SOLUTION BRIEF

Data Analytics & AI
Intel® Optimized Software

AI-Based Document Validation Speeds Business while Reducing Risk

BigData Corp uses Intel software tools to increase inference performance by up to 19.5X[1].

Useful data is scattered throughout the Internet, yet finding and aggregating the right data for a particular business need can be difficult and time consuming. BigData Corp is helping companies of all sizes solve this challenge. The Brazilian company processes hundreds of petabytes of data daily, capturing data from nearly a billion websites—including millions of public databases. The data covers more than a thousand attributes about hundreds of millions of people, companies, and products around the world.

BigData Corp structures the data for efficient analysis and provides easy-to-use APIs that allow its customers to extract real-time insights and integrate them into their existing applications and processes. Companies such as PayPal, Sanofi, and Certisign are using one or more of these tools to tailor their products, grow their customer bases, reduce risk, and improve their operational efficiency. BigData Corp offers the following API options.

- **BigBoost** for rich, targeted information about people and companies.
- **BigMarket** for deep insights into products and the competitive landscape.
- **BigWeb** for highly qualified business leads.
- **BigID** for identity validation to reduce fraud and increase sales conversion rates.

Using AI to Automate Document Validation

As a member of Intel® AI Builders, BigData Corp is taking advantage of the latest advances in AI to help its customers extract even higher value from the available data. One example is the company's document authentication application, which allows banks, credit unions, and other organizations to automate the complex but critical task of ensuring that an official document, such as a driver's license, passport, or foreign ID, is authentic (Figure 1).

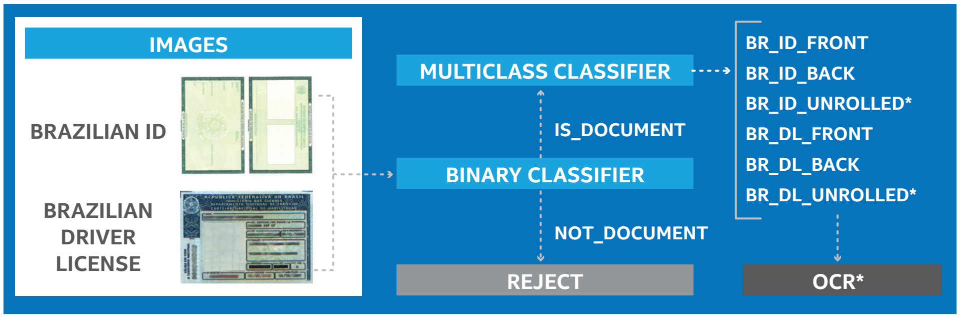


Figure 1. Using the AI-based document validation application, BigData Corp's customers can quickly and automatically 1) identify the type of document among two or more document classes, 2) determine whether the document is authentic, and 3) extract desired information using optical character recognition (OCR).

The customer submits a scanned image of the document using the BigID API. A customized, cloud-hosted AI application is then used to process the image data along with relevant data that already resides within BigData Corp databases. Depending on the specific customer's requirements, the AI application can identify the type of document, determine whether it is authentic, and extract desired information. Results are typically available within a second, which can help customers streamline many types of online transactions, while eliminating the costs and potential human error associated with manual document verification.

Factors that often complicate manual validation, such as changes in age or facial hair in a consumer's ID, are readily handled by the AI application. The application is also adept at detecting subtle image manipulations and other details that are indicative of fraud and are easily missed with traditional, manual processing.

Increasing Performance with Optimized Software

For many companies, especially those in the financial services industry, the simplicity and speed of their online services is key to their success. This makes automated, high-speed document validation a critical capability for select transactions. To improve the speed and throughput of its document validation service without overspending on hardware infrastructure, BigData Corp turned to Intel. The initial goal was to double performance through software optimization.

The document validation application is based on a customized convolutional neural network (CNN) topology that uses Keras* software with a TensorFlow* back end. Working with Intel, BigData Corp engineers upgraded their software stack to replace their off-the-shelf software distributions with Intel® Optimization for TensorFlow* and Intel® Distribution for Python*. These software packages are highly optimized for better performance on Intel® architecture. Many of their performance advantages are derived from using the Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN), which provides optimized algorithms for many of the most popular and compute-intensive operations used in neural network training and inference solutions.

The software optimizations in Intel MKL-DNN help to ensure that each algorithm makes efficient use of the available execution resources in the latest Intel® Xeon® Scalable processors, which can include up to 28 processor cores with two execution threads per core. Memory and cache usage are also optimized, so that required data gets to the cores quickly to avoid lags in processing.

The algorithms in Intel MKL-DNN are optimized to take advantage of Intel® Advanced Vector Extensions (Intel® AVX), which allow a single instruction to be executed simultaneously across multiple data points. Intel has enhanced this technology in successive processor generations to provide increasing levels of parallelism. The latest Intel Xeon Scalable processors support Intel® Advanced Vector Extensions 512 (Intel® AVX-512), which enables simultaneous processing of all the data elements stored in 512-bit vector registers. Optimizing software for this strategy is known as vectorization and can dramatically increase performance for operations that can be parallelized in this way. In addition to using Intel MKL-DNN for many key algorithms, BigData Corp engineers optimized portions of their own code to take advantage of this highly parallel processing capability.

Software Tuning for Even Higher Performance

After upgrading components of their document validation software, Intel and BigData Corp engineers worked together to tune the stack to further enhance performance and efficiency. Computing demands vary for CNN workloads, and there are several parameters in Intel Optimization for TensorFlow that can be tuned to achieve optimal performance. Parameters associated with thread management can be particularly important when running TensorFlow on Intel Xeon Scalable processors, since efficient threading is essential to fully utilize the large numbers of cores and threads. The benefits were substantial, enabling BigData Corp to deliver significantly higher document authentication speeds and capacity with a reduced infrastructure footprint. Since the company deploys its applications on pay-for-use cloud infrastructure, the performance gains are expected to provide direct cost savings in many cases.

During tuning, the engineering team had to consider two different workload types: inference workloads, in which the neural network is used to authenticate documents in production environments, and training workloads, in which BigData Corp engineers train the neural network to handle new document types. The workloads for training are different and much heavier than inference workloads. Inference can be performed in a fraction of a second per document. It typically takes six to eight hours to train the neural network to handle a new document.

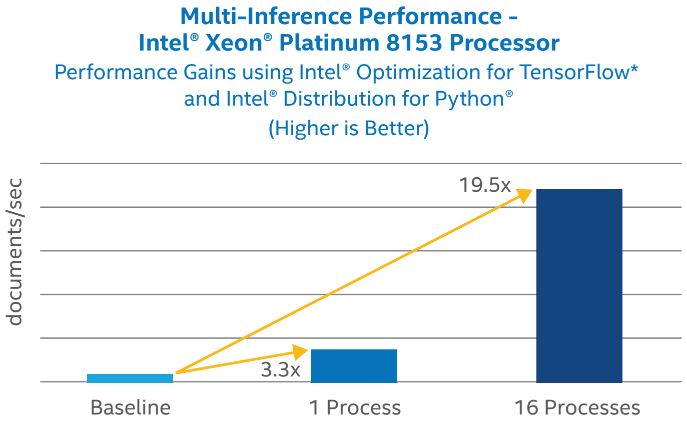
Optimizing the neural network for inference is most important, since the speed of authentication has a direct impact on the online experience provided by BigData Corp's customers. Although training performance is less important, accelerating training runs is useful, as it helps to reduce the time and cost associated with implementing new use cases.

Up to 19.5X[1] Higher Performance for Document Verification (Inference)

To quantify the performance gains, BigData Corp and Intel engineers ran the unoptimized and optimized software on a two-socket server configured with the Intel® Xeon® Platinum 8153 processor. These processors provide 16 cores per processor and support two threads per core, so a two-socket server can handle up to 64 simultaneous execution threads. As the test results show (Figure 2), upgrading to the Intel® software packages and tuning selected parameters in Intel Optimization for TensorFlow increased performance by 3.3X[1] versus the un-optimized baseline software.

Because the new, optimized software is well-threaded, the team was also able to increase the number of processes that run simultaneously on the server (each process corresponds to a single document verification workload). The team found that optimal performance was achieved by running 16 processes simultaneously, which equates to each process running on two physical cores. Using the optimized software and running 16 simultaneous processes increased performance by 19.5X[1] versus the original, unoptimized software running on the same server configuration. These performance gains will provide a major leap in speed and scalability for BigData Corp's document authentication application. It will help them to deliver better service to their customers, while potentially reducing their infrastructure requirements and the associated costs.

Up to 3.6X[1] Higher Performance for Neural



Multi-Inference Performance - Intel® Xeon® Platinum 8153 Processor

Performance Gains using Intel® Optimization for TensorFlow* and Intel® Distribution for Python®

(Higher is Better)

Figure 2. By upgrading to optimized software and tuning selected parameters, the engineering team improved the performance of the BigData Corp document validation application on the Intel® Xeon® Platinum 8153 processor by up to 19.5X[1]

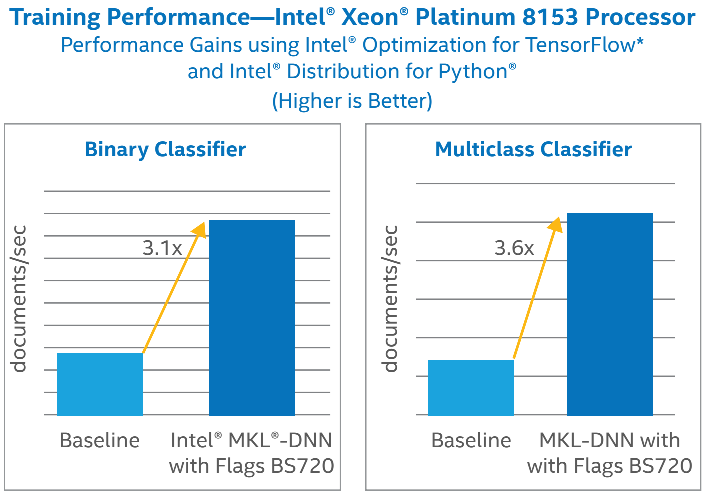
Network Training[2]

Training performance was determined by measuring the number of training images that could be processed per second. For a fixed training data set, this metric is inversely proportional to the total training time. By upgrading to Intel Optimization for TensorFlow and Intel Distribution for Python and then tuning selected parameters, the engineering team was able to increase performance by up to 1.8X[1]. (Figure 3.)

The increased efficiency of the optimized software running on the multicore hardware platform also enabled the team to increase the batch size of the training runs from 16 images when using the baseline software to 720 images when using the optimized software. Batch size refers to how many images are processed during each training iteration (images per batch x the number of batches per training run = total number of training images). Larger batch sizes reduce the number of training iterations required, but also place heavier computational loads on the server platform. There is no simple relationship between batch size and overall time to train, so testing is typically required to find the optimal batch size for a particular workload and server platform.

The team found that optimal performance was achieved with the optimized software when using a batch size of 720. With this configuration, performance was increased by 3.6X[1] versus the original, unoptimized software running on the same server configuration. These gains will potentially save many hours when training the neural network to handle new types of documents.

A Path to Scalable, Cost-Effective



Training Performance—Intel® Xeon® Platinum 8153 Processor

Performance Gains using Intel® Optimization for TensorFlow* and Intel® Distribution for Python®

(Higher is Better)

Binary Classifier

Multiclass Classifier

Figure 3. The software optimization effort delivered up to 3.6X[1] higher performance when training the document validation application to handle new documents.

AI Performance

As demonstrated by the performance tests, using optimized AI software can deliver immediate performance improvements for neural networks, enabling faster time-to-results for inference workloads and reduced training times for training workloads. These gains can be used to improve service levels or to reduce hardware requirements. Since Intel optimizes its software tools, frameworks, and libraries to leverage ongoing advances in each new Intel processor generation, these benefits tend to increase on a regular basis as new processors and platforms become available.

The engineers at BigData Corp are particularly interested in the upcoming Intel® Nervana™ Neural Network processor, which is purpose-built for the demands of deep learning using neural networks. With large amounts of built-in high-speed memory and with high-speed on-chip and off-chip interconnects, this processor is built to deliver a major leap in performance and scalability for neural network performance. It is also designed to give users the flexibility to support all deep learning primitives and to be used seamlessly in conjunction with Intel Xeon Scalable processors. With this addition to the Intel® processor product line, AI developers will have a unified, cost-effective platform that can handle all their application needs, including their largest and most demanding deep learning models.

Join the AI Revolution

The marriage of big data and AI represents one of the greatest technical and business opportunities of our era—perhaps of any era. BigData Corp brings the two technologies together in ways that are readily accessible to companies of all sizes. Many organizations are already realizing the benefits of the company's APIs and big data stores, including their fast, high-volume document validation solution.

Intel's goal is to help BigData Corp—and thousands of other companies—accelerate AI development by providing simpler, faster, and more affordable hardware and software solutions. Businesses need AI tools and systems that integrate easily with their existing applications and can scale cost-effectively to address future needs. Intel architecture-based software tools and server platforms address these needs today and are advancing rapidly to deliver unprecedented scalability and flexibility for ongoing AI innovation.

To learn more, visit Intel AI Builders at: https://builders.intel.com/ai

To learn more about BigData Corp, visit: https://www.bigdatacorp.com.br/

[1] Testing conducted by the Intel Technical Solution Engineering team on August 21, 2018 comparing default TensorFlow* to Intel® Optimization for TensorFlow* with and without Intel® Math Kernel for Deep Neural Networks Library (Intel® MKL-DNN) flag tuning. Tests for all software versions were performed in the Intel® AI DevCloud for AI Builders on a 2-socket server configured with 2 x Intel® Xeon® Platinum 8153 processor (2.00 GHz, 16 cores, 32 threads, 22 MB cache), 376 GB memory, CentOS* 7 operating system, TensorFlow 1.9 (tf.keras) with Intel Math Kernel Library.

[2] BigData Corp offers two versions of its document authentication algorithm. One version supports binary document classification (identifying the document among one of two types prior to authentication). The other supports multiclass use cases (identifying the document among several types prior to authentication). The gains shown are for the multiclass version.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

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.

Performance results are based on testing as of August 21, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

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 information go to www.intel.com/benchmarks.

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. 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. Cost optimization scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction. Check with your system manufacturer or retailer or learn more at intel.com.

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. Notice Revision #20110804.

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].Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.