AI Trends Data Science Basics TensorFlow

Search

Most Useful C/C++ ML Libraries Every Data Scientist Should Know

Get in Touch

Author: Posted On: September 23, 2020 Post Comments: 0

Source

Importance of C++ in Data Science and Big Data

Introduction and Motivation – Why C++

C++ is ideal for **dynamic load balancing**, **adaptive caching,** and **developing large big data frameworks,** and libraries. Google's **MapReduce, MongoDB**, most of the **deep learning libraries** listed below have been **implemented using C++. Scylla** known for its **ultra-low latency** and **extremely high throughput** is coded using C++ acts as a replacement to **Apache Cassandra and Amazon DynamoDB**.

With some of the **unique** advantages of C++ as a programming language, (including **memory management, performance characteristics,** and **systems programming**), it definitely serves as one of the most **efficient** tools for developing fast scalable Data Science and Big Data libraries.

Further, **Julia** (a **compiled** and **interactive** language – developed from **MIT**) is emerging as a potential competitor to Python in the field of scientific computing and data processing. Its fast processing speed, parallelism, static along with dynamic typing and C++ bindings for plugging in libraries, has eased the job for developers/data scientists to integrate and use **C++** as data science and big data library.

> *Let's take a closer look at different C++ libraries that can become useful to every data scientist for traditional and deep learning models.*

1. TensorFlow from Google AI

Popular Deep Learning Library developed by Google with its own ecosystem of tools, libraries, and community resources that lets researchers and developers build and deploy ML-powered applications easily

- https://www.tensorflow.org/lite/microcontrollers/library
- https://github.com/tensorflow/serving

2. Caffe from Berkeley

Convolutional Architecture for Fast Feature Embedding or Caffe is written in C++ for a deep learning framework, has been developed by the Berkeley Vision and Learning Center.

https://github.com/intel/caffe

3. Microsoft Cognitive Toolkit (CNTK)

Microsoft Cognitive Toolkit is a unified deep-learning toolkit that helps to translate neural networks as a series of computational steps via a directed graph.

4. mlpack Library

mlpack: It is a fast, flexible machine learning library, written in C++.that provides cutting-edge machine learning algorithms with Python bindings, Julia bindings, and C++ classes.

5. DyNet

Dynamic Neural Network Toolkit (supports computational graph on the fly) or DyNet is a high-performance neural network library written in C++ (with bindings in Python) that runs efficiently on CPU or GPU. It has support for natural language processing, graph structures, reinforcement learning, and other such.

6. Shogun

Shogun is an open-source machine learning library that offers a wide range of efficient and unified machine learning methods, like combination of multiple data representations, algorithm classes, and general-purpose tools for rapid prototyping of data pipelines.

7. FANN

Fast Artificial Neural Network is a multilayer artificial neural networks in C with support for both fully connected and sparsely connected networks. It has support for cross-platform execution in both fixed and floating points. In addition, it has support for evolving topology-based training and backpropagation based DL model training.

8. OpenNN

Open Neural Networks (OpenNN) is an open-source (C/C++) neural networks high-performance library for advanced analytics, with support for classification, regression, forecasting, among others.

9. SHARK Library

Shark is a fast, modular, general open-source machine learning library (C/C++), for applications and research, with support for linear and nonlinear optimization, kernel-based learning algorithms, neural networks, and various other machine learning techniques.

10. Armadillo

Armadillo is a linear algebra (C/C++) library with functionality similar to Matlab. The library is famous for the quick conversion of research code into production environments, for pattern recognition, computer vision, signal processing, bioinformatics, statistics, econometrics, among others.

11. Faisis

faiss: This library (C/C++) is used for efficient similarity search and clustering of dense vectors. It contains algorithms that search in sets of vectors of any size, up to ones that possibly do not fit in RAM. It also has support for optional GPU provided via CUDA, and an optional Python interface.

12. RandomForest

- https://github.com/zhufangzhou/RandomForest
- https://github.com/bjoern-andres/random-forest

13. Boosting

- XGBoost – A parallelized optimized general purpose gradient boosting library.
- ThunderGBM – A fast library for GBDTs and Random Forests on GPUs.
- LightGBM – Microsoft's fast, distributed, high-performance gradient boosting (GBDT, GBRT, GBM or MART) framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks.
- CatBoost – General purpose gradient boosting on decision trees library with categorical features support out of the box. It is easy to install, contains fast inference implementation and supports CPU and GPU (even multi-GPU) computation.

14. Recommendation Systems

- Recommender – A C library for product recommendations/suggestions using collaborative filtering (CF).
- Hybrid Recommender System – A hybrid recommender system based upon scikit-learn algorithms

15. Natural Language Processing

- BLLIP Parser – BLLIP Natural Language Parser (also known as the Charniak-Johnson parser).
- colibri-core – C++ library, command-line tools, and Python binding for extracting and working with basic linguistic constructions such as n-grams and skiagrams in a quick and memory-efficient way.
- CRF++ – Open source implementation of Conditional Random Fields (CRFs) for segmenting/labeling sequential data & other Natural Language Processing tasks. [Deprecated]
- CRFsuite – CRFsuite is an implementation of Conditional Random Fields (CRFs) for labeling sequential data. [Deprecated]
- CRF Models -Different deep learning-based CRF models.
- frog – Memory-based NLP suite developed for Dutch: PoS tagger, lemmatizer, dependency parser, NER, shallow parser, morphological analyzer.
- MeTA – ModErn Text Analysis is a C++ Data Sciences Toolkit that facilitates mining of big text data, including text tokenization, including deep semantic features like parse trees, topic models, classification algorithms, graph-algorithms, language models, multithreaded algorithms, etc.
- MIT Information Extraction Toolkit – C, C++, and Python tools for named entity recognition and relation extraction
- ucto – Unicode-aware regular-expression based tokenizer for various languages. Tool and C++ library. Supports FoLiA format.

16. Data Mining for Streams

- StreamDM : Useful for Mining Big Data Streams which has support for the following algorithms

SGD Learner and Perceptron : Naive Bayes,
CluStream,
Hoeffding Decision Trees,
Bagging,
Stream KM++
Data Generators HyperplaneGenerator,
RandomTreeGenerator,
RandomRBFGenerator,
RandomRBFEventsGenerator

17. Data structures/Graph/Dynamic Programming Algorithms

- https://github.com/TheAlgorithms/C-Plus-Plus

18. General-Purpose Machine Learning

- Darknet – Darknet is an open-source neural network framework written in C and CUDA, that supports CPU and GPU computation.
- cONNXr – An `ONNX` runtime is written in pure C (99) with zero dependencies focused on small embedded devices. Run inference on your machine learning models no matter which framework you train it with. Easy to install and compiles everywhere, even in very old devices.
- BanditLib – A simple Multi-armed Bandit library. **[Deprecated]**
- CUDA – This is a fast C++/CUDA implementation of convolutional deep Learning
- DeepDetect – A machine learning API and server written in C++11. It makes state of the art machine learning easy to work with and integrate into existing applications.
- Distributed Machine learning Tool Kit (DMTK) – A distributed machine learning (parameter server) framework by Microsoft. Enables training models on large data sets across multiple machines. Current tools bundled with it include: LightLDA and Distributed (Multisense) Word Embedding.
- DLib – A suite of ML tools designed to be easy to imbed in other applications.
- DSSTNE – A software library created by Amazon for training and deploying deep neural networks using GPUs which emphasizes speed and scale over experimental flexibility.
- DyNet – A dynamic neural network library working well with networks that have dynamic structures that change for every training instance. Written in C++ with bindings in Python.
- Fido – A highly-modular C++ machine learning library for embedded electronics and robotics.
- igraph – General purpose graph library.
- Intel(R) DAAL – A high performance software library developed by Intel and optimized for Intel's architectures. Library provides algorithmic building blocks for all stages of data analytics and allows to process data in batch, online and distributed modes.
- libfm – A generic approach that allows to mimic most factorization models by feature engineering.
- MLDB – The Machine Learning Database is a database designed for machine learning. Send it commands over a RESTful API to store data, explore it using SQL, then train machine learning models and expose them as APIs.
- mlpack – A scalable C++ machine learning library.
- MXNet – Lightweight, Portable, Flexible Distributed/Mobile Deep Learning with Dynamic, Mutation-aware Dataflow Dep Scheduler; for Python, R, Julia, Go, Javascript and more.
- proNet-core – A general-purpose network embedding framework: pair-wise representations optimization Network Edit.
- PyCUDA – Python interface to CUDA
- ROOT – A modular scientific software framework. It provides all the functionalities needed to deal with big data processing, statistical analysis, visualization and storage.
- shark – A fast, modular, feature-rich open-source C++ machine learning library.
- Shogun – The Shogun Machine Learning Toolbox.
- sofia-ml – Suite of fast incremental algorithms.
- Stan – A probabilistic programming language implementing full Bayesian statistical inference with Hamiltonian Monte Carlo sampling.
- Timbl – A software package/C++ library implementing several memory-based learning algorithms, among which IB1-IG, an implementation of k-nearest neighbor classification, and IGTree, a decision-tree approximation of IB1-IG. Commonly used for NLP.
- Vowpal Wabbit (VW) – A fast out-of-core learning system.
- Warp-CTC – A fast parallel implementation of Connectionist Temporal Classification (CTC), on both CPU and GPU.
- ThunderSVM – A fast SVM library on GPUs and CPUs.
- LKYDeepNN – A header-only C++11 Neural Network library. Low dependency, native traditional chinese document.
- xLearn – A high performance, easy-to-use, and scalable machine learning package, which can be used to solve large-scale machine learning problems. xLearn is especially useful for solving machine learning problems on large-scale sparse data, which is very common in Internet services such as online advertisement and recommender systems.
- Featuretools – A library for automated feature engineering. It excels at transforming transactional and relational datasets into feature matrices for machine learning using reusable feature engineering "primitives".
- skynet – A library for learning neural network, has C-interface, net set in JSON. Written in C++ with bindings in Python, C++ and C#.
- Feast – A feature store for the management, discovery, and access of machine learning features. Feast provides a consistent view of feature data for both model training and model serving.
- Hopsworks – An data-intensive platorm for AI with the industry's first open-source feature store. The Hopsworks Feature Store provides both a feature warehouse for training and batch based on Apache Hive and a feature serving database, based on MySQL Cluster, for online applications.
- Polyaxon – A platform for reproducible and scalable machine learning and deep learning.
- sara – C++ Computer Vision Library with easy-to-understand and efficient implementations of computer vision algorithms. [Mozilla Public License version 2.0]