GUIDEBOOK
PROGRAM: ANALYTICS
DOCUMENT NUMBER: T147

GUIDEBOOK
DEEP LEARNING ON AWS

ANALYSTS

Daniel Elman

Nucleus Research, Inc.
100 State Street, Boston, MA, 02109
+1 (617) 720-2000
www.nucleusresearch.com

THE BOTTOM LINE

Deep learning remains one of the hottest topics in the area of artificial intelligence (AI) and is rapidly moving from academia and theory to operationalized, value-add workloads. Advances in data infrastructure and compute power, as well as new classes of neural networks have helped make deep learning feasible for modern businesses to leverage, while it was still generally a pipe dream as little as one year ago. Partnering with cloud vendors like Amazon Web Services (AWS) and taking advantage of cloud-hosted machine learning services like Amazon SageMaker has been key for companies looking to accelerate deep learning projects from concept to production. In the past where the infrastructure wasn't as advanced and machine learning services were immature or not yet available, organizations without the budgets and expertise to grow AI systems from the ground up internally were left on the outside looking in as larger companies with deeper pockets and developer benches researched and implemented these AI capabilities.

To understand the state of deep learning adoption and usage today, as well as how it has changed since last year's report (Nucleus Research *s180 – Guidebook: TensorFlow on AWS* –November 2018), Nucleus conducted interviews and analyzed the experiences across 316 unique projects. For the second time in two years, the number of deep learning projects in production more than doubled. We found that 96 percent of deep learning is running in a cloud environment, with TensorFlow being the most popular framework, being used in 74 percent of deep learning projects. PyTorch was also used in 43 percent of projects (please note that most projects leverage multiple frameworks; MXNet, Keras, and Caffe2 also appeared, virtually always in conjunction with TensorFlow, PyTorch, or both), a significant increase in adoption from last year. Of the 316 total projects, only 9 percent were built with a singular framework. Most notably, of the cloud-hosted deep learning projects, 89 percent are running on AWS; a key driver of this is the breadth of framework choices on AWS along with its own continued investment in existing and new services.

We also found that 85 percent of cloud-based TensorFlow workloads are running on AWS, and 83 percent of cloud-based PyTorch projects are on AWS. Last year, about a third of the interviewed customers were either considering or using SageMaker, Amazon's managed service for building, training, deploying, and orchestrating deep learning models at scale; of the interviewed users this year, 63 percent of Amazon customers had begun using SageMaker.

THE SITUATION

Machine learning encompasses the technology where a computer analyzes data to "learn" from experience without human involvement. Deep learning is a subset of machine learning; in machine learning, the computer is given data with a designated set of features to analyze, whereas in deep learning the computer is presented with unstructured data like text, audio, or video, and it determines by itself which features are relevant to the analysis. Put simply, the computer is provided with pairs of sample inputs and the corresponding outputs and is able to work backwards to find what operations are necessary to transform the input to the output.

Modern deep learning models require massive amounts of compute and storage, making it prohibitive for most organizations to build these systems themselves. Thus, as we found throughout the course of this research, companies overwhelmingly look to leverage the cloud for deep learning projects. This approach allows businesses to buy the amount of data storage and compute power they need for the projects without having to purchase, configure, and maintain the infrastructure internally, producing significant cost savings over time.

With a diverse and quickly growing user base due to the hype and potential of AI, the technology landscape changes quickly. New tools and methodologies are constantly becoming available. Therefore, enabling users to develop on the platform with maximum flexibility is key to long term success in the cloud-based machine learning market. Simply put, cloud platforms need to support the myriad tools and development frameworks that are in use today and tomorrow, with the requisite security and availability to adhere to data handling and privacy regulations.

This is the third consecutive year of conducting this study, and over these three years we have seen transformative changes in model capabilities, compute power, and developer tools enabling new and exciting results. In the first year it was difficult to find organizations that had moved beyond preliminary development and proof-of-concept projects with deep learning. In 2018 we saw strong progress to this end with 14 percent of projects being classified as in production, handling live data. This year brought another leap forward to this effect with organizations ranging from 20-person startup to Fortune 100 global enterprises deploying deep learning to production, with 38 percent of percent of projects in production. Other aspects discussed in the interviews include:

- The goals and motivations behind the deep learning projects
- The deployment strategy and associated benefits
- The development frameworks, methods, and other tools being used
- The relative strengths and weaknesses of different models and frameworks

- The number of people involved and their respective roles on project teams

In total, Nucleus conducted in-depth interviews with 32 deep learning experts, many of whom were responsible for multiple projects concurrently, representing 316 unique projects.

DEEP LEARNING IN THE CLOUD

Production-scale deep learning workloads involve processing thousands or millions of example data to train the model. This is massively computationally expensive, especially for complex input data like images or video, and most organizations cannot afford to build and maintain high-performance computing systems with parallelized CPUs or GPUs to perform the calculations. As a result, organizations look to the cloud to access the resources and infrastructure they need. This year we found that 96 percent of deep learning projects are running in a cloud environment—this mirrors the finding last year, but with 177 projects in 2018 growing to 316 in 2019 it still demonstrates strong customer momentum to the cloud for deep learning. Of workloads that are in production on live data, 98 percent are in the cloud. For organizations that aren't fully in the cloud, a common deployment strategy involves developing a small-scale model on hardware on-site before migrating to the cloud for production.

BENEFITS OF THE CLOUD

With 96 percent of deep learning projects running in the cloud, clearly customers recognize the value of the approach. We asked respondents to identify the main benefits of running deep learning in the cloud. The responses clustered around three key themes:

- Cost savings from avoided hardware, personnel, and energy costs. This was the most common response, cited by every interviewee. Deep learning requires massive amounts of compute; building and maintaining hardware systems that can perform deep learning at scale requires dedicated IT professionals; physically running the hardware to train deep learning models consumes thousands of hours of CPU and GPU time—the electricity cost alone is often prohibitive. With the cloud, users pay for the resources they use without the associated costs.
- Ability to collaborate and work in distributed teams. Models deployed in the cloud can be accessible to all permissioned users, regardless of physical location. This speed up model development, especially across remote teams that are becoming increasingly common.

- Ability to leverage supplemental platform capabilities and tools. Security and availability were the most commonly mentioned aspects. Properly configured cloud systems benefit from the security investments of both the customer and the cloud provider. Additionally, the ability to run models on local data centers to stay in compliance with data protection laws like GDPR is important. Along with storage and compute, cloud vendors offer tools and platform capabilities to improve the developer experience. Tools like Amazon SageMaker provide great value for the cloud customers with fully-managed end-to-end machine learning workflow - from cleaning the data to training, building and deploying models.

DEEP LEARNING IS REAL

Last year demonstrated a step forward for the state of machine learning with 14 percent of projects in production. We saw a similar leap forward in 2019 with 38 percent of the 316 deep learning projects being classified as in production. 89 percent of deep learning projects in production are running on AWS. Seventy six percent of the projects in production leverage TensorFlow and 28 percent of projects use PyTorch. Keras and Apache MXNet were also seen in production settings as most projects have components built with multiple frameworks. Only 9 percent of projects were built with just one framework. As companies recognized that deep learning and other AI capabilities are reaching a level of maturity that allows them to deliver legitimate business value, they've scrambled to implement "low-hanging fruit," or simpler use cases that are well-demonstrated and quick to implement. Common examples include voice interfaces for websites and applications and recommendation engines for online shopping sites.

Companies are still exploring more complex deep learning applications as well, but many are still in testing. The results of this year's study suggests that companies expanded their overall deep learning investments from last year, continuing progress on more complicated, multi-year aspirational projects while adding quickly-implemented, value-add applications like recommendations, sentiment analysis in chat bots, and voice interfaces, to stay abreast of market trends and demonstrate the continued value and viability of deep learning in real-world usage.

DEEP LEARNING ON AWS

Nucleus found that the primary reasons for choosing AWS—the breadth of platform capabilities, the relationship with Amazon, and AWS' continued investment in deep learning services—remain unchanged since last year.

BREADTH OF AMAZON CAPABILITIES

Amazon supports the deep learning process from end-to-end, so there is no need to piece together best-of-breed components from disparate vendors. Customers can store data, build and deploy models, and create applications that leverage model outputs all on the platform. Customers have the flexibility to select specialized hardware optimized for their workloads; for example, they can access powerful GPU instances optimized for deep learning, like Amazon EC2 P3 and G4, on-demand. Amazon has regional data centers around the world, so customers can localize their data and operations as needed and comply with regional data sharing regulations. Platform-level security adds another layer of protection to the data and applications. Other integrated tools and services on the platform like containerization services lend themselves well to large-scale projects. Users said:

- *"Other cloud providers cannot match the maturity and completeness of the AWS platform. We get much more than technology from the partnership; they have the expertise and additional services like S3 and CloudFormation to help us bring technology projects from concept to production."*
- *"We are constantly evaluating other cloud vendors, but so far nobody comes close to AWS. Nobody else provides comparable value in cloud services and that's without mentioning the best-in-class platform security and availability."*

User profile – Biotechnology company

A biotechnology company is using deep learning to develop production-grade classifiers for predicting cancer from genomic data. Since it is still in the research phase, it is building models to test different hypotheses in parallel, so the flexibility to quickly implement and test new ideas is highly important.

There are four different approaches being tested in parallel; two are built on TensorFlow, one on PyTorch, and the other was developed with entirely custom code. The frameworks were chosen strategically—TensorFlow was chosen for two projects for the pre-built estimators and loss functions; PyTorch was chosen for one because the approach was inspired by a project in academia based on PyTorch.

The company chose AWS for its deep learning projects for a number of reasons. Amazon offers different compute options that are best suited for different use cases. Deep learning on genomic data requires extensive pre-processing, so the company needed machines optimized for memory to pre-process the training data, and GPU-accelerated machines to train the model. Working with petabytes of training genomic data, the cloud partner's ability to support a project at that scale was critical. Since the projects use different frameworks, it was important that the cloud provider supported multiple frameworks as first-class citizens on its infrastructure without requiring extensive setup. Amazon was a natural fit for the

performance it offers, even at petabyte-scale, and the flexibility of supported machines and its work to optimize all frameworks to perform well.

RELATIONSHIP WITH AMAZON

Interviewed customers also referenced their relationship with AWS as a driving factor behind the business decision. Many customers were already using AWS in other areas of the business and grew their investment to include deep learning. Deep learning requires a lot of data, and organizations already trust AWS with their data. Amazon also provides customers with the technology and best practices to complete their projects without locking them into any particular solution. Customers were quick to emphasize the flexibility of the platform—they can use the frameworks and libraries they find most comfortable without worrying that they are incompatible with AWS technology. One customer referenced the fact that Amazon is interested in studies like this as a demonstration of its commitment to maximizing customer value. Other users said:

- *"We built our entire business on AWS, so it would take a lot to motivate us to choose another provider. Our AWS [architecture] has grown with us, and the support has been solid every step of the way. We really feel like they're invested in our success which is key to any long-term partnership."*
- *"With Amazon, we have access to customer use cases and industry expertise for our machine learning projects. The support team was instrumental to our efforts—they were able to answer all of our questions about frameworks, model choice, and infrastructure requirements."*

User profile – Digital Media and Mobile Game Company

A digital media company that designs and sells mobile games uses deep learning to balance game difficulty with the projected revenue generated. It does this by using deep reinforcement learning to train a bot to play each game. Then it can monitor the bot's performance to estimate the difficulty of each level which it then uses to predict how many users will quit playing at each level. Of course, when a user stops playing, they stop generating ad revenue for the game vendor, so the company uses this process to balance game difficulty with projected revenue.

AWS has been the company's cloud provider since 2008, so it was a no-brainer to use AWS for its deep learning. Both projects are built on PyTorch—PyTorch is the industry standard for deep reinforcement learning, and much of the current research is published with an accompanying PyTorch implementation.

The company used SageMaker to deploy the model responsible for predicting player attrition into production. It was chosen for because it gives the project leadership a "birds' eye view" of the model and affords project leadership a centralized location to view and control all the models in the deployment. With technology like SageMaker combined with the flexible compute resources offered by Amazon, the interviewed expert said, "it would take a compelling business case to persuade us to leave AWS and so far no other cloud provider has been able to offer us the machine learning-specific tools along with storage and compute for a [better value]."

AWS INVESTMENT IN DEEP LEARNING

Customers know that Amazon is developing and using its own deep learning technology. Deep learning experts referenced the ongoing improvements to documentation, framework support, and cloud services like Amazon SageMaker as primary factors in choosing AWS over other cloud providers. SageMaker became available in 2017; it is a fully-managed cloud service that covers the entire machine learning workflow - from building, training and deploying machine learning models.

SageMaker adoption is growing fast as developers realize the how it can reduce complexity and accelerate model deployment. Last year, approximately one third of the respondents were using or exploring the use of SageMaker to automate aspects of their deep learning projects; this year that figure nearly doubled with 63 percent of customers using or considering SageMaker. In the course of this study, we found customers who were migrating their homegrown TensorFlow deployments to a managed service on AWS via SageMaker, as well as customers who built their systems from the ground up entirely with SageMaker. Users said:

- *"We don't need to source dedicated hardware to run large-scale deep learning projects. Without AWS and specifically SageMaker, we would need to buy hardware, train the model locally, then store and host the model on an internal server so it would be accessible when I want it for forecasting. Just to get started, this could take weeks and comes with a ton of added costs for hardware, electricity, and personnel, not to mention all the lost time from building capabilities internally that are pre-built on SageMaker."*
- *"We focus on implementing conversational machine learning for our clients and SageMaker streamlines model building. On average we have five projects in progress for a client on a given week. Since starting to use SageMaker last year, we found that we save about two hours per project from automating distributed model training."*

User profile – Enterprise Software Company

A global enterprise software company that produces applications primarily for sales and service teams adopted Amazon SageMaker to manage its TensorFlow deployment. Its deep learning efforts are mainly around sentiment analysis and classifying customer interactions in order to effectively understand how different types of outreach affect the customer's likelihood to churn or buy again.

As a large technology company, it has been ahead of the curve with its deep learning efforts compared to the greater market. Before SageMaker, it custom built the bulk of its TensorFlow-based deep learning infrastructure. At the start of the year, it decided to migrate the self-managed TensorFlow deployment to SageMaker where it could be managed as-a-service. While the effort is ongoing and not yet fully complete, the organization has been able to reassign 3 FTEs so far that were primarily responsible for managing the TensorFlow ecosystem. Additionally, the speed of training models is dramatically increased since SageMaker automatically distributes the compute load across multiple CPUs or GPUs in parallel. The company reported that deploying a new model with SageMaker takes less than 50 percent of the time needed to do it in a self-managed environment.

User profile – Application Development Company

An application development company that specializes in creating voice-integrated games playable on Amazon Alexa, the smart speaker, built a deep learning project entirely on Amazon SageMaker that recommends games to keep user engagement high. The system uses data from previous games played on the system to predictively recommend other similar games to the user.

The company built its business on the AWS platform, so it chose to leverage SageMaker for the native integration with existing architecture on AWS, particularly Amazon S3 and AWS Lambda for accessing stored data and serverless compute. Additionally, since the company was already on AWS, user permissions and devops procedures had already been formalized, allowing them to avoid duplicating that effort. The parallelized model training and control-level UI made training and evaluating much faster than it would be manually. The customer estimates that using SageMaker makes model training three times faster and model deployment four times faster than manually managing the system.

CONCLUSION

Last year we saw deep learning projects moving *en masse* out of the classroom and into the boardroom as advances in machine learning education, compute technology, cloud infrastructure, and suitable datasets helped make deep learning commercially feasible. This year, the trend continued with more and more companies looking to implement deep learning at scale to solve real-world problems. For the second time in two years, the number of deep learning projects in production more than doubled, primarily because of the same key factors as last year

- Increased availability of cloud infrastructure and services from vendors like AWS to support data-heavy, compute-intensive processes like deep learning.
- Advances to the state-of-the-art in deep learning with improved techniques, network architectures, and datasets that make neural networks more accurate and capable.
- Continued investment in the community to share experience and enable other deep learning researchers through online forums and documentation, open source libraries and frameworks, and cloud offerings like pre-built models and specialized hardware optimized for machine learning.

As the initial cost to explore deep learning decreases, we see more and more businesses looking to join the fray. Rather than looking to re-invent the wheel, the most efficient strategy to this end is to partner with a cloud vendor that has the infrastructure, expertise, and additional services to bring deep learning from concept to completion. From our analysis, we found that Amazon's reputation as the most mature and sophisticated enterprise cloud technology provider along with its field-specific investments in machine learning services and platform flexibility to support the customer's choice of network architecture, development framework, or data sources make it the cloud platform of choice for deep learning professionals.