SOLUTION BRIEF

Samsung SDSA Brightics AI Accelerator

Simplify, automate, and accelerate your machine learning projects with Samsung SDS America, NetApp AI, and Insight CDCT.

NetApp | SAMSUNG SDS | Brightics AI ACCELERATOR

Challenges in data science and AI workflows

Not all data points are created equal. After a data point is labeled, labeling of similar data points costs the same amount of human effort but provides only marginal extra information to the AI learning process. In fact, we find that the first 10% to 20% of data points provide (nearly) all the information that's present in the dataset. The other 80% to 90% don't offer any significant additional information. They provide only additional work. It's the well-known law of diminishing returns applied to AI projects.

As another challenge, often teams want to speed up training by using multiple GPUs for a single task. Data scientists and AI engineers then quickly discover that scaling a training process from one to several GPUs inside a single server is challenging and is accomplished only by scaling to many GPUs over several servers. It's not clear whether the orchestration and IT work that this effort requires pays off in the end—either in hardware cost, human cost, or total project duration.

Surpassing the competition is yet another challenge. In virtually all use cases, AI is a fiercely competitive field that's evolving at great speed. Getting to market faster than the competition can often be the deciding factor between commercial success or failure. The other side of the equation is the effort that's required to provide the model. This effort is dominated by coding complex tasks, figuring out the right features, selecting the best model, and tuning the hyperparameters of the learning algorithm to the task. Most of the time, these tasks are performed through a clever combination of trial and error and human intelligence. However, automating these processes is possible and frees up the scientists. Feature engineering, for instance, is the critical step that identifies and defines the most informative parts of the dataset. Usually the result of several months of conversation between data scientists and domain experts, the important features of a dataset can be discovered by specialized AI methods and therefore be automated.

Throughout the process, the data must be stored piece by piece and be made available at the fastest possible speed to the various hardware resources (CPUs and GPUs). Therefore, it's critical to have an integrated hardware-software pipeline that's designed for AI processing.

Key benefits

Improve productivity

- Eliminate 80% of manual data-labeling effort by using the autoLabel feature.
- Reduce 80% of the work to achieve state-of-the-art artificial intelligence (AI) model accuracy by exploiting distributed automated machine learning (AutoML).
- Improve time to market by using many GPUs simultaneously on a single task.
- Save on infrastructure costs by reducing total GPU-hours.

Reduce complexity with an easy-to-use platform

- Simplify projects with low-code or no-code needed to train and to deploy AI models at scale.
- Orchestrate complex training tasks and compare experiments easily through a single interface.
- Focus on data science rather than on infrastructure orchestration, development, and maintenance.

Deliver performance with flexibility

- Jump-start AI with fully automated feature engineering, model selection, and hyperparameter tuning.
- Fine-tune AI by running many parameter experiments simultaneously and intelligently comparing them.

Build an integrated, end-to-end AI toolchain

- Switch between workloads and resources easily without moving data.
- Build enterprise ML that is portable between on-premises and cloud environments.
- Collaborate, compare, and reproduce results with interactive workspaces, dashboards, dataset organization, and experiment tracking and visualization.

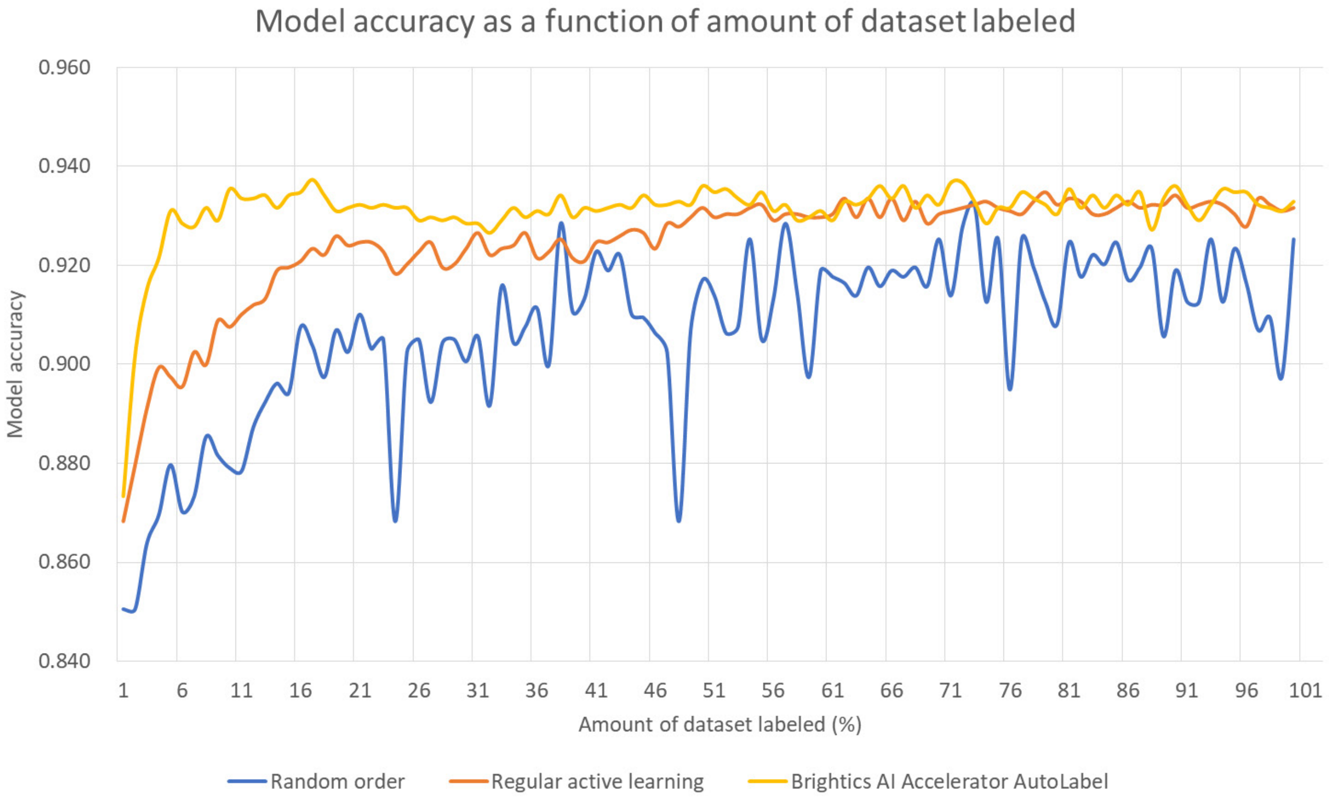

Figure 1: AutoLabel achieves a significantly faster convergence to the highest possible accuracy.

The solution

Now you can fully streamline the AI workflow by labeling only the data that provides valuable information. By creating and selecting the right features, by tuning the best hyperparameters, and by using an arbitrarily large hardware infrastructure to execute it, you can easily label the most valuable data. And you can do it all in an easy-to-use environment that powered by Samsung SDSA Brightics AI Accelerator, NetApp Trident, and NetApp® AFF A-Series. With Trident, you get a dynamic persistent storage orchestrator for containers, and AFF A-Series gives you cloud-connected flash storage that's powered by NetApp ONTAP® software. You can use models at inference within the platform or exported to the edge.

The Brightics AI Accelerator is built for data scientists by data scientists. It's an easy-to-use, automated ML platform to train and to deploy AI models at scale. Because the platform uses Kubernetes containerization, models are portable between the cloud and on-premises data centers. Brightics AI Accelerator gives you a unified interface for orchestrating large, distributed clusters to train deep learning (DL) models by using a TensorFlow, Keras, or PyTorch framework. You can also use AutoML with scikit-learn to unify AI workloads, to simplify deployment, and to accelerate return on investment.

Improve productivity

It's well known that about 80% of the total workload of a data science or AI project lies in producing a clean, labeled dataset. Our primary feature, autoLabel, is dedicated to minimizing this effort.

The autoLabel feature in Samsung SDSA Brightics AI Accelerator is a human-in-the-loop interactive process in which the human labelers label the data in relatively small batches. After each batch, an AI system learns the patterns and sorts the remaining unlabeled data in order of confidence. The data that the model is least confident about forms the next batch. After 10% to 20% of the data is labeled, the confidence is typically so high that no further labeling is necessary, and the remaining 80% to 90% of the dataset can be labeled automatically by the system (Figure 1).

Figure 1 shows that as the dataset is being labeled 1% at a time, the autoLabel feature keeps the unlabeled data sorted; therefore, the informative data is labeled first. The graph shows that this approach achieves a significantly faster convergence to the highest possible accuracy—6% in this example. Accuracy is much lower in the normal labeling of data in a random order and in the regular active learning that's found in other AI platforms.

Another time-consuming part of the AI workflow is the design of synthetic features and the selection of the right features for modeling. Known together as feature engineering, this step often involves multiple data scientists and domain experts. Brightics AI Accelerator has fully automated this step by intelligently designing and selecting features. Although these features can't fully replace human domain expertise, they can reduce the feature engineering effort by 70%.

Reduce complexity with an easy-to-use platform

A fundamental barrier to AI progress is the time that it takes to train an AI model. After the considerable effort to produce a clean dataset and obtain its features, model training consumes very little human effort but quite a lot of compute resources and time. One way to cut the computation time is to use multiple computers, or multiple GPUs, simultaneously for the same task. Like any teamwork, this approach requires managing the cluster of hardware resources. Significant budgets and time are spent on designing, developing, and maintaining the software infrastructure that AI scientists need to build and to deploy AI models in production.

With the Samsung SDSA Brightics AI Accelerator solution, your AI team can devote 100% of their attention to science rather than to software infrastructure development and maintenance. The orchestration, management, and cleanup of the infrastructure of hundreds of GPUs are fully automated, and the solution is as easy to use and to deploy as it is to do work on your desktop.

DL training routines demand massive amounts of compute power. Faster training can cut down on overall compute costs while accelerating AI innovation and productivity. With Brightics AI Accelerator, your team can also migrate their models from the desktop to the cloud or to the data center with no code changes. Your team can exploit large clusters of resources that save both time and money in training AI models. And this increased speed enables automated data labeling, feature engineering, model selection, model training, and hyperparameter search with native Kubernetes cluster orchestration and meta-scheduling.

In one example, a DistilBERT natural language model was trained on 8, 32 and 64 GPUs. By using Brightics AI Accelerator, this 8-fold increase in resources resulted in 13.7 times faster training. Figure 2 displays the amount of time that it took to go through three training epochs for a DistilBERT model by using several different methods. With a normal 8-GPU machine, the time was about 400 minutes. The enhanced inter-GPU communication of Brightics AI Accelerator lowered this training time to about 200 minutes on the same hardware. Going to 32 and 64 GPUs further lowered this time to 66 minutes and 32 minutes, respectively. In total, this approach sped up training by 13.7 times as compared with a linear scaling on ordinary infrastructure.

As this example shows, it's possible to scale your resources super-linearly and thus reduce the total number of GPU-hours, saving hard costs on your compute infrastructure. If you're interested in the details, read this LinkedIn article.

The Brightics AI Accelerator platform is easy to use in any scenario. If you have code already, you can lift and shift it to Brightics AI Accelerator unchanged. If you have a new task, you can solve it with very few lines of code because the platform automates all the necessary steps. But all this automation doesn't prevent you from having full flexibility to dive into the detail and to fine-tune the last few percentage points of accuracy. You can still use the AI toolchains that you're accustomed to, such as PyTorch, TensorFlow, Keras, and Python. The platform thus provides you with maximum comfort without sacrificing any details.

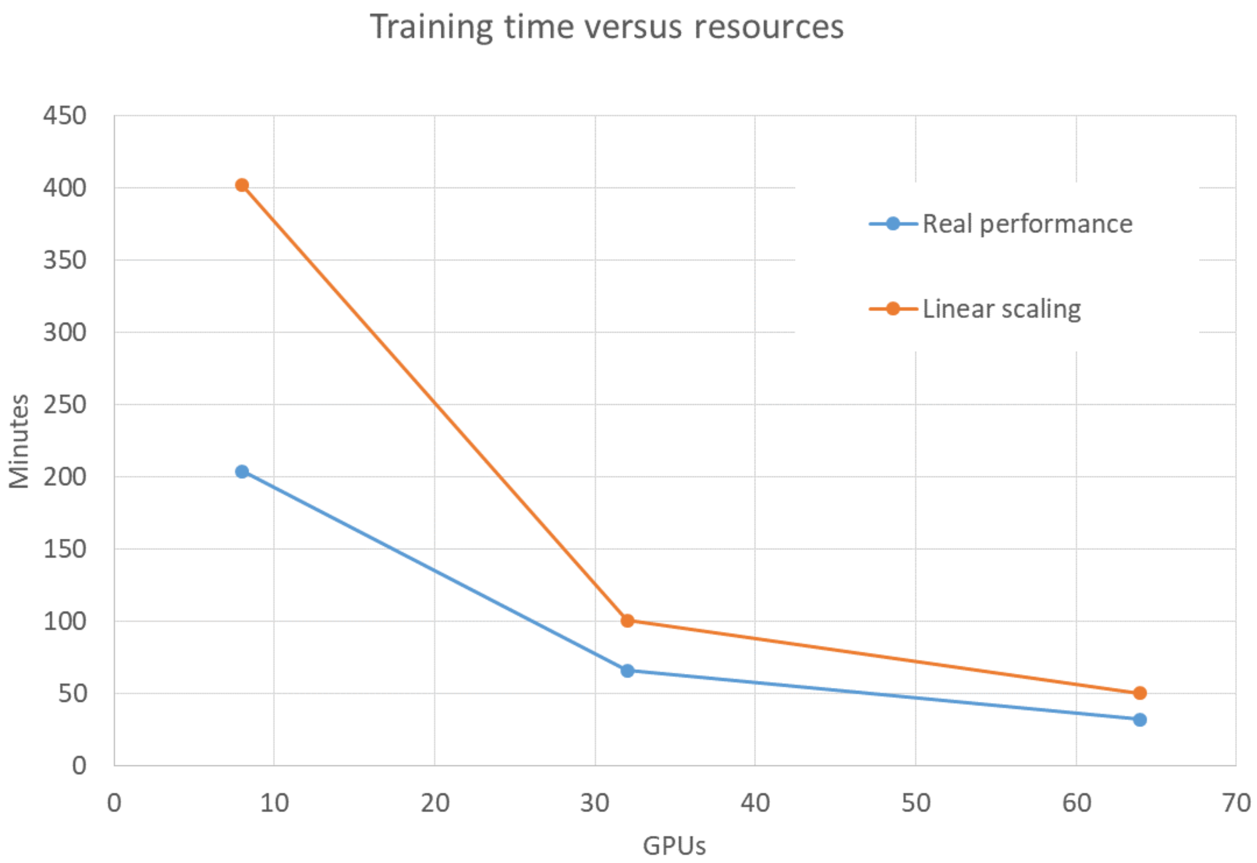

Figure 2: With Brightics AI Accelerator and 64 GPUs, a DistilBERT model completed three training epochs 13.7 times faster than with 8 GPUs.

Investing in state-of-the-art compute also demands state-of-the-art tools that simplify complex tasks and facilitate collaboration. You need a high-performance AI platform that keeps up with your most demanding DL training workloads. Brightics AI Accelerator integrates NetApp Trident APIs to speed the flow of data and to facilitate collaboration among AI teams. It also simultaneously provisions data all the way up to Kubernetes containers with data refresh and code versioning to make it easier to train and to deploy AI models.

Deliver performance with flexibility

The rapid pace of AI innovation makes it challenging to design and to train accurate AI models. With Brightics AI Accelerator, you can eliminate guesswork and get started faster by automating the initial tasks of labeling data, designing features, and finding the right hyperparameters. These automated initial steps get you most of the way to optimal accuracy.

Then you can fine-tune your model by running multiple experiments in parallel before you scale up to exploit large, distributed compute clusters.

You can run 10 to 100 single-server experiments simultaneously or train one model on up to 512 CPUs or GPUs to identify the parameters that you need, then you start pushing the envelope of the state of the art. Brightics AI Accelerator reduces all this server orchestration complexity down to a single line of code through a single Jupyter Notebook, PyCharm integrated development environment (IDE), or CLI.

Build an integrated, end-to-end AI toolchain

The Brightics AI Accelerator is a Kubernetes, container-based application that offers model portability between the cloud and on-premises data centers. You get a unified interface to orchestrate large, distributed clusters to train DL models by using a PyTorch, TensorFlow, or Keras framework, and you can use AutoML with scikit-learn. It's easy to build an integrated AI toolchain that spans from the core to the cloud to the edge.

With this integration, now your organization can eliminate knowledge silos of science that either are underused or inhibit AI innovation. Brightics AI

Accelerator provides a universal AI infrastructure solution that's certified to work with NetApp Trident storage volumes to consolidate analytics, training, and inference onto one platform. Facilitate collaboration, reproduction of results, and innovation with interactive workspaces, dashboards, dataset organization, and experiment tracking and visualization.

Collaboration also benefits from this unified interface. Data science and AI are not driven by individuals but by teams of diverse professionals who work hand in hand. With Brightics AI Accelerator, you get seamless integration of collaboration features with advanced AI so that your teams can work together on a single project. Your data scientists organize the data, AI engineers develop the code and run jobs on the infrastructure, and other scientists track the experiments and visualize the results—all on one platform.

Samsung SDSA, NetApp AI, and Insight CDCT: Driving innovation together

Samsung, NetApp, and Insight CDCT solutions can help reduce the time that it takes you to get from data to insights, actions, and outcomes. And you can accelerate your journey to AI with a data fabric that makes the right data available in the right place at the right cost.

At the heart of the Brightics AI Accelerator is an intelligent and flexible automation platform that reduces complex AI model training orchestration down to a single line of code. Brightics AI Accelerator is use-case agnostic and covers training of all AI models by applying AutoML to tabular, comma-separated values (CSV), time-series, image, or natural language data. It enables analytics; computer vision classification, detection, and segmentation; and natural language processing (NLP) use cases. The distributed clustering technology powers AI models to run at inference efficiently in the cloud and on premises.

Our partners also enable Brightics AI Accelerator–trained models to run fast and accurately on constrained edge devices, such as field-programmable gate arrays (FPGAs) or Raspberry Pi systems. The Brightics AI Accelerator platform can be applied across industry verticals, including but not limited to healthcare, retail, automotive, aerospace, communications, finance, marketing, manufacturing, any industry that uses Internet of Things (IoT), and academia for fundamental science.

By using automated model selection, feature synthesis, and hyperparameter search optimization, Brightics AI Accelerator AutoML software automates and accelerates model training on tabular data. AutoML with synthetic feature generation exploits up to 512 CPU cores simultaneously to produce a model in less than 1 hour, versus 2 months by using manual feature engineering methods. The Brightics AI Accelerator platform also automates ML and model deployment. When model training is complete, you can simply specify the number of GPUs and point to an address to automatically deploy the model and perform inference calculations.

Brightics AI Accelerator AutoDL software automates and accelerates deep learning model training by using data-parallel, distributed synchronous Horovod Ring-AllReduce, Keras, TensorFlow, and PyTorch frameworks with minimal code. AutoDL exploits up to 512 GPUs per training job to produce a model in 1 hour versus in 3 weeks by using traditional methods. AutoDL eliminates installation of any software or configuration per job and offers a painless experience in provisioning, running, monitoring, and cleaning up jobs. For computer vision projects, AutoDL uses automated training and grid search–based hyperparameter optimization to shrink the entire lifecycle from as many as 9 months to only a couple weeks in a properly sized cluster. AutoDL also includes automated transfer learning for image data that's fitted to all models in the model zoo with a hyperparameter search.

With NetApp AI, you can build a seamless AI pipeline no matter where your data lives—or where it moves to—from the edge to the core to the cloud. Powered by NetApp ONTAP data management software, AFF A-Series storage systems deliver industry-leading performance, superior flexibility, and best-in-class data services and cloud integration to help you accelerate, manage, and protect your business-critical data in the hybrid cloud. To efficiently orchestrate, provision, and deprovision persistent volumes in Kubernetes, deploy NetApp Trident. And the NetApp AI Control Plane, a full-stack AI data and experiment management solution, provides extreme scalability, streamlined deployment, and nonstop data availability—when and where you need it.

Insight CDCT offers three Research & Innovation Hubs to streamline the assessment process, giving the data and insights needed to move ahead with technology initiatives and to future-proof your business. Insight CDCT continually identifies, builds out, and tests innovative solutions to provide fresh insights into how to effectively address your most complex business challenges.

Solution components

- Samsung SDSA Brightics AI Accelerator software platform
- NetApp AFF A-Series storage systems with ONTAP 9
- NetApp Trident
- NetApp AI Control Plane

About Samsung SDS America

Samsung SDS is a global leader in enterprise AI, digital transformation, AI transformation, and innovation solutions. Scientific consulting services from Samsung SDS America can help you transform your business and customer engagement workflows to capitalize on the promise and potential of AI. Learn more at www.samsungsds.com and Samsung SDS AI YouTube channel.

About Insight CDCT

Insight CDCT helps clients modernize and secure critical platforms to transform IT. They believe data is a key driver, hybrid models are accelerators, and secure networks are well integrated. Their end-to-end services empower companies to effectively leverage technology solutions to overcome challenges, support growth and innovation, reduce risk, and transform the business. Learn more at: insightCDCT.com.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services and applications to the right people—anytime, anywhere. www.netapp.com