Hewlett Packard Enterprise | NVIDIA

Transforming financial services with AI technologies

Drive business value with NVIDIA® GPU-accelerated deep learning

Contents

The financial services industry is undergoing one of the largest transformational shifts in decades, driven by the development of new digital products and services, broadening availability of powerful computing solutions, and increased customer adoption of cloud, mobile, and web-based technologies. As ever-expanding regulatory requirements and evolving customer preferences shake up the traditional business models that many financial firms have relied upon for years, there is a race to implement the latest technologies and keep pace with new trends that promise to deliver disruptive competitive advantage and drive new sources of revenue. In this ultra-competitive environment, where market fluctuations are a given and speed is paramount to success, survival is increasingly dependent on rapidly sorting and processing large volumes of complex datasets to gain deep insights that can help firms protect themselves against fraud, streamline operational processes, and better serve their customers.

The global financial crisis of 2008 left both financial organizations and regulators badly shaken and unsure of how to move forward. To help prevent another crisis, regulators ramped-up bank stress test requirements and developed new standards that financial institutions must adhere to in order to control risk, better understand their level of exposure at a given time, and demonstrate capital adequacy. At the same time, customer preferences are rapidly changing and financial institutions must provide a multi-channel experience along with a variety of ways for customers to manage their finances and connect with advisors. These trends are demanding that all financial applications are accelerated and streamlined, from market risk models to collateral management, capital investing, and many other areas. Accomplishing this requires that today's financial firms rely heavily on lightning-quick insights derived from voluminous quantities of data, made possible by massively parallel processing power.

As the financial industry increasingly realizes the impact of faster analytical insights on overall business strategy, artificial intelligence (AI) techniques like machine learning are permeating nearly every industry. Deep learning, the fastest-growing field in machine learning, leverages many-layered deep neural networks (DNNs) to learn levels of representation and abstraction that make sense of data such as images, sound, and text. This technique is showing great promise for automating a variety of operational processes and ushering in disruptive new business models for the industry. Deep learning can effectively combine both structured and unstructured data to rapidly glean new insights, and companies that can integrate this data will be able to leverage real-time data models to achieve a huge competitive differentiation.

However, these newfound capabilities are quickly pushing conventional computing architectures to their limits. With in-memory databases now providing fast access to data, the bottleneck now falls at the computing level as traditional infrastructures struggle to sort and compare massive datasets on the fly. The only way to resolve these bottlenecks is to employ massively parallel computing. Graphics processing units (GPUs) provide the computing power required for many complex financial applications and stretch the limits of what has been possible with traditional computing. This is causing a large-scale adoption of GPU-based parallel computing infrastructure among enterprise data centers, as GPU computing is by far the most efficient form of parallel compute.

With demands for compute performance expanding, the high performance computing (HPC) industry is shifting to a hybrid computing model where GPUs and CPUs work in tandem to perform general-purpose computing tasks. The parallel processing capabilities of GPUs make them more capable of handling large volumes of complex datasets, both structured and unstructured, because the task can be split into thousands of pieces and calculated simultaneously. As the financial industry increasingly adopts new techniques like AI and deep learning, GPU-accelerated computing will become more pervasive and critically important among financial institutions and enterprise data centers to enable new applications that would simply not be feasible without the tremendous compute power of GPUs.

GPUs: The fuel for modern financial services applications

The financial services industry can be cleanly broken into three segments – capital market banking, consumer banking, and insurance. Though it has not been widely recognized, companies in each segment have been adopting GPU technologies and implementing deep learning techniques to fuel many financial services activities like fraud detection, credit risk assessment, natural language processing, consumer sales and marketing, and much more.

Capital market banking

The aftermath of the 2008 recession ushered in new strict regulations for controlling risk and gauging exposure for many banking institutions. Risk models that may have previously been run on a daily or weekly basis now must be completed several times a day, sometimes in near real-time. GPU computing is powering many of these complex risk controls, allowing analysts to calculate many market scenarios simultaneously. Many Tier 1 global investment banks even program their own risk management code in-house, and many institutions are using GPU hardware to speed up their applications and better understand their exposure through a variety of different scenarios.

The same is true for algorithmic trading, a field where computers follow a defined set of rules and instructions to place trade orders at a speed and frequency that far exceeds the capabilities of a human trader. To better predict market dynamics and make trading recommendations, deep learning models powered by GPUs comb through thousands of news articles each day, or scan social media platforms to discern customer sentiment toward a specific brand or company. This information can then be fed into advanced algorithms that can predict with great accuracy the most profitable trade or investment option.

Similar techniques can also be used to monitor for insider threats in the capital market banking sector. In fact, some large capital market banks are already using compliance monitoring techniques to scan employee emails and check for signs of collusion, and employing deep learning models to detect anomalies in trading patterns that would indicate a rogue trader.

Consumer banking

GPU computing has rapidly gained a strong foothold in this industry segment as deep learning progressively takes over the field of fraud detection. There is no denying that the rapid rise in the number of digital transactions, growing amounts of financial data, and the increasing sophistication of fraud techniques have widened the attack surface for financial services companies. And the repercussions of fraud can not only permanently impact a company's reputation, but they also cost the industry billions of dollars per year. Deep learning techniques are now being exploited to improve fraud detection rates with fewer false positives, by quickly isolating anomalies or patterns of behavior that signal abnormal, and possibly fraudulent, card activity.

In mortgage and retail banking, deep learning can be used as an engine for credit risk assessment analysis for mortgages, credit cards, auto, and small business loans. Advanced algorithms can quickly incorporate a variety of factors, such as credit scores, previous defaults, and debt-to-income ratios to help determine a person's creditworthiness for a specific type of loan.

Deep learning algorithms can also be used to improve personalization and customer service. The most classic use case for deep learning in a variety of industries is a version of the "recommendation engine" used by companies like Amazon and Netflix. In financial services, the same technique can be applied for clustering and segmenting specific customer groups by behavior, preferences, and purchase history to allow companies to better serve their customers and sell to "like" groups. Another vast area is voice analytics, which can be applied to call center audio recordings to better inform future customer service decisions, resulting in higher satisfaction and less churn.

Insurance

Insurance companies are quickly realizing the potential of deep learning for automating the claims process, enhancing customer service, and streamlining policy pricing and underwriting. Claims processing is one of the most complex, costly, and time-consuming functions that any insurance company can undertake. However, deep learning algorithms can leverage data from past claims to quickly pinpoint characteristics of a straightforward claim that can be automatically processed, or anomalies that signal a claim should be flagged for human review. This can carry tremendous benefits for both the customer and the insurer. Customers can access information about their claims more quickly and get their claims closed and paid in hours or days versus weeks, which helps to improve overall customer satisfaction. For the insurers, automating claims processing can reduce the overall cost of claims and result in better loss ratios, and free up valuable agent time so they can concentrate more on the complicated claims where human expertise is required.

AI techniques can also improve policy underwriting, by helping insurers better assess each customer's risk in terms of a specific policy and price that policy accordingly. Today, insurance companies are feeding data from social media platforms, historical records, and closed claims into deep learning models to pinpoint characteristics that indicate the level of risk associated with insuring a particular customer. While statistical models have been used for years to evaluate a variety of factors that go into policy pricing, deep learning is dramatically accelerating this process and rapidly replacing these commonly-used techniques.

Driving improvements to core business functions

Deep learning has become so transformative for the financial services industry that organizations are driving improvements to nearly every area of the company's balance sheet. For example, applying deep learning techniques to outward-facing customer sales and marketing activities can have a direct impact on the top line, helping to drive new business, increase personalization, and improve customer satisfaction. Using deep learning to automate operational processes and free up valuable staff time can increase the agility and speed of operations and shrink the bottom line. Many financial companies are seeing the application of these techniques result in direct improvements to every aspect of their core business functions, from policy underwriting, to claims management, pricing investments, and trading desk decisions.

According to a recent report, today's financial sector has the technical potential to automate activities currently consuming up to 43 percent of its workers' time.[1] The same report references an example from the mortgage banking segment, where it's estimated that more sophisticated verification processes for reviewing documents and credit applications could reduce the time that mortgage brokers spend on processing applications from over 90 percent to just 60 percent. From a cost standpoint, it's estimated that deep learning and AI technologies have the potential to enable access to roughly $34-$43 billion per year in cost savings as well as create tremendous new revenue opportunities by 2025.

As deep learning technologies continue to proliferate throughout the industry, financial firms are investing in powerful GPU hardware solutions so they are positioned to capitalize on the speed advantages of GPUs to accelerate these new capabilities. The ability of GPUs to execute many FSI applications much faster than CPUs means workloads that previously would run over the course of a month can now be run overnight. Advanced models and algorithms are arriving at insights faster and helping firms make better, quicker, and more data-driven decisions. These capabilities are truly transformational for the financial services industry.

A crucial first step to any deep learning implementation is to train the system by exposing it to a large group of labelled examples. DNNs require this intense training period before they can be useful and begin to iteratively learn and adapt without being explicitly programmed. GPUs have quickly become the platform of choice for training DNNs, and have been shown to significantly accelerate this process as compared to CPUs. This is because the training process for DNNs involves multiple matrix multiplications, convolution, and other compute-intensive operations that can take advantage of a GPU's massively parallel architecture. Performing these tasks on extremely large datasets can take days or even weeks to run on a single processor; however, offloading these tasks to multiple GPUs and running these operations in parallel can reduce training times from days to just hours. The broadening popularity of GPUs for this process is one of the main drivers behind many of the recent successes in deep learning.

Demystifying deep learning for faster insights across all organizations

Of course, there are lingering challenges that continue to hinder the widespread adoption of AI and deep learning among many enterprises. Recent research found that two thirds of U.S. financial services companies felt they were held back by operations, regulations, budgets, or resource limitations.[2] While the desire to take advantage of these newfound capabilities has resulted in a variable explosion of supporting technology and frameworks, there is still a shortfall of expertise in the deep learning field that persists. NVIDIA® and Hewlett Packard Enterprise (HPE) are working together to demystify deep learning techniques and provide the foundation, blueprint, and expertise that organizations need to leverage new technologies and arrive at faster, data-driven insights.

Leveraging capabilities from their recent acquisition of HPC leader SGI, HPE now offers greater choice for large-scale, dense GPU environments. The purpose-built HPE Apollo server portfolio maximizes performance, scale, and efficiency for AI and deep learning environments, while HPE's deep expertise in technology integration helps customers simplify new AI implementations. These new, optimized GPU compute platforms boost application performance by integrating the latest NVIDIA GPU technologies. HPE's new optimized GPU compute platforms support next-generation NVIDIA® Tesla® GPUs based on the NVIDIA® Volta® architecture and maximize performance, scale, and efficiency for deep learning applications.

The rise in online and electronic transactions, growing sophistication of fraudulent activities, and an increase in compliance regulations have many financial companies looking to deep learning to bolster their fraud detection and prevention efforts. By leveraging next-generation predictive analytics, organizations can now detect and stop fraudulent transactions easier, more accurately, and in real-time across various platforms and datasets. HPE's Fraud Detection and Prevention solution, designed with the help of NVIDIA and Kinetica, is a real-time, AI-enhanced predictive analytics platform that is delivered production-ready to provide immediate deployment and results.

The solution incorporates HPE's most trusted HPC systems, NVIDIA GPU accelerators, and Kinetica's GPU-accelerated analytics database. HPE's Fraud Detection and Prevention solution is built on the industry-leading HPE ProLiant DL380 Gen9 and Gen10 servers, HPE Apollo 2000 Gen10 Servers, or the HPE Apollo 6500 Gen9. All three platforms deliver the highest levels of performance, expandability, reliability, security, serviceability, and availability in a dense form factor. NVIDIA® Tesla® V100 GPUs deliver the compute power needed to fuel today's modern data centers. And Kinetica has architected a scalable in-memory database capability around GPU technology that provides acceleration of traditional data analytics and provides optimized performance for AI algorithms. Together, these three powerhouse technologies provide powerful predictive analytics based on deep learning to combat the complexities of today's fraud, and allow organizations to confidently secure transactions with real-time, deep learning fraud detection.

[1] McKinsey & Company, Where machines could replace humans—and where they can't (yet), July 2016
[2] PricewaterhouseCoopers, 2016 Global Data and Analytics Survey: Big Decisions™, June 2016

The NVIDIA® CUDA® Deep Neural Network library (cuDNN) is a GPU-accelerated library of primitives for DNNs. The cuDNN library makes it easy to achieve state-of-the-art performance with DNNs by providing tuned implementations of routines that arise frequently in DNN applications, such as convolution, pooling, softmax, and neuron activations. Particularly for capital market banks that are writing their own code using these routines, cuDNN allows them to very easily integrate GPUs into their code, and maximize the speed of their hardware. These capabilities allow developers to focus more on training DNNs and developing software applications rather than low-level GPU performance tuning, helping to speed the adoption rate of DNNs across the financial industry. The NVIDIA Collective Communications Library (NCCL) implements multi-GPU and multi-node collective communication primitives that are performance optimized for NVIDIA GPUs. This allows developers of deep learning frameworks and HPC applications to rely on highly optimized, MPI-compatible, and topology-aware routines to take full advantage of all available GPUs within and across multiple nodes.

Partnering to bring disruptive technologies to a broader base of customers

When considering any new technology, all companies eventually face the decision of whether to invest in bringing the technology in-house or rolling out their own solution from the ground up. Regardless if the business has decided to buy or build their own solution, NVIDIA and HPE offer the industry-leading hardware solutions that are optimized for deep learning applications, and tuned perfectly to the open source frameworks that support deep learning projects. NVIDIA considers these frameworks an extension of their software, and has a team of engineers working constantly to optimize the way these frameworks can be run on NVIDIA GPU solutions. Once a performance gap is identified, the issue is quickly resolved, and changes are pushed back into the framework code so they are constantly evolving and improving. HPE's large footprint helps bring solid, optimized GPU hardware to a broader base of customers so they can leverage the compute power of GPUs to improve their offerings and enhance their business. Whether it's for a solo project, a small team, or at-scale, NVIDIA and HPE can help financial organizations easily kick off their deep learning projects, and accelerate the adoption of technologies that provide real-time insights from massive data volumes.

The recently-enhanced partnership between NVIDIA and HPE is geared to help enterprises deploy, manage, and optimize their GPU computing infrastructure to realize the benefits of AI and deep learning for their business more quickly. The collaboration between HPE and NVIDIA will jointly address the GPU technology integration and deep learning expertise challenges commonly experienced by companies as they attempt to adopt new technologies. In addition to delivering a number of global Centers of Excellence (CoE) for benchmarking, code modernization, and proof of concept initiatives, the collaboration will also offer an early access program for Volta-based NVIDIA® Tesla® SXM2 GPU systems that will be made available for select HPE customers in late 2017.

Conclusion

In the fast-moving and highly competitive world of financial services, companies are increasingly turning to GPU-optimized compute solutions and deep learning tools to accelerate financial services applications and increase operational agility. From collateral management, to fraud detection, to insurance claims processing, AI and deep learning techniques are dramatically altering traditional financial services business models and helping firms better manage risk, increase customer satisfaction, and automate a variety of operational processes. HPE and NVIDIA are teaming up to provide the expertise, support, and end-to-end solutions that will rapidly increase the accessibility of deep learning capabilities, and help the financial industry accelerate the conversion of massive datasets into deep insight that can improve nearly every aspect of the business.

Learn more at

nvidia.com/deep-learning-ai

hpe.com/high-performance-computing/deep-learning

a00038863ENW, December 2017