AWS White Paper - Machine Learning Foundations

https://d1.awsstatic.com/whitepapers/machine-learning-foundations.pdf

Machine Learning Foundations

Evolution of Machine Learning and Artificial Intelligence

February 2019

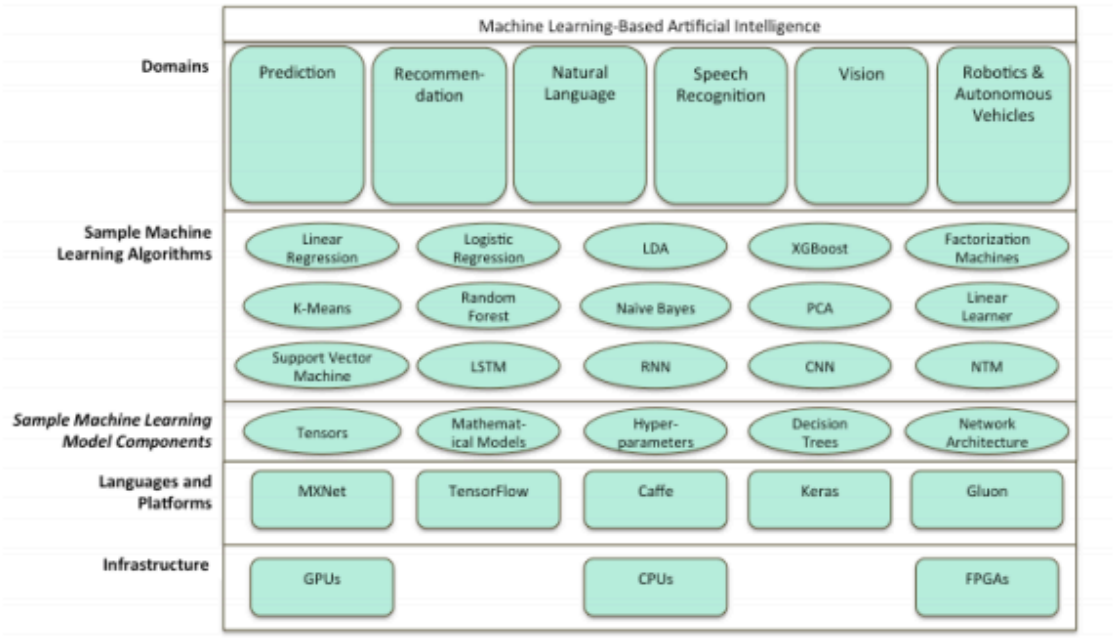

Figure 2: Machine Learning as a foundation for Artificial Intelligence

Table 3: Strengths and Limitations of ML Based AI

Strength	Limitation
Easy to train new solutions given data and tools	Experiencing hype, and researchers and practitioners need to properly set expectations
Large number of diverse algorithms to solve many types of problems	Requires large amounts of clean, potentially labeled data
Solves problems in all AI domains, often approaching or exceeding human level of capability	Problems in data, such as staleness, incompleteness, or adversarial injection of bad data, can skew results
No human expertise or complex knowledge engineering required, solutions are derived from examples	Some, especially statistically-based ML algorithms, rely on manual feature engineering

Strength	Limitation
Deep learning extracts features automatically, which enables complex perception and understanding solutions	System logic is not programmed and must be learned. This can lead to more subjective results, such as competing levels of activation, where precise answers are needed (e.g., specific true or false answers for compliance or verification problems).
Trained ML models can be replicated and reused in ensembles or components of other solutions	Selecting the best algorithm, network architecture, and hyperparameters is more art than science and requires iteration - though tools for hyperparameter optimization are now available
Making predictions or producing results is often faster than traditional inferencing or algorithmic approaches	Training on complex problems with large data sets requires significant time and compute resources
Algorithms for training ML models can be engineered to be distributed and one-pass, improving scalability and reducing training time	It is often difficult to explain how the model derived the results by looking at its structure and results of its training.
Can be trained and deployed on scalable, high-performance infrastructure	Most algorithms solve problems in one step, so no chains of reasoning or partial results are available, though outputs can reflect numeric "confidence"
Deployed using common mechanisms like microservices / APIs for ease of integrations with other systems	

AWS and Machine Learning

AWS is committed to democratizing machine learning. Our goal is to make machine learning widely accessible to customers with different levels of training and experience, and to organizations across the board. AWS innovates rapidly, creating services and features for customers prioritized by their needs. Machine Learning services are no exception. In the diagram below, you can see how the current AWS Machine Learning services map to the other AI diagrams.

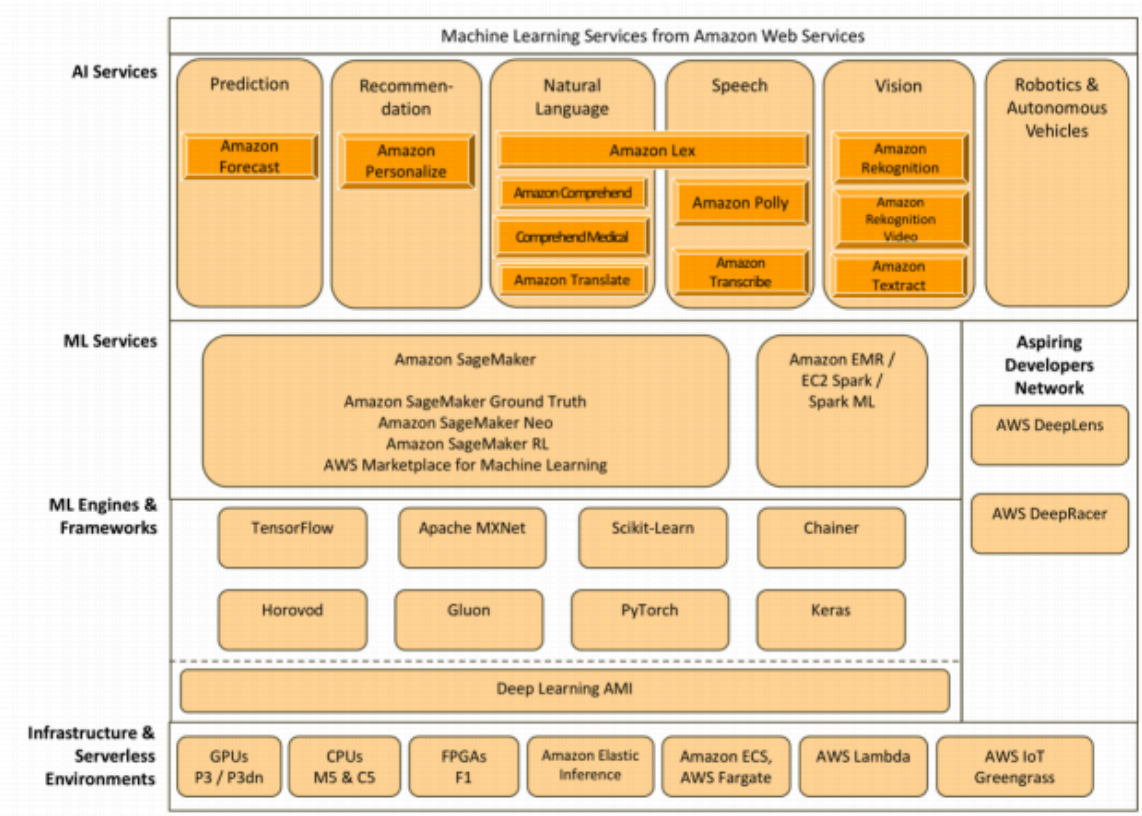

Figure 3: AWS Machine Learning Services

Amazon Forecast

Amazon Forecast is a fully managed service that delivers highly accurate forecasts, and is based on the same technology used at Amazon.com. You provide historical data plus any additional data that you believe impacts your forecasts. Amazon Forecast examines the data, identifies what is meaningful and produces a forecasting model.

Amazon Personalize

Amazon Personalize makes it easy for developers to create individualized product and content recommendations for customers using their applications. You provide an activity stream from your application, inventory of items you want to recommend and potential demographic information from your users. Amazon Personalize processes and examines the data, identifies what is meaningful, selects the right algorithms, and trains and optimizes a personalization model.

Amazon Lex

Amazon Lex is a service for building conversational interfaces into any application using voice and text. Amazon Lex provides the advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, to enable you to build applications with highly engaging user experiences and lifelike conversational interactions. With Amazon Lex, the same deep learning technologies that power Amazon Alexa are now available to any developer, enabling you to quickly and easily build sophisticated, natural language, conversational bots ("chatbots").

Amazon Comprehend

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. Amazon Comprehend identifies the language of the text; extracts key phrases, places, people, brands, or events; understands how positive or negative the text is and automatically organizes a collection of text files by topic.

Amazon Comprehend Medical

Amazon Comprehend Medical is a natural language processing service that extracts relevant medical information from unstructured text using advanced machine learning models. You can use the extracted medical information and their relationships to build or enhance applications.

Amazon Translate

Amazon Translate is a neural machine translation service that delivers fast, high-quality, and affordable language translation. Neural machine translation is a form of language translation automation that uses deep learning models to deliver more accurate and more natural sounding translation than traditional statistical and rule-based translation algorithms. Amazon Translate allows you to localize content - such as websites and applications - for international users, and to easily translate large volumes of text efficiently.

Amazon Polly

Amazon Polly is a service that turns text into lifelike speech, allowing you to create applications that talk, and build entirely new categories of speech-enabled products. Amazon Polly is a Text-to-Speech service that uses advanced deep learning technologies to synthesize speech that sounds like a human voice.

Amazon Transcribe

Amazon Transcribe is an automatic speech recognition (ASR) service that makes it easy for developers to add speech-to-text capability to their applications. Using the Amazon Transcribe API, you can analyze audio files stored in Amazon S3 and have the service return a text file of the transcribed speech.

Amazon Rekognition

Amazon Rekognition makes it easy to add image and video analysis to your applications. You just provide an image or video to the Rekognition API, and the service can identify the objects, people, text, scenes, and activities, as well as detect any inappropriate content. Amazon Rekognition also provides highly accurate facial analysis and facial recognition. You can detect, analyze, and compare faces for a wide variety of user verification, cataloging, people counting, and public safety use cases.

Amazon Textract

Amazon Textract automatically extracts text and data from scanned documents and forms, going beyond simple optical character recognition to identify contents of fields in forms and information stored in tables.

AWS Machine Learning Services for Custom ML Models

The *ML Services* layer in Figure 3 provides more access to managed services and resources used by developers, data scientists, researchers and other customers to create their own custom ML models. Custom ML models address tasks such as inferencing and prediction, recommender systems and guiding autonomous vehicles.

Amazon SageMaker

Amazon SageMaker is a fully-managed machine learning (ML) service that enables developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale. Amazon SageMaker Ground Truth helps build training data sets quickly and accurately using an active learning model to label data, combining machine learning and human interaction to make the model progressively better.

SageMaker provides fully-managed and pre-built Jupyter notebooks to address common use cases. The services come with multiple built-in, high-performance algorithms, and there is the AWS Marketplace for Machine Learning containing more than 100 additional pre-trained ML models and algorithms. You can also bring your own algorithms and frameworks that are built into a Docker container.

Amazon Sagemaker includes built-in, fully managed Reinforcement Learning (RL) algorithms. RL is ideal for situations where there is not pre-labeled historical data, but there is an optimal outcome. RL trains using rewards and penalties, which direct the model toward the desired behavior. SageMaker supports RL in multiple frameworks, including TensorFlow and MXNet, as well as custom developed frameworks.

SageMaker sets up and manages environments for training, and provides hyperparameter optimization with Automatic Model Tuning to make the model as accurate as possible. Sagemaker Neo allows you to deploy the same trained model to multiple platforms. Using machine learning, Neo optimizes the performance and size of the model and deploys to edge devices containing the Neo runtime. AWS has released the code as the open source Neo-AI project on GitHub under the Apache Software License. SageMaker deployments run models spread across availability zones to deliver high performance and high availability.

Amazon EMR/EC2 with Spark/Spark ML

Amazon EMR provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances. You can also run other popular distributed frameworks such as Apache Spark - including the Spark ML machine learning library, HBase, Presto, and Flink in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB. Spark and Spark ML can also be run on Amazon EC2 instances to pre-process data, engineer features or run machine learning models.

Aspiring Developers Framework

In parallel with *ML Services*, is the *Aspiring Developers Framework* layer. With a focus on teaching ML technology and techniques to users, this layer is not intended for production use at scale. Currently, the aspiring developers framework consists of two service offerings.

AWS DeepLens

AWS DeepLens helps put deep learning in the hands of developers, with a fully programmable video camera, tutorials, code, and pre-trained models designed to expand deep learning skills. DeepLens offers developers the opportunity to use neural networks to learn and make predictions through computer vision projects, tutorials, and real world, hands-on exploration with a physical device.

AWS DeepRacer

AWS DeepRacer is a 1/18th scale race car that provides a way to get started with reinforcement learning (RL). AWS DeepRacer provides a means to experiment with and learn about RL by building models in Amazon SageMaker, testing in the simulator and deploying an RL model into the car.

ML Engines and Frameworks

Below the ML Platform layer is the *ML Engines and Frameworks* layer. This layer provides direct, hands-on access to the most popular machine learning tools. In this layer are the AWS Deep Learning AMIs that equip you with the infrastructure and tools to accelerate deep learning in the cloud. The AMIs package together several important tools and frameworks, and are pre-installed with Apache MXNet, TensorFlow, PyTorch, the Microsoft Cognitive Toolkit (CNTK), Caffe, Caffe2, Theano, Torch, Gluon, Chainer and Keras to train sophisticated, custom AI models. The Deep Learning AMIs let you

ML Model Training and Deployment Support

The *Infrastructure & Serverless Environments* layer provides the tools that support the training and deployment of machine learning models. Machine learning requires a broad set of powerful compute options, ranging from GPUs for compute-intensive deep learning, to FPGAs for specialized hardware acceleration, to high-memory instances for running inference.

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2 provides a wide selection of instance types optimized to fit machine learning use cases. Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources, whether you are training models or running inference on trained models.

Amazon Elastic Inference

Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances for making predictions with your model. Rather than attaching a full GPU, which is more than required for most models, Elastic Inference can provide savings of up to 75% by allowing separate configuration of the right amount of acceleration for the specific model.

Amazon Elastic Container Service (Amazon ECS)

Amazon ECS supports running and scaling containerized applications, including trained machine learning models from Amazon SageMaker and containerized Spark ML.

Serverless Options

Serverless options remove the burden of managing specific infrastructure, and allow customers to focus on deploying the ML models and other logic necessary to run their systems. Some of the serverless ML deployment options provided by AWS include Amazon SageMaker model deployment, AWS Fargate for containers, and AWS Lambda for serverless code deployment.

AWS Fargate is a serverless compute engine for containers that works with both Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS). Fargate makes it easy for you to focus on building your applications. Fargate removes the need to provision and manage servers, lets you specify and pay for resources per application, and improves security through application isolation by design.

ML at the Edge

AWS also provides an option for pushing ML models to the edge to run locally on connected devices using Amazon SageMaker Neo and AWS IoT Greengrass ML Inference. This allows customers to use ML models that are built and trained in the cloud and deploy and run ML inference locally on connected devices.