The Race for Faster Machine Learning - Intel Artificial Intelligence Technical Update.

Pawel Gepner, Robert Adamski

Intel Corporation, Pipers Way, Swindon, Wiltshire, SN3 1RJ, United Kingdom
Intel Technology Poland, Al. Jerozolimskie 146C, Warszawa, Poland

email: pawel.gepner@intel.com; robert.adamski@intel.com

Keywords: Machine Learning, Artificial Intelligence, Intel Deep Learning SDK, Intel Xeon Phi Processors

1. Introduction

If we look at any ranking of successful business using the for instance lists of the Fortune 500, or the Global 2000 – We can see a significant difference in the last two decades. Companies have fallen off the list, or risen to the top of the list based on their ability to benefit from data, which drives value to their business. These insights are a competitive advantage to business. The use of data analytics is moving to the mainstream of business and the technologies that are driving it are quickly evolving as more enterprises adopt them.

According to the McKinsey research we can see that in Europe, many of the banks have replaced older statistical-modeling approaches with machine-learning techniques and, in some cases, experienced 10 percent increases in sales of new products, 20 percent savings in capital expenditures, 20 percent increases in cash collections, and 20 percent declines in churn. Is not just banks phenomena it is more wider approach and consequently we will continue to see companies keen to use more analytics to make decisions.

1. Problem statement

Let’s start with the definition of artificial intelligence (AI): this is machines that can sense, learn, reason, act, and adapt to the external world without explicit programming after they have been created. Machine learning is a sub set of AI and uses a set of algorithms or mathematical models that “learn” from data that is they improve their performance based on experience. Under the umbrella of machine learning there are a lot of algorithms. Deep learning is one of them. Deep learning involves multiple layers of data and compute representing neurons in our human brain. The other type of machine learning is more statistical in nature and has been around for some time now.

Whether we are doing deep learning or statistical/other machine learning there are two key steps in machine learning. First data scientist have to train a model to “learn”. This is where they are trying to create a model based on data that is already tagged. Once a model has learned to do a task with high accuracy, data scientists can then deploy that model to make predictions on new data. Scoring is also known as inference.

For training, performance is important and time to train is a critical metric. For scoring, throughput is most important and most IT managers want to see good throughput/TCO. There are some use cases where IT managers don’t consider time to score as a critical component – e.g. when they use low workload timeframes such as overnight to complete their scoring. Nevertheless it is important to know the usage case for the customer.

Key point is that today training runs in the data center and the key metric is time to train. This is count in days/weeks and sometimes hours. Conclusion can happen instantaneously at the edge or in the data center and generally happens instantly. For conclusion throughput/TCO is really critical.

Training of the model is supervised and usually requires expert input. It takes time to train the model. This is an iterative process. The larger the data set that is available for the training results in increased accuracy. While training the model accuracy will increase over time until it ready to be released. Inference (sometimes called Conclusion or Scoring) is unsupervised and the output classification can be fed into a number of different usages including – a dashboard for visualization or a decision tree for automatic decision making. What is the Intel’s strategy and product portfolio to address this market segment will be address during the talk.

2. Description of a problem solution

Intel strategy starts with the silicon innovation. We continue to bring the best platform performance for both single node and multimode benchmarks and workloads. We also continue to refine our roadmap to ensure leadership in this space. Beyond silicon, we are making Intel® Math Kernel Library (MKL) and Intel® Data Analytics Acceleration Library (Daal) libraries optimized for key machine learning primitives. MKL has been widely used by HPC developers and Daal is now open source and become the popular package for data analytics library.

Intel® MKL 2017 introduces the DNN (Deep Neural Networks) domain, which includes functions necessary to accelerate the most popular image recognition topologies, including AlexNet, VGG, GoogleNet and ResNet, on frameworks Caffe, Tensorflow, Torch, Theano. These DNN topologies rely on a number of standard building blocks, or primitives, that operate on data in the form of multidimensional sets called tensors. These primitives include convolution, normalization, activation and inner product functions along with functions necessary to manipulate tensors. Performing computations effectively on Intel architectures requires taking advantage of SIMD instructions via vectorization and of multiple compute cores via threading. Both library MKL and Daal are free.

For the existing frameworks, we are working with the open source community to ensure the frameworks are optimized for Intel CPUs. As this space evolves, we will continue to ensure all the key frameworks are optimized for next generation of Intel CPUs as well.

Intel is not just a hardware company, we have looked at AI at a holistic level and we make libraries and software packages to be optimized for AI frameworks for multimode parallel computing. These libraries are open source and free and we believe that collaboration with active communities make them even more robust and sophisticated.

3. Conclusions

Artificial Intelligence is all around us – we encounter it in our daily tasks such as talk-to-text and photo tagging, and see it contributing to cutting-edge innovations such as precision medicine, injury prediction and autonomous cars. Artificial Intelligence I is the next big revolution in computing and holds the promise to provide insights previously unavailable while also solving the world's biggest challenges. Intel is the partner for Artificial Intelligence today and in the future, and is committed to driving this transformation by offering a complete portfolio to deliver end-to-end Artificial Intelligence solutions. Intel is democratizing Artificial Intelligence innovations by increasing the accessibility of data, tools, training, and intelligent machines, while collaborating across industries to improve society.