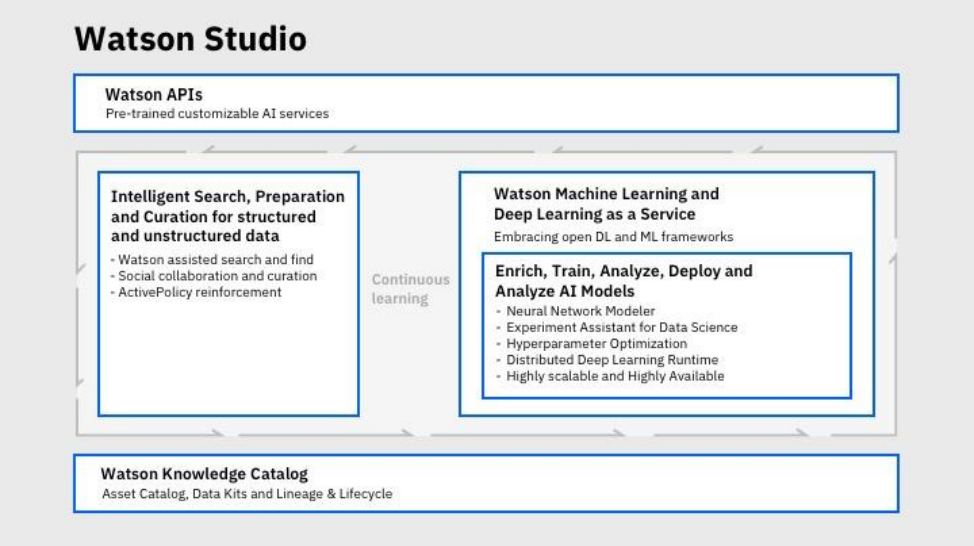
Deep Learning as a Service, IBM makes advanced AI more accessible for users everywhere

Artificial intelligence will be the most disruptive class of technologies over the next decade, fueled by near-endless amounts of data, and unprecedented advances in deep learning. The rise of deep learning has been fueled by three recent trends: the explosion in the amount of training data; the use of accelerators such as graphics processing units (GPUs); and the advancement in the training algorithms and neural network architectures.

Training of deep neural networks, known as deep learning, is currently highly complex and computationally intensive. It requires a highly-tuned system with the right combination of software, drivers, compute, memory, network, and storage resources.To realize the full potential of this rising trend, we want this technology to be more easily accessible to developers and data scientists so they can focus more on doing what they do best –concentrating on data and its refinements, training neural network models with automation over these large datasets, and creating cutting-edge models.

Today, I'm excited to announce the launch of Deep Learning as a Service within Watson Studio. Drawing from advances made at IBM Research, Deep Learning as a Service enables organizations to overcome the common barriers to deep learning deployment: skills, standardization, and complexity. It embraces a wide array of popular open source frameworks like TensorFlow, Caffe, PyTorch and others, and offers them truly as a cloud-native service on IBM Cloud, lowering the barrier to entry for deep learning. It combines the flexibility, ease-of-use, and economics of a cloud service with the compute power of deep learning. With easy to use REST APIs, one can train deep learning models with different amounts of resources per user requirements, or budget.

Our goal is to make it easier for you to build your deep learning models. Deep Learning as a Service has unique features, such as Neural Network Modeler, to lower the barrier to entry for all users, not just a few experts. The enhancements live within Watson Studio, our cloud-native, end-to-end environment for data scientists, developers, business analysts and SMEs to build and train AI models that work with structured, semi-structured and unstructured data —while maintaining an organization's existing policy/access rules around the data.



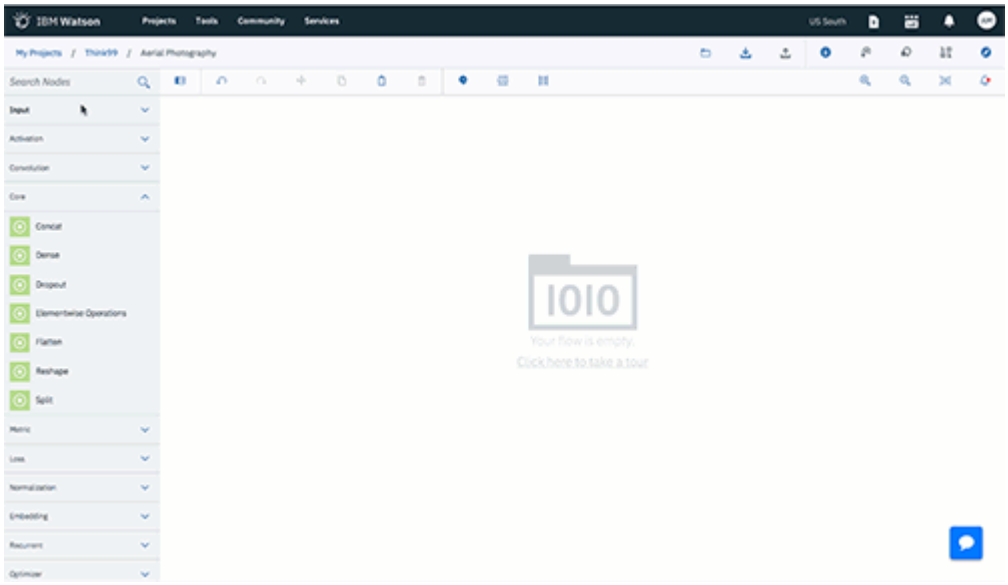
Making Deep Learning More Accessible, and Easier to Scale

Deep learning involves building and training a “neural network,” a machine learning model inspired by the human brain. Once a neural network is trained on a dataset, it can be used for a variety of recognition tasks —from identifying objects in an image and recognizing intention in an expression, to recognizing trends in a set of data.

For example, deep learning can help an insurance company determine how much a car has been damaged after an accident. How? An image of the damaged car can be included within a dataset trained at detecting not only the car make and model, but also where the car has sustained damage. Once the deep learning AI system recognizes the car, it compares the image of the damaged car to its dataset —and then classifies that damaged car as, for example, missing a bumper.

But developing deep learning models is a painstakingly iterative and experimental process –often requiring hundreds, even thousands of training runs that need very large amount of computing power to find the right combination of neural network configurations and hyperparameters. This may take weeks or even months.

This training process has been a challenge for data scientists and developers. To simplify this neural network building process and making it possible even for professionals without deep coding experience to do it, Deep Learning as a Service now includes a unique Neural Network Modeler. Neural Network Modeler is an intuitive drag-and-drop interface that enables a non-programmer to speed up the model-building process by visually selecting, configuring, designing and auto-coding their neural network using the most popular deep learning frameworks.



Automating Processes to Reduce Complexity

We've also abstracted out the complex, time-intensive and costly parameter optimization and training process. This Deep Learning as a Service is an experiment-centric model training environment, meaning users don't have to worry about getting bogged down with planning and managing training runs themselves. Instead, the entire training life-cycle is managed automatically and the results can be viewed in real-time and revisited later. Each training run is automatically started, monitored, and stopped upon completion, saving users time and money as they only pay for the resources they use.

The new feature also dramatically simplifies the often-arduous process of hyper-parameter selection. Instead of selecting hyper-parameters based on intuition, hyperparameter optimization provides an objective and automated method of exploring a complex problem-space. This results in a higher likelihood of identifying a more ideal model than most data scientists could find using traditional methods, like grid search. Therefore, users spend less time on experiments with little valuable output and more time developing even more sophisticated and powerful neural networks.

In addition, to accelerate experimentation for large training jobs, distribution across multiple machines and GPUs is critical in handling the large amount of training data for complex neural networks. The Deep Learning capability in Watson Studio is built upon IBM's distributed deep learning technology and the latest open source framework technologies, handling compute across many servers, each with multiple GPUs. Our consistent focus in Watson Studio is ease of use, and for distributed training, we hide the complex mechanisms of deep learning task distribution across various compute nodes and GPUs.

Furthermore, to develop a vibrant community around deep learning fabric, we are open sourcing the Fabric for Deep Learning (pronouncedFfDL), the core of Deep Learning as a Service. Leveraging the power of Kubernetes, FfDL provides a scalable, resilient, and fault-tolerant deep-learning framework. The platform uses a

distribution and orchestration layer that facilitates learning from a large amount of data in a reasonable amount of time across compute nodes.

References: IBM Deep Learning Service, B. Bhattacharjee, S. Boag, C. Doshi, P. Dube, B. Herta, V. Ishakian, K. R. Jayaram, R. Khalaf, A. Krishna, Y. B. Li, V. Muthusamy, R. Puri, Y. Ren, F. Rosenberg, S. Seelam, Y. Wang, J. M. Zhang, L. Zhang. Arvix, 2017.