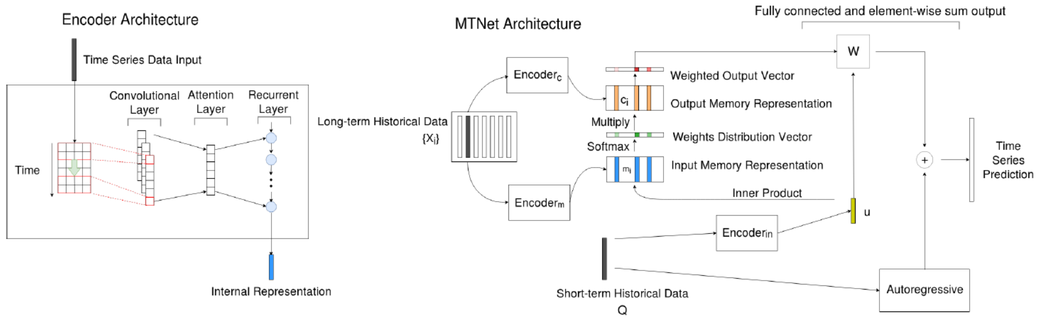


Figure 1: An overview of Memory Time-series network (MTNet) on the right and the details of the encoder architecture on the left

A Memory-Network Based Solution for Multivariate Time-Series Forecasting
https://arxiv.org/abs/1809.02105

Nov 11, 2019

Future Work

- Time Series
  - Additional models (e.g. statistical, MLP, transformer, etc.)
  - Additional features (e.g. auto-encoder, etc.)
- AutoML
  - Model Ensemble
  - Neural Architecture Search

More Information about AutoML+TimeSeries in Analytics Zoo

- **Resources**
  - Source code as a branch of analytics-zoo repo @ https://github.com/intel-analytics/analytics-zoo/tree/automl
  - README @ https://github.com/intel-analytics/analytics-zoo/blob/automl/pyzoo/zoo/automl/README.md
  - A demo notebook @ https://github.com/intel-analytics/analytics-zoo/blob/automl/apps/automl/nyc_taxi_dataset.ipynb
  - Blog https://medium.com/riselab/scalable-automl-for-time-series-prediction-using-ray-and-analytics-zoo-b79a6fd08139

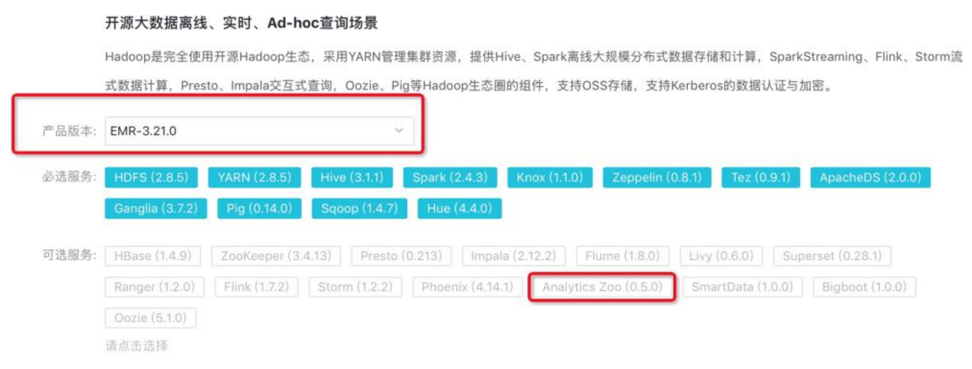
- **Contact *AnalyticsZoo* team or community**
  - Discuss it in analytics-zoo user-group @ https://groups.google.com/forum/#!forum/bigdl-user-group
  - Raise issues or questions @ https://github.com/intel-analytics/analytics-zoo/issues
  - Contact me @ shan.yu@intel.com

Nov 11, 2019

在阿里云E-MR上使用Analytics Zoo

ANALYTICS ZOO + Alibaba Cloud aliyun.com

Analytics Zoo已经集成在阿里云E-MR平台：



开源大数据离线、实时、Ad-hoc查询场景

Hadoop是完全使用开源Hadoop生态，采用YARN管理集群资源，提供Hive、Spark离线大规模分布式数据存储和计算，SparkStreaming、Flink、Storm流式数据计算，Presto、Impala交互式查询，Oozie、Pig等Hadoop生态圈的组件，支持OSS存储，支持Kerberos的数据认证与加密。

产品版本：EMR-3.21.0

必选服务：HDFS (2.8.5) YARN (2.8.5) Hive (3.1.1) Spark (2.4.3) Knox (1.1.0) Zeppelin (0.8.1) Tez (0.9.1) ApacheDS (2.0.0) Ganglia (3.7.2) Pig (0.14.0) Sqoop (1.4.7) Hue (4.4.0)

可选服务：HBase (1.4.9) ZooKeeper (3.4.13) Presto (0.213) Impala (2.12.2) Flume (1.8.0) Livy (0.6.0) Superset (0.28.1) Ranger (1.2.0) Flink (1.7.2) Storm (1.2.2) Phoenix (4.14.1) Analytics Zoo (0.5.0) SmartData (1.0.0) Bigboot (1.0.0) Oozie (5.1.0)

请点击选择

获取技术和资源支持，请联系：

邮件： wesley.du@intel.com
钉钉：

Nov 11, 2019

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.
-