Generative Ai Fundamentals

Introduction and Applications

Part 1:

- There are two approaches to AI:
  - Discriminative AI:
    - Learns to distinguish different classes of data
    - Different, classify, identify patterns, draw conclusions
  - Generative AI:
    - Create new content based on training data
    - Starts with a prompt; output is a new text, image, audio, video, code, data
  - Both use Deep Learning
- GenAi Models:
  - Generative Adversarial Networks (GANs)
  - Variational autoencoders (VAEs)
  - Transformers
  - Diffusion Models
- Capabilities of GenAi:
  - Text generation
  - Image generation
  - Audio generation
  - Video generation
  - Code generation
  - Data generation and augmentation
  - Virtual worlds

Part 2. Foundation Models and Platforms

A. Deep Learning and Large Language Models

- Deep Learning:
  - Performed with the help of Artificial Neural Networks (ANNs):
    - ANNs are neurons

    - Each neuron in the hidden layer contains inherent bias parameters
    - Connection between 2 neurons establishes weight parameters
  - Organized in 3 layers:
    - Input layer
    - Hidden layers
    - Output layer
- Supervised DL:
  - Predefined output
- Unsupervised DL:
  - Applications:
    - Clustering
    - Dimensionality Reduction
- DL Neural Architecture differentiates the level of responses produced by the algo:
  - Convolutional Neural Network (CNN):
    - Conducts a convolution / mathematical operation on a previous layer
    - Able to extract useful information from images to recognize patterns, classify images and segment pics
    - Useful:
      - Image processing
      - Video recognition
      - Natural language processing
  - Recurrent Neural Network (RNN):
    - Efficient in processing sequential data such as text and speech
    - Process memory component which enables them to capture dependencies and contextual information over time
    - Useful:
      - Machine translation
      - Sentiment analysis
      - Speech recognition
  - Transformer-based models:
    - Have a 2 stack structure where there is an encoder and decoder process – an exceptionally high number of parameters

- Analyzes and captures the context and meaning of words in a hierarchical sequence and predicts the next word in the output sequence
- This leads to creation of LLM

B. Core Generative Ai Models

- There are 4 Gen Ai Models with different DL architectures and use a probabilistic approach:
  - Variational Autoencoders (VAE):
    - Work with diverse range of data – images, audio, text
    - Rapidly reduce dimensionality
    - Original Input -> Encoder -> Latent Space -> Decoder -> Reconstructed Output
      - Encoder – studies the probability distribution
        - Isolates the most useful data variables which creates the most compressed representation of the data and stores in latent space
      - Latent Space:
        - Mathematical space
        - Stores large dimensional data in a compressed format
      - Decoder:
        - Decompresses the compressed the data in the latent space to generate the desired output
    - VAE is trained in a static environment but the latent space is continuous
      - This generates new samples by random sampling
      - Produce realistic varied images
      - Usage:
        - Image synthesis: create game maps; generate Anime Avatars
        - Data compression: forecast the volatility surfaces of stocks
        - Anomaly detection: detect diseases using electrocardiogram signals
  - Generative Adversarial Networks:
    - Uses imagery and textual input
    - Uses 2 CNNs compete with each other in an adversarial game

- 1 CNN plays the role of a generator and is trained on a vast dataset to produce data samples
- Other CNN plays the role of a discriminator and tries to distinguish between real and fake samples
- Based on the discriminator's responses, the generator seeks to produce more realistic data samples
- Applications:
  - Creates new, realistic images
  - Style transfer
  - Image-to-image translation
  - Deep fakes
  - Finance – loan pricing or generating time series
  - Creates video game characters
- Challenges:
  - Difficult to train
  - Large amount of data
  - Heavy computational power
  - Can create false material

- Transformer-based Models:
  - Focuses on valuable text and filter unnecessary text
- Diffusion Models:
  - Recent addition to GenAi
  - Addresses the systematic decay of data due to noise in the latent space by appling the principles of diffusion to prevent information loss
  - Diffusion process – moves molecules from high-density to low-density similarly the diffusion models moves noise to and from the data sample
    - Step 1: Forward Diffusion: algorithm gradually add random noise to a training data
    - Step 2: Reverse Diffusion: turn the noise to around to recover the data and generate the desired output
  - Benefits:
    - Can train unlimited layers
    - Remarkable for image and video data

C. Foundation Models

- Foundation Model is a large general purpose self-supervised pre-trained on a vast amounts of unlabeled data with billions of parameters

- Large Language Models:
    - Trained on NLP
    - Develop independent reasoning which allows them to respond to queries uniquely

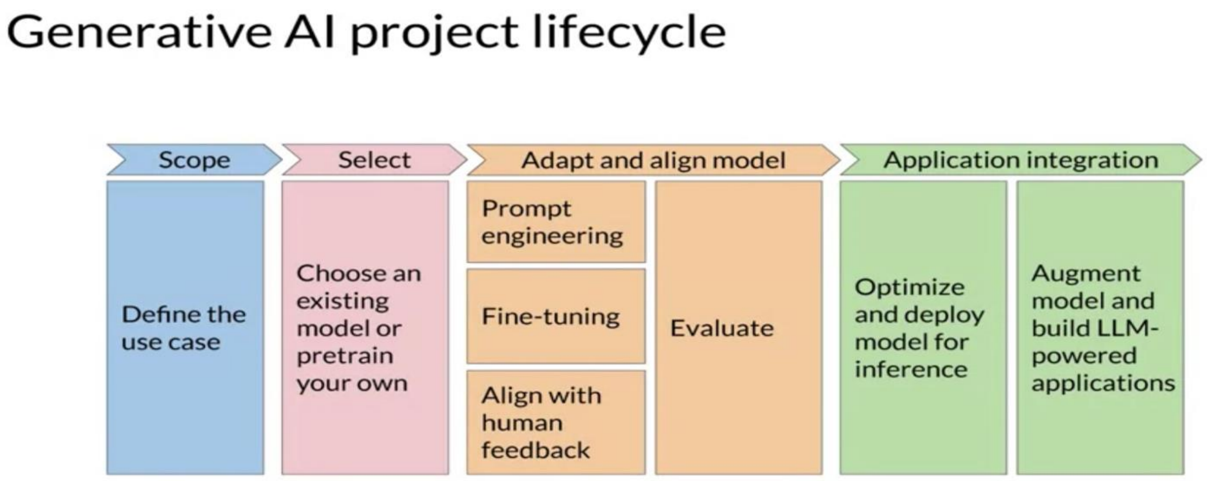
- Pre-Trained Models: Text-to-Text Generation
    - ML model
    - Trained on large corpus of text
    - Types of Model:
        - Statistical Model:
            - Markov Chain
                - Takes a sequence of states
        - Neural Network Models:
            - Use artificial neural networks to generate text
            - Represents complex relationships between data
            - Trained on a large corpus
            - Generates text similar to the text they are trained on
        - Use Sequence-2-sequence models or transformer models
        - Seq2Seq (Sequence to Sequence):
            - First encode the input text into a sequence of numbers then decode into a new number sequence
                - The new number sequence represents the generated text
            - Used for tasks with sequential data
            - Used:
                - Summarization
                - Speech Recognition
                - Machine Translation
        - Transformer Models:
            - Directly map the input text to the generated text

Gen Ai and Large Language Model Lifecycle

- Gen Ai is the subset of machine learning
- These models have learned to find statistical patterns in massive datasets
- Foundational / Base models:
        - GPT

- Bloom
- FLAN-T5
- PaLM
- LLaMa
- BERT

<u>Generative AI Project Lifecycle:</u>



Generative AI project lifecycle

Scope

Define the use case

Select

Choose an existing model or pretrain your own

Adapt and align model

Prompt engineering

Fine-tuning

Align with human feedback

Evaluate

Application integration

Optimize and deploy model for inference

Augment model and build LLM-powered applications

1. Scope – Define the use case:
    a. LLMs are capable of carrying out many tasks but their abilities depend strongly on the size and architecture of the model
    b. Tasks:
        i. Essay writing
        ii. Summarization
        iii. Translation
        iv. Information retrieval
        v. Invoke APIs and action
2. Select – Choose an existing model (foundational model) or pretrain your own:
3. Adapt and align model:
    a. Prompt Engineering: prompt engineering is the process where you guide LLM to generate desired output.
        i. Zero Shot
        ii. One Shot
        iii. Few Shot
    b. Fine-tuning: supervised learning process
    c. Align with human feedback

    d. Evaluate
4. Application Integration:
    a. Optimize and deploy model for inference
    b. Augment model and build LLM powered applications