ANN, DL and Generative AI: Overview

Tessema Mengistu(Ph.D.)

mengistu@cs.vt.edu

Outline

- Overview of Artificial Neural Network
- Overview of Deep Learning
- Overview of Generative AI
- AWS offering: DL and Generative AI

Overview of Artificial Neural Network

- Artificial Neural Network (ANN)
  - Is a computational system consisting of a large number of interconnected units called artificial **neurons**
    - The smallest processing units of the ANNs
    - Input signal x_i (x_1, x_2, …, x_n)
  - The connection between artificial neurons can transmit signal from one neuron to another

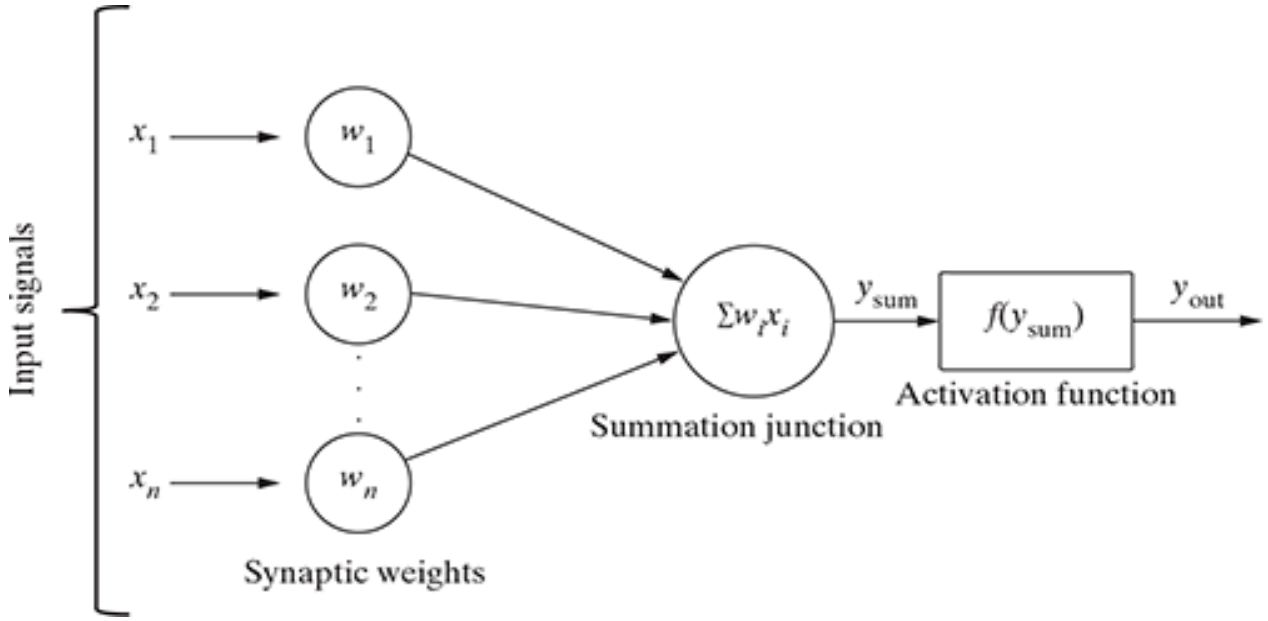
Overview of Artificial Neural Network

- Each neuron has three major components:
  - A set of ‘i’ **synapses** having weight w_i
  - A **summation junction** for the input signals is weighted by the respective synaptic weight

$$y_{\text{sum}} = \sum_{i=1}^{n} w_i x_i$$

  - A threshold activation function (or simply **activation function**
    - Results in an output signal only when an input signal exceeding a specific threshold value comes as an input

Overview of Artificial Neural Network

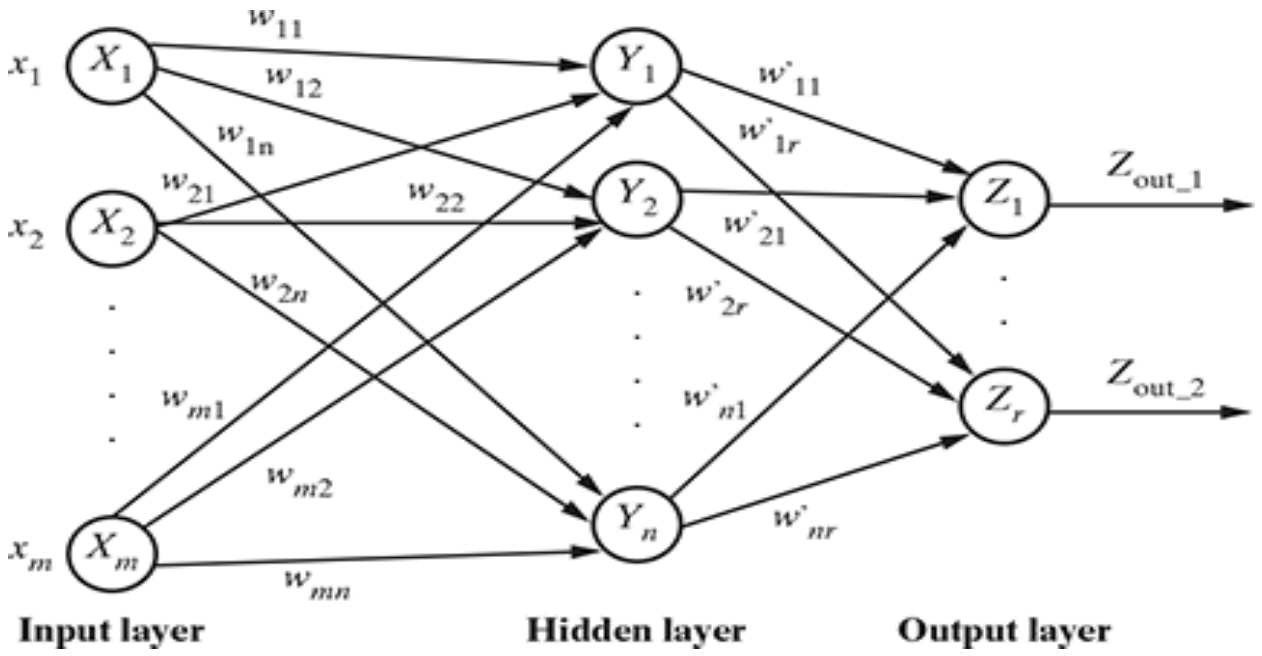


Overview of Artificial Neural Network

- Multi-layer feed forward is the simplest and most basic architecture of ANNs
  - The input layer consists of a set of 'm' input neurons X_1, X_2, …, X_m
  - The output layer consists of 'n' output neurons Y_1, Y_2, …, Y_n
  - One or more intermediate layers of neurons between the input and the output layers
    - May have varying number of neurons
  - The connections carry weights w_{11}, w_{12}, …, w_{mn}
  - The net signal input to the neuron in the hidden layer is given by

$$y_{in\_k} = x_1 w_{1k} + x_2 w_{2k} + \ldots + x_m w_{mk} = \sum_{i=1}^{m} x_i w_{ik},$$

Overview of Artificial Neural Network



Overview of Artificial Neural Network

- There are four major aspects which need to be decided for learning in ANN:
  - The number of layers in the network
  - The direction of signal flow
  - The number of nodes in each layer
  - The value of weights attached with each interconnection between neurons

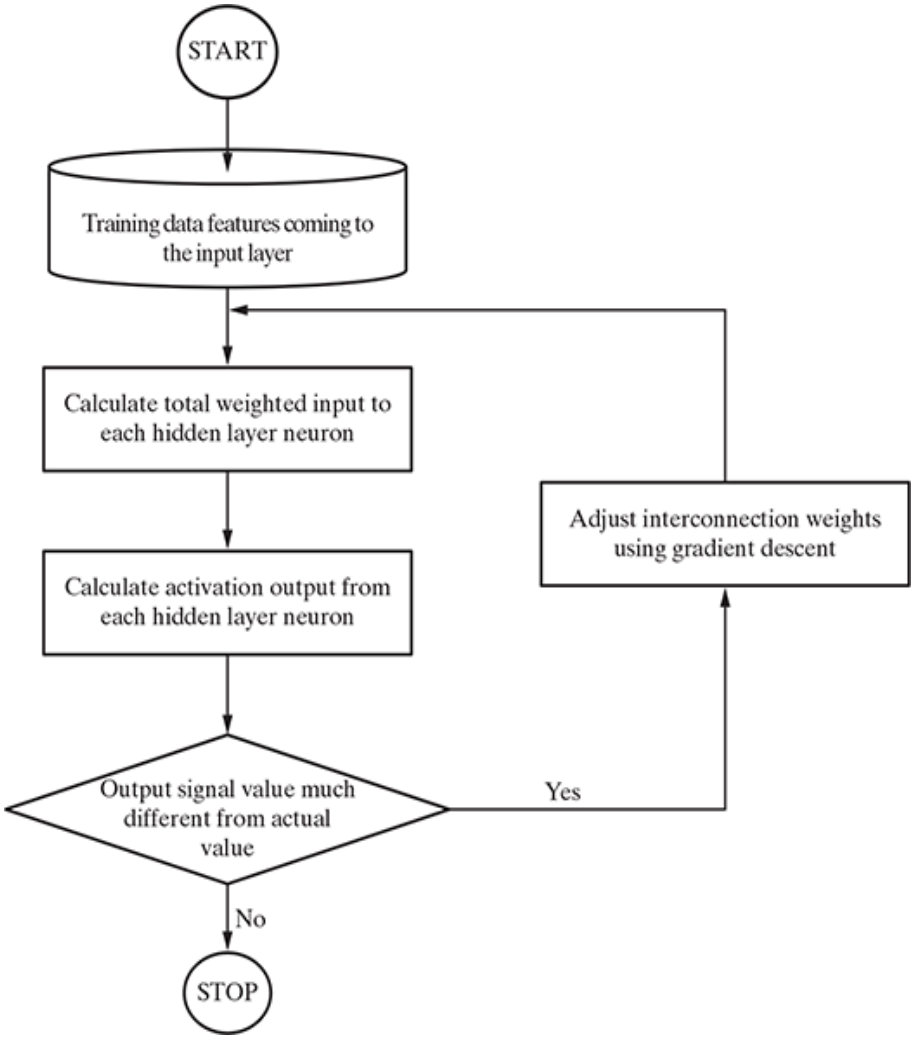
Overview of Artificial Neural Network

- Back Propagation
  - Applicable for multi-layer feed forward networks
  - It is a supervised learning algorithm which continues adjusting the weights of the connected neurons with an objective to reduce the deviation of the output signal from the target output
    - **Errors**, i.e., difference in output values of the output layer and the expected values, are propagated back from the output layer to the preceding layers
  - Consists of multiple iterations, also known as **epochs**
  - One main part of the algorithm is adjusting the interconnection weights
    - **Gradient descent**

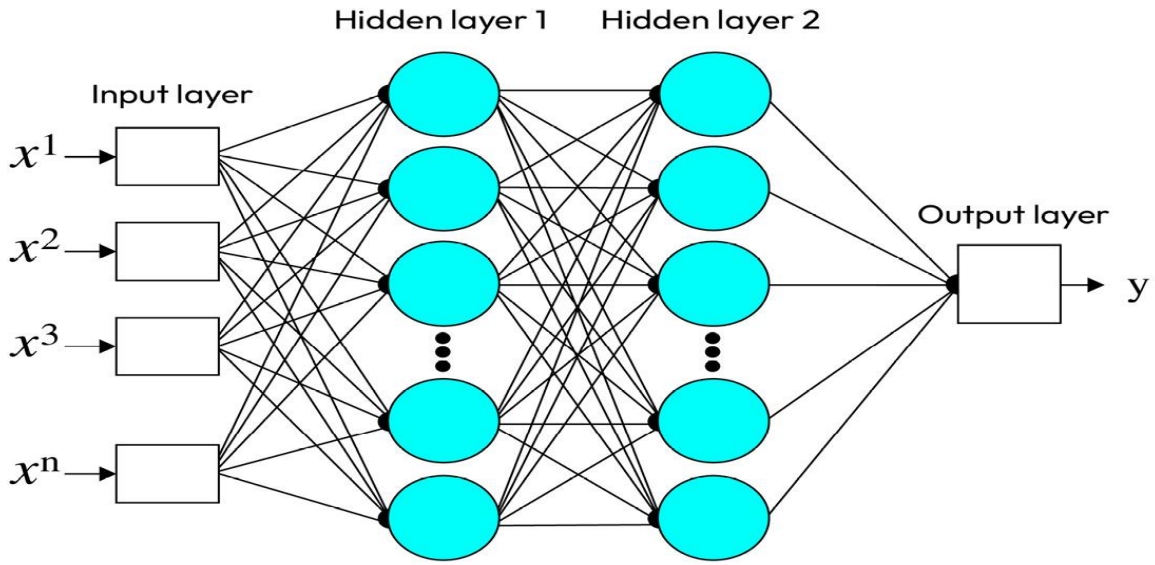
Overview of Artificial Neural Network

- Each epoch consists of two phases
  - A **forward** phase:
    - The signals flow from the neurons in the input layer to the neurons in the output layer through the hidden layers
    - The weights of the interconnections and activation functions are used during the flow
    - In the output layer, the output signals are generated
  - A **backward** phase:
    - The output signal is compared with the expected value
    - The computed errors are propagated backwards from the output to the preceding layers
    - The errors propagated back are used to adjust the interconnection weights between the layers

Overview of Artificial Neural Network



- Deep Learning
  - A multi-layer ANN with:
    - An input layer
    - At least two hidden layers
    - An output layer



Overview of Deep Learning

- Deep Learning
  - A specialized subset that uses neural networks with multiple layers to analyze complex factors in data
  - Efficient at handling unstructured data like images and text,
    - Led to breakthroughs in numerous complex tasks such as image and speech recognition

Overview of Deep Learning

- In a typical DL model, we define the following to construct a neural network:
  - The layers of the model (input layer, hidden layers, and output layers)
  - The activation function for each layer (such as ReLU or softmax)
  - The optimizer, which is the DL algorithm used to train the model
  - The loss function (such as MAE, MSE, and categorical_crossentropy)
  - The dropout rate, which is the percentage of nodes and their incoming/outgoing connections to be temporarily removed from the network
    - To avoid model overfitting

Overview of Deep Learning

- Different types of DL
  - Convolutional Neural Networks (CNNs)
    - Primarily in computer vision and image classification applications, can detect features and patterns within an image, enabling tasks, like object detection or recognition
  - Recurrent Neural Network (RNNs)
    - Typically used in natural language and speech recognition applications as it leverages sequential or times series data

Overview of Deep Learning

- Advantages of DL
  - Can analyze large amounts of data more deeply and reveal new insights for which it might not have been trained
  - Scalability - large data
  - Generalization - can generalize well to new situations or context
- Challenges
  - Large quantity of quality data
  - Large processing power
  - Lack of interpretability: DL models with many layers, can be complex and difficult to interpret

DL on AWS

- AWS managed DL services
  - Natural Language Processing
    - Amazon Comprehend
      - Uses machine learning to find insights and relationships in text
  - Computer vision
    - Amazon Rekognition
      - Image and video analysis to your applications deep learning technology
    - Amazon Textract
      - Extracts text and data from scanned documents

ML on AWS

- Speech
  - Amazon Polly
    - Text-to-speech - uses advanced deep learning technologies to synthesize speech that sounds like a human voice
  - Amazon Transcribe
    - Speech-to-text - Automatic speech recognition (ASR) service
- Chatbots
  - Amazon Lex
    - Provides the advanced deep learning functionalities
      - Automatic speech recognition (ASR) for converting speech to text
      - Natural language understanding (NLU) to recognize the intent of the text

ML on AWS

- Forecasting
  - Amazon Forecast
    - Uses machine learning to combine time series data with additional variables to build forecasts
- Recommendation
  - Amazon Personalize
    - Create individualized recommendations for customers
    - Privacy

Overview of Generative AI

- Generative AI
  - A type of AI that create new content
    - Conversations, music, video, image, etc.
  - Powered by pretrained ML models – Foundation Models (FMs)
    - Pretrained on a broad spectrum of generalized and unlabeled data to perform a wide range of tasks
      - Text/video/audio generation, data summarization , Q&A, programming code, etc.
      - In contrast with the traditional ML models that trained for a specific task
      - Example: OpenAI trained GPT-4 using 170 trillion parameters and a 45 GB training dataset
      - Can be:
        - Unimodal
        - Multimodal
  - Based on the **transformer architecture**

Attention is all you Need: https://arxiv.org/abs/1706.03762

Overview of Generative AI

Artificial intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically

Deep learning (DL)

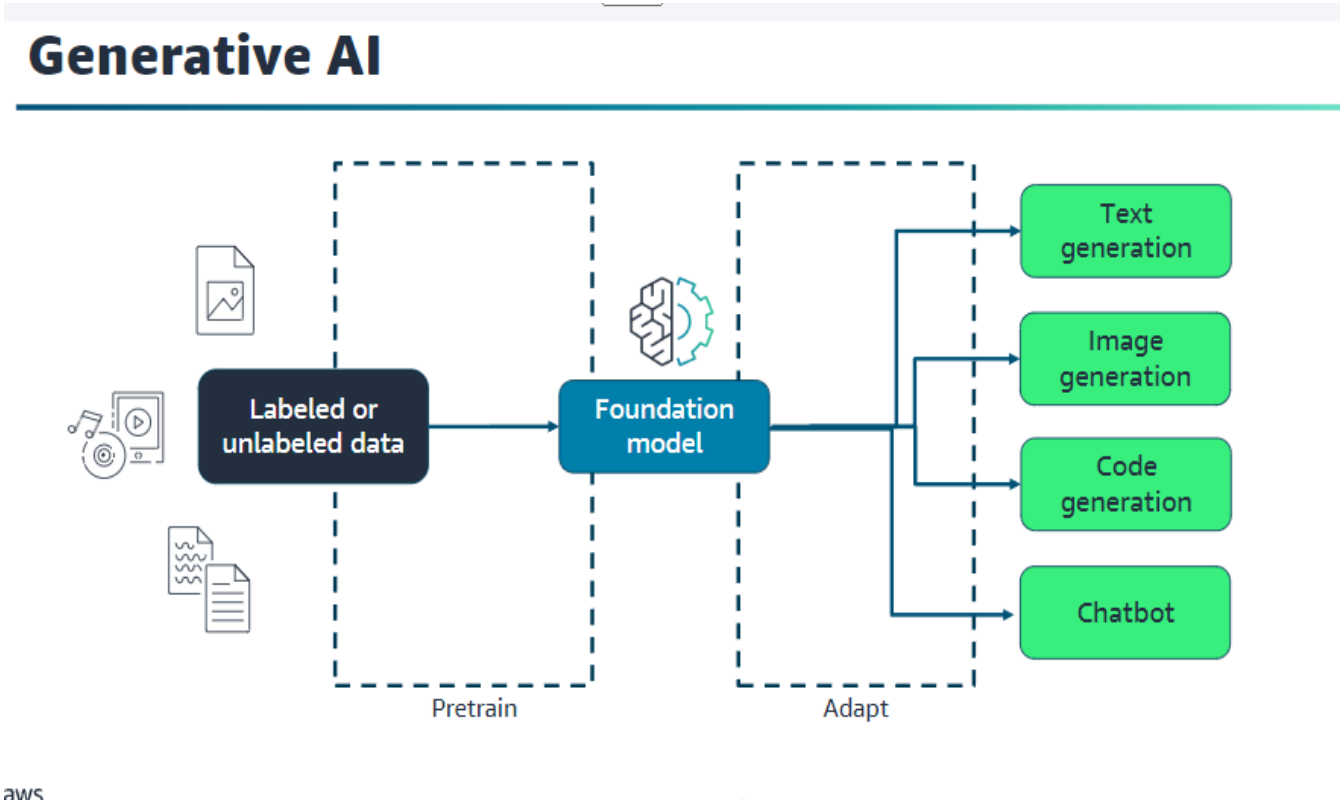
A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

Generative AI

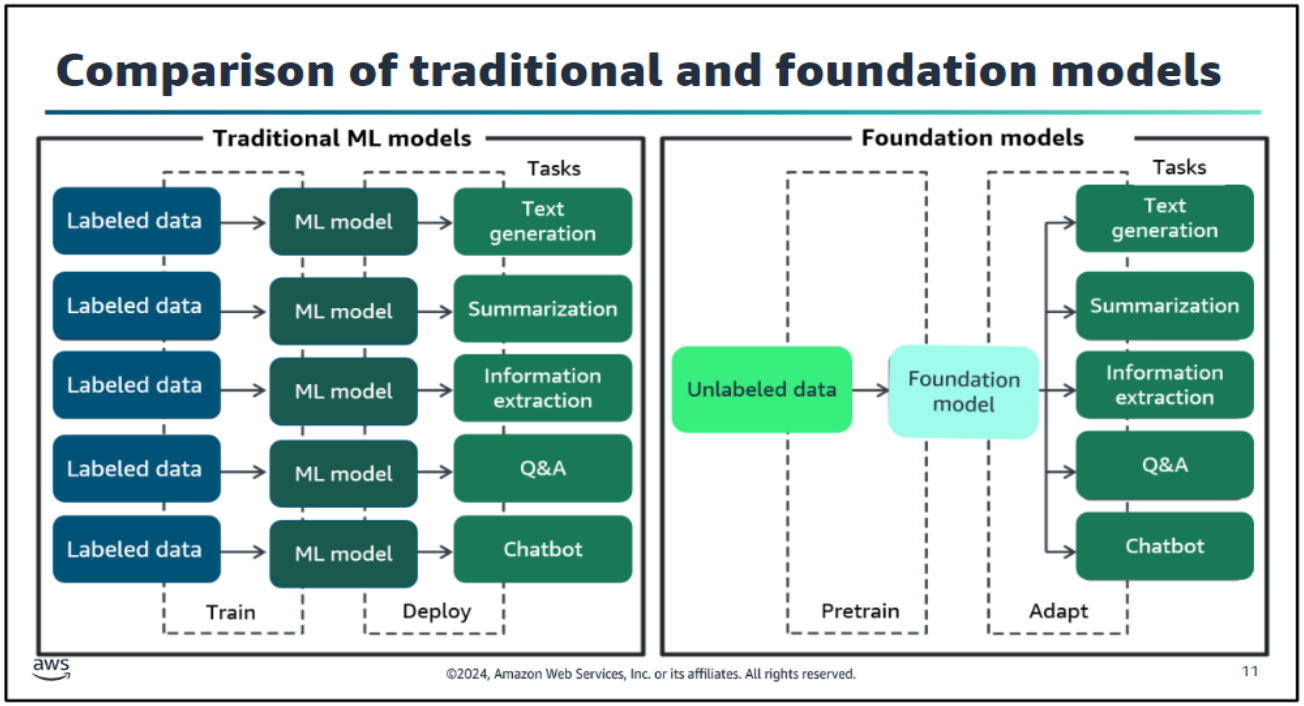
Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

Source: https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/aws-caf-for-ai.html

Overview of Generative AI



Overview of Generative AI



Overview of Generative AI

- LLM – Large Language Model
  - Subset of FM
  - Trained on very large words corpus
  - Can understand, learn, and generate text
  - Uses **prompt engineering** to insatiate an action from the FM
    - Input stimuli for the LM to generate specific outputs
  - Example: GPT models

Overview of Generative AI

Overview of Generative AI

- BERT - Bidirectional Encoder Representations from Transformers
  - A bidirectional model that analyzes the context of a complete sequence then makes a prediction
  - Trained on a plain text corpus and Wikipedia using 3.3 billion tokens (words) and 340 million parameters
  - Usage: Answer questions, predict sentences, and translate texts
- GPT - Generative Pre-trained Transformer
  - GPT-3 has a 96-layer neural network and 175 billion parameters and is trained using the 500-billion-word Common Crawl dataset
  - The popular ChatGPT chatbot is based on GPT-3.5
  - GPT-4, the latest version, launched in late 2022
- Mistral, Claude, Amazon Titan, etc.

DL and Generative AI services

- Amazon Elastic Inference
  - Allows you to attach low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances to run deep learning inference
    - Reduce the cost of running by up to 75%.
  - Supports TensorFlow, Apache MXNet, PyTorch, and ONNX models

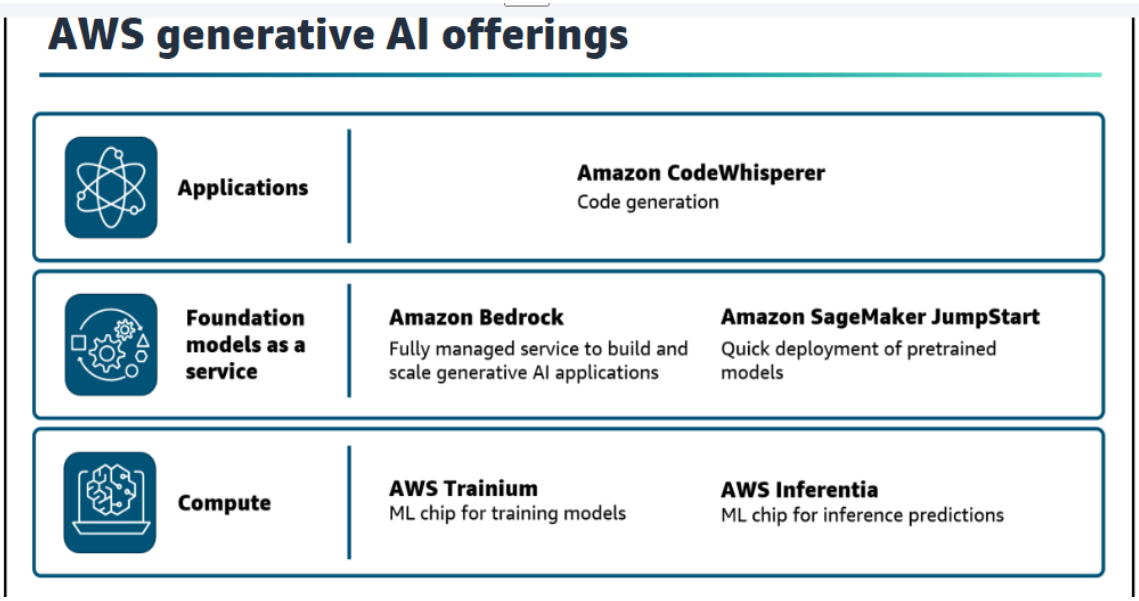
DL and Generative AI services

- Apache MXNet on AWS
  - Fast and scalable training and inference framework with an easy-to-use, concise API for machine learning
  - MXNet includes the Gluon interface that allows developers of all skill levels to get started with deep learning on the cloud

DL and Generative AI services

- AWS Deep Learning AMIs
  - Provide the infrastructure and tools to accelerate deep learning in the cloud, at any scale.
  - Train sophisticated, custom AI models, experiment with new algorithms, or to learn new skills and techniques.
  - Provides Amazon EC2 instances pre-installed with popular deep learning frameworks such as Apache MXNet and Gluon, TensorFlow, Microsoft Cognitive Toolkit, Caffe, Caffe2, Theano, Torch, PyTorch, Chainer, and Keras

DL and Generative AI services



DL and Generative AI services

- AWS Bedrock
  - A fully managed service that makes foundational models (FMs) from Amazon and leading AI startups available through an API
  - Allows private customization with your own data, and seamlessly integrate and deploy FMs into your AWS applications

DL and Generative AI services

- Amazon SageMaker JumpStart
  - Discover, explore, and deploy open source FMs or even create your own
  - Provides managed infrastructure and tools to accelerate scalable, reliable, and secure model building, training, and deployment

DL and Generative AI services

- AWS Inferentia
  - A custom ML chip designed by AWS
  - Accelerators are designed to deliver high performance at the lowest cost in Amazon EC2 for deep learning (DL) and generative AI inference applications
- AWS Trainium
  - Second-generation machine learning (ML) accelerator that AWS purpose built for deep learning training of 100B+ parameter models.

DL and Generative AI services

- GenAI Services:
  - AWS
    - Amazon Q - Latest addition
  - GCP
    - Gemini
  - Azure
    - OpenAI

Reference

- Machine Learning. S. Chandramouli, S. Dutt, A. K. Das, Pearson Education India, 2018
- https://aws.amazon.com/what-is/foundation-models/