GUIDEBOOK
PROGRAM: DATA AND ANALYTICS
DOCUMENT NUMBER: S180 NOVEMBER 2018

TENSORFLOW ON AWS

ANALYST

Rebecca Wettemann

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

THE BOTTOM LINE

Deep learning (DL) may be one of the most common buzzwords in artificial intelligence (AI) today, but until recently it has been mostly conceptual and experimental in nature. Deep learning is a subset of machine learning (ML): ML uses algorithms to analyze data, learn from that data, and make decisions or recommendations based on that learning in a linear fashion. Deep learning structures similar algorithms in layers to create an "artificial neural network," where the layers are put in a matrix, or tensor, with each layer being weighted, and through an iterative computation process the weights are adjusted to reach the desired outcomes. Deep learning typically requires thousands of iterations in the training phase, and very large data sets, often with complex data such as images – so the amount of compute power to operationalize deep learning is significant. Deep learning performs complex tasks, such as voice and facial recognition, sentiment analysis, natural language processing (NLP), image classification, fraud detection, and complex recommendations, better than traditional linear machine learning.

Deep learning frameworks have been in development since the 1940s but have only recently been operationalized in a meaningful way because of the emergence of two critical components: broad data sets for experimentation, and access to computing power. It is not surprising that most deep learning has moved to the cloud given its ability to scale and provide flexible, nearly-limitless CPUs as needed for model training and production. To better understand the state of deep learning adoption and usage today, Nucleus analyzed the experiences of more than 30 DL experts managing more than 177 unique projects. We found that 96 percent of deep learning today is running in the cloud, with Tensorflow being the most popular DL framework, being used in 89 percent of all DL projects. We also found that 85 percent of Tensorflow projects running on the cloud are running on Amazon Web Services (AWS), because of AWS's breadth of capabilities, support, and ongoing investments in services such as Amazon SageMaker (a managed service for building, training, and deploying machine learning and deep learning models at scale with minimal coding and configuration) that drive faster time to execution of large-scale DL projects.

THE SITUATION

Although there is significant hype around deep learning today, in the past 12 months many organizations have moved from hype and experimentation to reality. Medical research firms, consumer goods companies, marketing firms, university researchers, and various startups are increasingly turning to deep learning because it can solve a diverse variety of complex tasks more effectively than traditional machine learning with little or no human intervention.

The "secret" of deep learning is its iterative learning process, with multiple layers of algorithms processed in a single tensor. The results from each processing layer are fed back into the model to refine its results until the desired level of accuracy is reached. For

example, in medical image processing, a DL model might review hundreds of thousands of MRIs until it can effectively identify a certain brain abnormality with a less than 2 percent false positive rate.

Deep learning has evolved from simple tensors with one or two processing layers of algorithms to more complex tensors that may have five or more processing layers. Until recently, the two main barriers to effective deep learning have been the availability of broad and appropriate data sets and the accessibility of cost-effective computing. However, in the past 12 months Nucleus has seen significant advancements on both fronts:

- The rise of data development initiatives such as The Experience Project, the ability to use simple AI to anonymize medical record data so it can be anonymously shared and aggregated, and the move by cloud application users to opt in, in some cases, to allowing vendors to use their anonymized data for analysis have increased the quality and availability of data sets for research and experimentation.
- The investments that Amazon Web Services (AWS) and other cloud providers have made in both providing cloud computing power priced and optimized for deep learning (specifically, the rise of graphic processing units (GPUs)), as well as tools and services, support and documentation, and access to communities have made it easier and faster for organizations to execute on broad-scale DL projects.

When we first undertook this study in 2017, it was difficult to find data scientists that had moved beyond the experimentation phase with DL models and frameworks. In 2018, we found many more with DL expertise – both in academia, large enterprises, and startups and services firms – that were able to talk about:

- The business challenges and goals addressed by their projects
- The frameworks, methods, tools, and data libraries being used
- Their model experimentation, training, and deployment strategies
- The relative strengths and weaknesses of different models and frameworks
- Projects actually in production – meaning they were either delivering recommendations or data-driven results to end customers or providing a direct impact on decision making

Nucleus conducted in-depth interviews with 33 deep learning experts, many of whom had multiple teams working on DL projects. In all, this represented more than 177 unique projects.

DEEP LEARNING FRAMEWORK ADOPTION

Although some deep learning professionals develop their own deep learning frameworks, the availability of frameworks such as TensorFlow, originally developed in 2011, shortens the

time between concept and experimentation and training, with built-out capabilities for backpropagation and other capabilities that generate neural networks with higher speed and accuracy. TensorFlow was released under the Apache 2.0 open-source license in 2015.

We found that TensorFlow is still the most commonly-used DL framework today, with 89 percent of all projects using TensorFlow in some capacity. Of the total projects analyzed:

- Three percent were using only TensorFlow.
- More than two thirds of projects were applying TensorFlow as well as other frameworks (such as Apache MXNet, PyTorch, Caffe, Chainer, or other tools or frameworks). That trend has accelerated from last year, when 43 percent were using multiple frameworks.

Many of the deep learning professionals using more than one framework still considered TensorFlow to be their primary framework for deep learning, with Tensorflow being the clear framework of choice for image processing and recognition.

Deep learning continues to move to cloud, with only 4 percent of projects running on premise – and even fewer in production on premise.

DEEP LEARNING IN THE CLOUD

Given the compute demands of deep learning, it is not surprising that organizations going beyond the experimental or training phase to deliver deep learning in customer-facing or other professional applications are moving to the cloud, both to avoid the cost of building out their own data centers and to take advantage of cloud providers' investments in flexibility, scalability, and performance.

When we reviewed the state of the field last year, we found that momentum was building for moving DL workloads to the cloud. In last year's research, fewer than 10 percent of projects were being run on premise. That trend has accelerated, with only 4 percent of projects running on premise in 2018. We also found that AWS is still the predominant cloud choice for deep learning – and is gaining ground as the choice for all deployments:

- Eighty-one percent of all deep learning is running on AWS.
- Eighty-five percent of all Tensorflow-based projects in the cloud are running on AWS.

DEEP LEARNING IS REAL

We also found that deep learning is moving beyond experimentation to actual production, either in customer-facing applications or in actionable models driving business decisions. While only a handful of projects were beyond the experimentation stage in last year's study, in 2018, 14 percent of projects were actually in production. On a scale of one to five, deep learning professionals Nucleus interviewed rated their projects' maturity at an average of 3.8.

Although on-premise continues to be an option for experimentation, organizations moving to production are doing so in the cloud, and the cloud of choice is predominantly AWS. Although some startups were pursuing a multi-cloud strategy for commercial reasons (some other cloud vendors' startup partners, for example, receive cloud credits and sales support for their solutions), the vast majority of them were still relying on AWS for deep learning. Of the projects in production included in our analysis, four out of five were running on AWS.

> Four out of five deep learning projects in production are running on AWS.

USER PROFILE – CONSUMER TECHNOLOGY COMPANY

A global consumer technology company has adopted Tensorflow and other frameworks to support gaming-related personalized guidance and voice interaction for its gaming technology, which is in production today. It is also exploring the application of deep learning for image recognition for a more personalized gaming experience.

The company runs all of its deep learning on AWS, except for some initial experimentation and development work that is of small enough scale that it can be performed on developers' PCs. The head of machine learning in its engineering division rated its maturity of use of DL frameworks as a four out of five. The company is not using Amazon SageMaker yet, but has adoption on its roadmap in the next few quarters.

It chose AWS because of its scalability, proven history, and customer support. The company's lead engineer cites support for its mix of frameworks, AWS investment in professional support, and services such as SageMaker as drivers for accelerated time to production and a direct impact on increasing customer engagement and revenue.

WHY TENSORFLOW ON AWS

Nucleus found that there were three main reasons why DL experts chose to run their projects on AWS instead of other cloud providers: the breadth of AWS's capabilities, the Amazon relationship, and AWS's investments in the DL space.

BREADTH OF AMAZON CAPABILITIES

Nucleus found that, particularly as experts working in deep learning sought to move models and applications to production, Amazon's data storage, support for multiple frameworks, compute locations, and other resources that could support large-scale projects were important factors in choosing AWS. Users said:

- *"The decision was based on technology. AWS is far ahead of its competitors in terms of maturity of services. We've had other cloud providers and transitioned to AWS."*
- *"We're comfortable with the Amazon stack and it was the best."*
- *"I'm thinking ahead. I have models that detect what you're doing. Do I want it to talk to you, listen to you? All those technologies come from Amazon. Google may or may not have them but I don't want to get into mixing and matching different things. AWS has the integration and the breadth of services."*
- *"Amazon simply has the largest cloud computing resources."*

Many organizations already had large amounts of data stored in Amazon S3 or one of AWS's other database or storage services, and found that conducting analysis where the data already existed drove greater speed and cost effectiveness. Additionally, because machine learning and deep learning were only part of their analytics workflow, AWS provided them with a broad range of compatible options for data lake storage, analytical tools, and security, as well access to compute instances based on NVIDIA's V100 GPUs, which provide significant benefits for both training and running deep learning.

THE AMAZON WEB SERVICES (AWS) RELATIONSHIP

Experts in deep learning also cited their relationship with AWS, as well as demand from customers who had a relationship or trusted that AWS would support their performance, data security, and privacy needs, as another key reason for choosing AWS. Nucleus also found that organizations' existing relationship with AWS made using AWS for TensorFlow a natural extension of their existing footprint (particularly their existing investments in application workloads, storage, databases, and analytics), making adding compute for deep learning relatively seamless.

Beyond AWS's cloud computing capabilities, many teams cited previous positive experiences using AWS to support other deep learning projects as a reason for trusting AWS for new projects. Users said:

- *"The support you get and the community is really important."*
- *"The whole point is we don't need to worry about infrastructure and we can focus on what we're developing here. To change our cloud provider, they'd have to give it to us for free."*
- *"Our team had prior experience with AWS and building out our own data center was not something we wanted to focus on."*
- *"With AWS, there's a proven history, and good customer and executive support."*

It is interesting to note that a number of the organizations interviewed for this study had experience with other cloud partners; in fact, a number had been members of other vendors' incubator or startup programs and were still using AWS to run part or all of their production despite receiving cloud credits from other providers. One CEO of a startup said, "*We still have 100K in IBM cloud credits but can't find a way to use them – it just doesn't make sense.*"

AWS DEEP LEARNING INVESTMENTS

Leaders in the deep learning space also cited ongoing investments AWS is making in framework support, documentation, and services like Amazon SageMaker as key reasons for using AWS to support their deep learning. SageMaker, announced in late November 2017, is a fully managed service that enables developers and data scientists to build, train, and deploy machine learning and deep learning models at any scale without much of the technical configuration work typically required for executing projects. The service allows TensorFlow users to author models in a hosted Jupyter notebook environment (or import existing models), rapidly train across virtually unlimited CPUs or GPUs, automatically tune models, and deploy and host models with minimal coding. Users said:

- *We're constantly benchmarking performance of AWS with every system we can find. With Amazon and SageMaker, we realized SageMaker had so many people on it and the performance is so good. The partnership is great too, the investments they're making. If we had gone to Google I doubt they would have done it."*
- *"We decided based on R and D – for production we use very powerful AWS instances that we trigger to SageMaker. We didn't consider others – it would just be so inconvenient to use infrastructure from another cloud provider."*

USER PROFILE – RESTAURANT CHAIN

A North American fast-food restaurant chain has adopted Amazon Comprehend, Tensorflow, and MXNet to support a number of projects including image recognition and natural language processing (NLP) to monitor and predict food safety. It has NLP projects in

production today that distribute food safety recommendations and alerts to franchisees and management based on an analysis of social media posts.

The company is running all of its DL on AWS and is using SageMaker not only to accelerate time to production, but also to support streamlined distribution of its future applications and models to its hundreds of locations. It cited the breadth of the AWS portfolio and its existing relationship (AWS is its primary cloud provider for business applications) as key reasons for its DL cloud platform choice. The head of innovation rated its maturity of use of DL frameworks as a 3.5.

Key benefits the company has achieved from projects in production include improved management visibility and faster time to resolve potential food safety issues.

SageMaker drove a 20 percent average faster time to insight by accelerating setup and configuration and optimizing execution.

SAGEMAKER

Although SageMaker is a relatively new product, it is rapidly gaining adoption and delivering clear benefits for users. With roughly a third of researchers either using SageMaker or considering its use in the near future, Nucleus found that its use drove a 20 percent average acceleration in time to insight by both shortening the setup and configuration process and optimizing deep learning execution on AWS. Users said:

- *"SageMaker has reduced a lot of the heavy lifting for us. When you want to train a model on very big data you need to log on install packages, set up all the environments you want – it's time consuming and has nothing to do with data science. SageMaker does this for you – it brings up instances and manages performance, giving you more time to do research. Without it, it would take 20 percent more time and the computing cost would be much higher."*
- *"SageMaker gives us parallel scale. Because the data scientist isn't limited to their local compute, which of course maxes out, they can run variations of a model in parallel, testing different hypotheses to drive convergence to a single best model. This means they don't execute variant 1, wait for it to finish, execute variant 2, wait for it to finish, etc., but that they launch them all in parallel. For the experimentation phase, this gives us a lot of benefits."*
- *"With SageMaker, once the data scientist has arrived at a preferred model, they switch from parallel execution to vertical execution. This means using a P3 instance,*

rather than a P2 instance, which can deliver the training on a large imaging data set in hours rather than six to seven days."

- *"SageMaker provides you with all the integrations with data services and prebuilt models if you want to try something so you don't have to build it yourself, packaged in a nice little product – it gives us, easily, a 20 percent time savings."*
- *"There are a lot of dependencies with TensorFlow. You have to manage five or six different libraries of software. With SageMaker it's managed for you – five clicks, five minutes and it's all set up instead of five days."*

USER PROFILE – RESEARCH UNIVERSITY

A US-based university is investing in four major areas of DL research including NLP, image recognition for medical images, character profiling and mining, and semantic mapping. It is primarily using TensorFlow and MXNet and is running all of its deep learning on AWS.

Although most of the work is experimental in nature today, researchers at the university are applying their knowledge and development work on preprocessing models to productize NLP as a service in the cloud to extend the scope and productivity of others' NLP research. It chose AWS to support its deep learning because of AWS's experience in the NLP space and the breadth and scalability of AWS's cloud computing resources. The lead professor managing deep learning at the university rated their maturity of use of DL frameworks as a 4.5.

Main benefits the university has achieved include accelerated time to decision and the ability to support more complex projects with fewer skilled data scientists.

CONCLUSION

Once just an interesting area of machine learning clearly in the research and development camp, the adoption of deep learning for actual business use is accelerating at a rapid pace, with the number of actual deep learningh projects in production more than doubling in the past 12 months. In our research, we found that this was driven by three main factors:

- The availability of cloud computing from vendors such as AWS to support the computing-intensive needs of deep learning, the growth of documentation and services (such as SageMaker), and the support expertise to help DL professionals accelerate time to insight. This is coupled with the recognition that deep learning is just one component of an overall data-driven application strategy, and that seamless access to a broader cloud ecosystem of storage and computing services and workloads is the most cost-effective and rapid way to execute on deep learning insights.

- Greater understanding in the deep learning community of the best techniques, frameworks, and data sets to use to bring their concepts through experimentation to training to production, and the growth and availability of data sets for sharing and collaborative research.
- Investments in the community to help deep learning teams take advantage of the experience of others, with emerging "DL as a service" models, such as NLP platforms. These leverage the cloud model to enable teams with limited expertise in some areas to take advantage of additional capabilities built and managed in the "DL stack." We expect both shared "open source" services and services from AWS and others – such as Amazon Rekognition and Amazon Lex – to drive further operationalization of deep learning for less-experienced developers and teams.

The field of deep learning is clearly still evolving. As experts look to operationalize their research for measurable business outcomes, they are looking for not just compute power and data sets but an overall cloud infrastructure that can be optimized for deep learning training and execution and, ultimately, production. In our analysis, we found that AWS's reputation and relationship as a trusted enterprise computing partner, its continued investment in deep learning services and resources (such as SageMaker), and its breadth of services made it the cloud platform of choice for deep learning professionals.

Nucleus Research, Inc. | Boston, MA

Nucleus Research provides the ROI, insight, benchmarks, and facts that allow clients to understand the value of technology and make informed decisions. All research is built on an in-depth, case-study research approach that analyzes the results of actual deployments to provide technology advice built on real-world outcomes, not analyst opinions. Learn more at

NucleusResearch.com