جامعة الحسين التقنية
Al Hussein Technical University

Center of Professional Development and Community Outreach

Generative AI Training Program - Professional Development (Reskilling) Training Program

Generative AI - Foundations of Generative AI

جامعة الحسين التقنية
Al Hussein Technical University

Module1:
Foundations of Generative AI
(From AI history to modern architectures)

Dr. Abeer Al-Hyari
Abeer.alhyariups@htu.edu.jo

Module-1

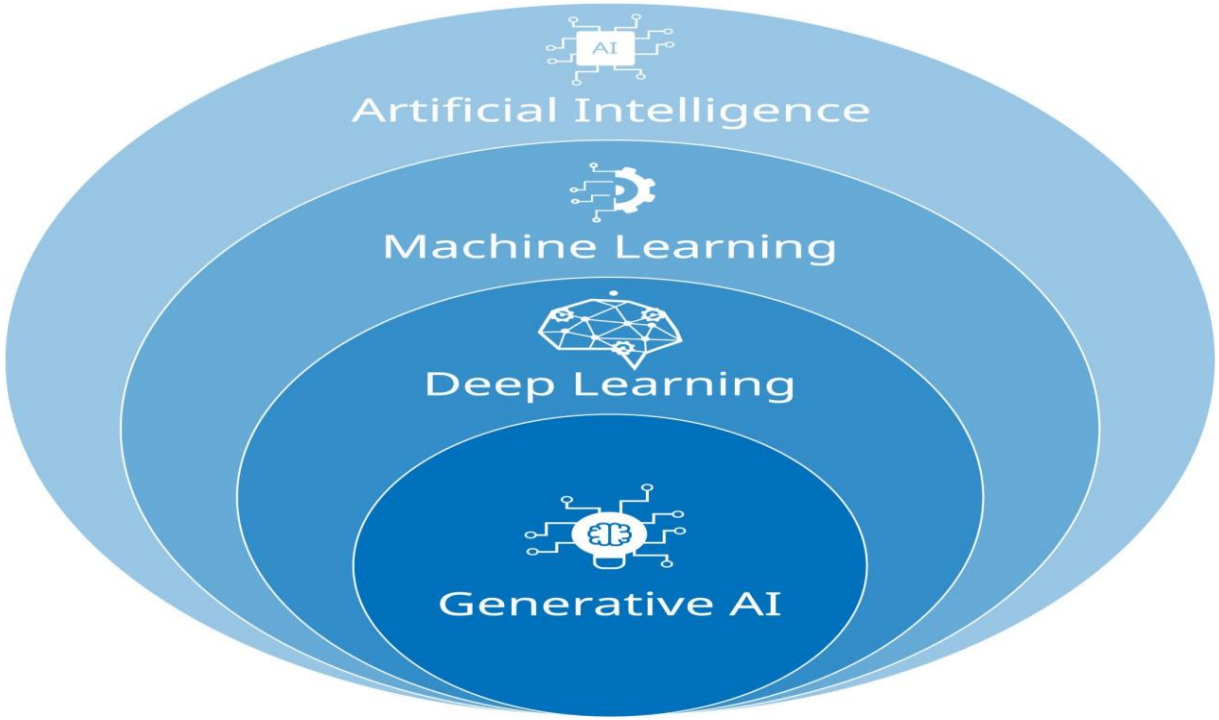
Module	Module Name and Focus	Key Topics to Cover	PLOs Addressed
Module 1	Foundations of Generative AI (From AI history to modern architectures)	1. AI, ML, and Deep Learning Review (Supervised vs. Unsupervised). 2. Introduction to Generative vs. Discriminative Models. 3. Overview of Key Generative Model Types (GANs, VAEs, Flow Models, Transformers). 4. The Pre-training and Scaling Hypothesis. 5. High-Level Look at Transformer Architecture and the Attention Mechanism.	Foundations (Primary)

Topics Covered

جامعة الحسين التقنية
Al Hussein Technical University

- AI, ML, and Deep Learning Review (Supervised vs. Unsupervised)
- Introduction to Generative vs. Discriminative Models
- Overview of Key Generative Model Types (GANs, VAEs, Flow Models, Transformers)
- The Pre-training and Scaling Hypothesis
- High-Level Look at Transformer Architecture and the Attention Mechanism.

The AI Landscape: A Hierarchy



AI

- AI: A field that has evolved since the 1950s. AI encompasses a variety of strategies and techniques for making machines perform tasks that typically require human intelligence.

ML

جامعة الحسين التقنية
Al Hussein Technical University

- ML: A group of AI techniques that use existing data to train a mathematical model. The model learns to recognize patterns in the training data so that it can make an accurate prediction when faced with new input data.

Supervised vs. Unsupervised ML

جامعة الحسين التقنية
Al Hussein Technical University

- Supervised learning is like learning under the guidance of a teacher.
- Unsupervised learning is like exploring data on your own without a teacher, trying to find inherent structure.

Feature	Supervised Learning	Unsupervised Learning
Input	**Labeled** (Input-Output Pairs)	**Unlabeled** (Only Input Features)
Goal	**Prediction** or **Classification**	**Pattern Discovery** or **Data Grouping**
Complexity	Generally **easier** to evaluate and implement.	Often **more complex** algorithmically; results can be subjective.
Typical Use	Spam filters, predictive maintenance, image recognition.	Market segmentation, recommendation systems, data visualization.

Deep Learning

جامعة الحسين التقنية
Al Hussein Technical University

- Uses deeply layered neural network models that have been trained on very large amounts of data
- Excels at working with unstructured data and modeling complex relationships between inputs and outputs
- Is designed for a specific domain

Example use cases

- Self driving cars
- Weather prediction
- Personalized recommendations

Generative vs. Discriminative Models

جامعة الحسين التقنية
Al Hussein Technical University

- Discriminative models draw a boundary (e.g., is this an apple or an orange?), while Generative models learn the underlying data distribution to create new data (e.g., generate a new apple).

The Discriminative Approach

جامعة الحسين التقنية
Al Hussein Technical University

- The **Discriminative Approach** in machine learning is a fundamental paradigm that focuses on **classifying or predicting** outcomes based on input data. It is the core methodology behind traditional machine learning and is distinct from the Generative Approach used in GenAI.
- **Core Principle: Learning the Decision Boundary**

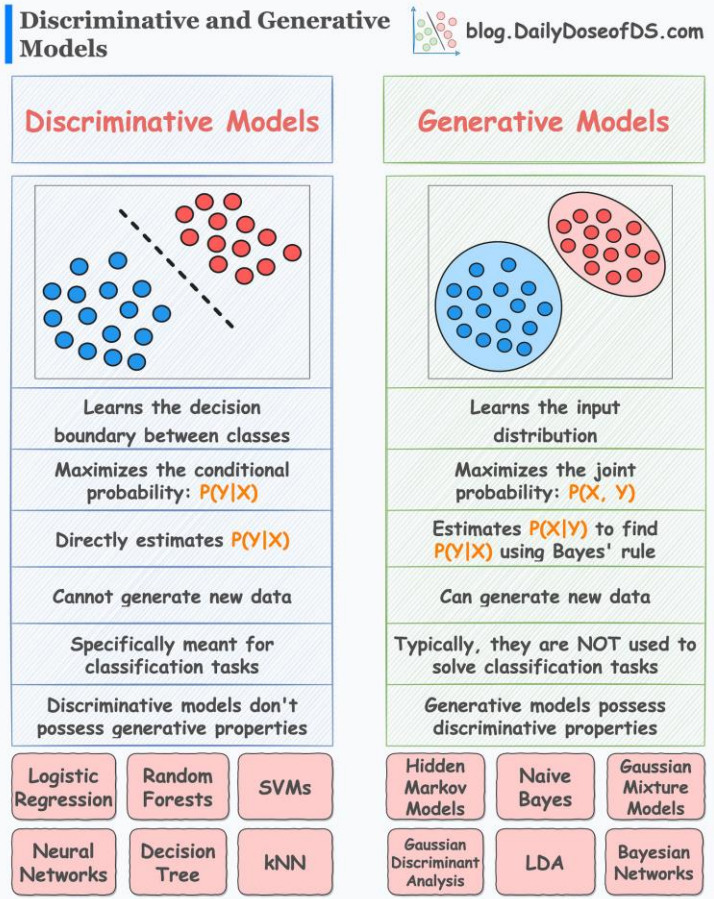
The Generative Goal

جامعة الحسين التقنية
Al Hussein Technical University

- The **generative goal** in AI is to create new, original data that is statistically and qualitatively similar to the data on which the model was trained.
- This core objective—**synthesis and creation**—is what fundamentally distinguishes generative models from traditional discriminative models.

Discriminative and Generative Models

https://youtu.be/hjsZSmL67Ck



Defining GenAI

جامعة الحسين التقنية
Al Hussein Technical University

- **Generative AI (GenAI)** refers to a class of artificial intelligence models designed to create new content—such as text, images, audio, video, or code—based on patterns learned from existing data. Unlike traditional AI systems that classify or predict, GenAI can produce original outputs that resemble human creativity.

Core Characteristics of GenAI

جامعة الحسين التقنية
Al Hussein Technical University

- **Content Creation**: Generates essays, poems, artwork, music, software code, and more.
- **Foundation Models**: Often built on large-scale transformer architectures (e.g., GPT, BERT, DALL·E) trained on massive datasets.
- **Multimodal Capabilities**: Can work across different types of media—text-to-image, image-to-text, audio-to-text, etc.
- **Interactive and Prompt-Driven**

Applications of GenAI

جامعة الحسين التقنية
Al Hussein Technical University

- **Education**: Personalized tutoring, curriculum design, automated grading.
- **Healthcare**: Drafting clinical notes, summarizing patient data.
- **Business**: Marketing copy, product descriptions, customer support automation.
- **Art & Design**: Visual concept generation, music composition, fashion prototyping.
- **Software Development**: Code generation, debugging, documentation

Taxonomy of Generative Models

جامعة الحسين التقنية
Al Hussein Technical University

- Introduce the main families: **Likelihood-based** (VAEs, Flow Models), **Implicit** (GANs), and **Sequence-based** (Transformers/Diffusion).

Variational Autoencoders (VAEs)

جامعة الحسين التقنية
Al Hussein Technical University

A VAE is a popular deep learning architecture that consists of two main parts:

- **Encoder:** Maps the input data (X) to a **probabilistic distribution** over a lower-dimensional representation called the **latent space** (Z). The encoder outputs the mean μ and variance (σ^2) of this distribution, typically a Gaussian.
- **Decoder:** Generates a new sample (X') by taking a random point sampled from the latent space (Z) and transforming it back into the original data space (e.g., an image).

Variational Autoencoders (VAEs)

جامعة الحسين التقنية
Al Hussein Technical University

How it's Likelihood-Based: VAEs cannot compute the exact log-likelihood of the data, as the math is too complex (intractable). Instead, they maximize the **Evidence Lower Bound (ELBO)**, which is a mathematically tractable approximation of the log-likelihood.

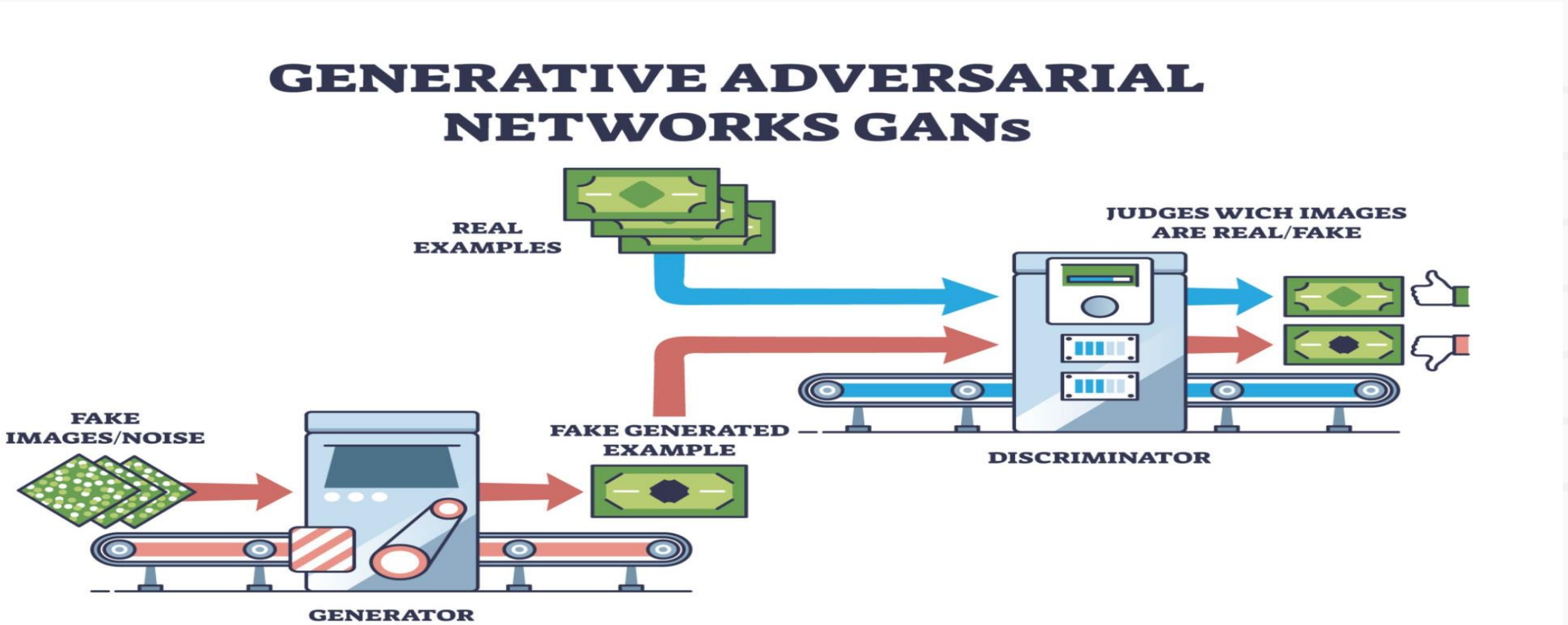
•**Trade-offs:** VAEs are **stable and easy to train**, and they create a **smooth, structured latent space** that is great for controlled generation and interpolation (e.g., morphing one face into another). However, they historically tend to produce **blurry outputs** compared to GANs due to the way their loss functions average pixels.

Generative Adversarial Networks (GANs)

جامعة الحسين التقنية
Al Hussein Technical University

- **Concept:** The two-player game (Generator vs. Discriminator). Focus on the core idea of learning through competition. **Outputs:** Realistic images.

Generative Adversarial Networks (GANs)



Flow Models & Other Likelihood Models

Flow-based generative models are a type of likelihood-based model that explicitly learns a complex probability distribution by transforming a simple, known distribution (like a standard Gaussian/normal distribution) through a sequence of invertible transformations.

- **Key Innovation:** The transformations used are designed to be **invertible** and have a **tractable Jacobian determinant**.
- **Training:** They are trained directly by maximizing the log-likelihood of the data.
- **Trade-offs:** They provide **exact density estimation** and guarantee efficient, non-iterative sampling. However, the requirement that all transformations must be invertible and have a tractable Jacobian often limits the model's architectural flexibility (expressiveness) compared to non-likelihood models like GANs.

What Is A transformer?

جامعة الحسين التقنية
Al Hussein Technical University

- A **Transformer** is a powerful neural network architecture that forms the backbone of nearly all modern Generative AI models, including Large Language Models (LLMs) like GPT, Gemini, and Llama.
- Its significance lies in its fundamental shift from older, slower sequential models (like RNNs/LSTMs) to a highly **parallel** structure built entirely on the **Attention Mechanism**. This change made it possible to train models with billions of parameters on massive datasets, directly leading to the current era of generative AI.

What Are the Most Popular Models?

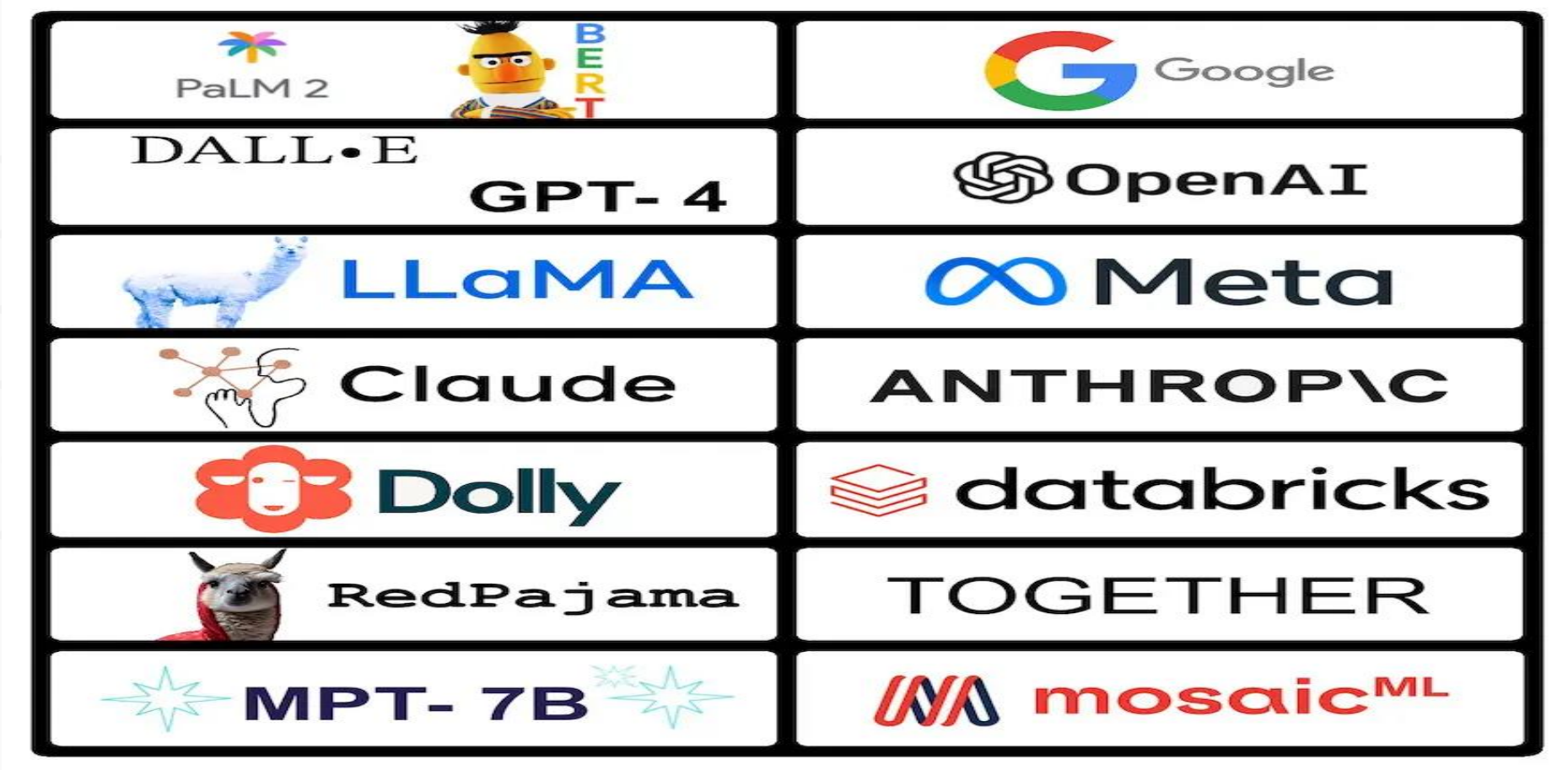


Image source: https://arize.com/blog-course/foundation-models-ai/

The Rise of the Transformer

جامعة الحسين التقنية
Al Hussein Technical University

Why the Transformer Replaced RNNs/LSTMs

The traditional sequence models (Recurrent Neural Networks and their variants, Long Short-Term Memory networks) processed data **sequentially** (one word at a time). The Transformer architecture abandoned this recurrence, leading to two major breakthroughs:

1. Parallel Processing (Speed and Scale)

- **The Problem with RNNs/LSTMs:** Since they process tokens sequentially, they could not be trained efficiently across large clusters of GPUs, creating a massive **computational bottleneck**. The calculation for the third word had to wait for the second word's result.
- **The Transformer Solution:** By relying *only* on **attention mechanisms**, the Transformer can process the entire input sequence **in parallel**. This unlocked the potential of modern hardware (GPUs and TPUs), drastically cutting down training time from weeks to hours and making it computationally feasible to train models with billions of parameters. This efficiency is the key enabler of the **Scaling Hypothesis**.

The Rise of the Transformer

جامعة الحسين التقنية
Al Hussein Technical University

2. Superior Context Modeling (Long-Range Dependencies)

- **The Problem with RNNs/LSTMs:** In very long sequences (e.g., a long document or a complex paragraph), RNNs/LSTMs often suffered from the **vanishing gradient problem**, causing them to "forget" information from the start of the sequence. They struggled to connect a pronoun to a noun that appeared 50 words earlier.
- **The Transformer Solution:** The core mechanism, **Self-Attention**, allows every token to directly and simultaneously access the information of *every other token* in the sequence, regardless of distance. The path between the first word and the last word is direct (a single attention step), not sequential (50 steps). This provides a **global understanding of context** unmatched by predecessors.

Pre-training and Fine-Tuning Paradigm

جامعة الحسين التقنية
Al Hussein Technical University

1. **Pre-training** on massive, general dataset (Foundation Model).
2. **Fine-Tuning** on smaller, specific task data. Analogy: General knowledge (Pre-train) vs. Specialization (Fine-tune).

The Scaling Hypothesis

جامعة الحسين التقنية
Al Hussein Technical University

- The hypothesis focuses on the simultaneous scaling of three primary factors:
  1. **Model Size(N):** The number of **parameters** (weights) in the neural network. Larger models have greater capacity to store knowledge and recognize complex patterns.
  2. **Dataset Size (D):** The amount and diversity of **training data** (tokens or images). More data provides a richer, more comprehensive view of the world.
  3. **Compute (C):** The total computational resources, typically measured in **FLOPs** (Floating-point Operations), used for training the model.

Foundation Models & LLMs

جامعة الحسين التقنية
Al Hussein Technical University

- **Foundation Models** (FM) as large, pre-trained models adaptable to many downstream tasks (e.g., GPT, BERT, Gemini).
- Explain LLM as the specific FM type for language.

Sequence-to-Sequence

جامعة الحسين التقنية
Al Hussein Technical University

- The **Sequence-to-Sequence (Seq2Seq)** concept is the foundational architectural pattern for many Generative AI models, particularly those dealing with language and translation tasks. It is designed to transform an input sequence into a corresponding output sequence, where the lengths of both sequences can be variable.

The Encoder-Decoder Architecture

جامعة الحسين التقنية
Al Hussein Technical University

The Seq2Seq model, in its most classic form (originally built using Recurrent Neural Networks, or RNNs), consists of two core components, often referred to as the **Encoder-Decoder architecture**:

1. The Encoder

- **Purpose:** The encoder processes the entire input sequence (e.g., a sentence in English) and converts it into a rich, compact, numerical representation, historically known as the **context vector** (or a set of hidden states).
- **Function:** It reads the entire source sequence to capture its **meaning, context, and semantic information**.

The Encoder-Decoder Architecture

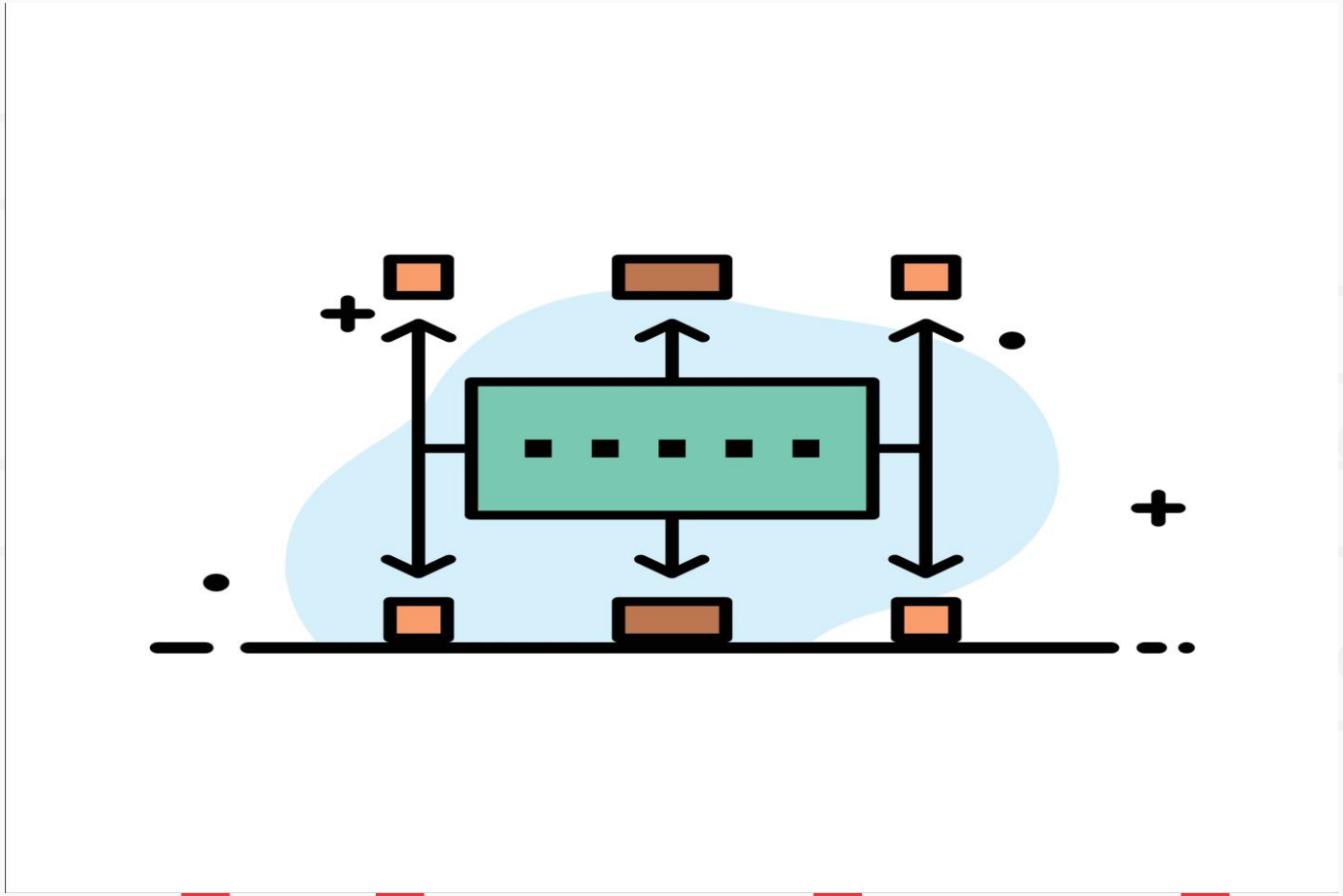
جامعة الحسين التقنية
Al Hussein Technical University

2. The Decoder

- **Purpose:** The decoder takes the context vector generated by the encoder and generates the output sequence (e.g., the translated sentence in French) one element (or token) at a time.
- **Function:** It operates in an **autoregressive** manner, meaning that each newly generated word is based on the previously generated words and the original context vector.

- https://youtu.be/zbdong_h-x4

The Encoder-Decoder Architecture



The Attention Mechanism

جامعة الحسين التقنية
Al Hussein Technical University

- **Concept:** The revolutionary idea of weighting the importance of every input token relative to every other token. Analogy: "Focusing" on the most relevant words in a sentence.
- https://youtu.be/KMHkbXzHn7s

Attention Is All You Need

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Core Concept: Query, Key, and Value (QKV)

جامعة الحسين التقنية
Al Hussein Technical University

The attention mechanism works by mathematically calculating a **weighted average** of the input elements (tokens). For every element in the sequence, the model computes three learned vector representations:

- **Query (Q):** Represents **"What am I looking for?"** (The current token's request for information).
- **Key (K):** Represents **"What information do I have?"** (The identifying tag or content of every other token in the sequence).
- **Value (V):** Represents **"What is the actual information?"** (The content/meaning of every other token).

How the Weights are Calculated:

جامعة الحسين التقنية
Al Hussein Technical University

The process to determine how much attention to pay to other words for a given word is as follows:

1. **Scoring:** The **Query** vector is compared against all **Key** vectors using a similarity function (typically the **dot product**). This produces a raw score indicating how relevant each token is to the current token being processed.

2. **Softmax:** The raw scores are scaled and then passed through a **Softmax** function to convert them into **Attention Weights**—a set of probabilities that sum to 1. This is the **dynamic weighting** that tells the model where to focus.

How the Weights are Calculated:

جامعة الحسين التقنية
Al Hussein Technical University

The process to determine how much attention to pay to other words for a given word is as follows:

3. **Output:** The final output is calculated as the **weighted sum** of the **Value** vectors, where the weights are the calculated attention probabilities. This creates a new, contextualized representation for the current token.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Self-Attention and Multi-Head Attention

Mechanism	Description	Location in Transformer
Self-Attention	Allows a token to attend to all other tokens in the same sequence (and itself) to calculate a refined contextual representation. It enables long-range dependencies to be captured directly and efficiently.	Used in both the Encoder and Decoder blocks.
Multi-Head Attention	The process is run multiple times in parallel using different, independently learned Q, K, and V matrices (heads). This allows the model to capture different types of relationships simultaneously (e.g., one head may track syntactic relationships, another semantic relationships).	Used throughout the Transformer architecture for deep contextual understanding.

Tokenization and Encoding

جامعة الحسين التقنية
Al Hussein Technical University

- Briefly explain how text is converted into numerical tokens (e.g., BPE). Mention the role of **Positional Encoding** (giving the model context about word order without recurrence).

Module Key Takeaways

1. Generative AI fundamentally learns data distribution.
2. Key architectures include GANs and VAEs.
3. Modern GenAI is dominated by **Transformers** and the **Scaling Hypothesis**.
4. The **Attention Mechanism** is the engine of the Transformer.

Reading for Module 1

Module Topic	Resource Type	Title / Author / Provider
1. AI, ML, and Deep Learning Review (Supervised vs. Unsupervised)	**Academic Textbook**	**Deep Learning** by Ian Goodfellow, et al. (MIT Press)
	Online Guide	**Supervised vs. Unsupervised Learning** (Google Cloud or IBM Think)

Reading for Module 1

جامعة الحسين التقنية
Al Hussein Technical University

Module Topic	Resource Type	Title / Author / Provider
2. Introduction to Generative vs. Discriminative Models	**Conceptual Book**	*Generative Deep Learning* by David Foster (O'Reilly Media)
3. Overview of Key Generative Model Types (GANs, VAEs, Flow Models)	**Online Course**	**Introduction Course to Autoencoders, VAEs, and GANs** (Simplilearn on Coursera)
	Review Paper (Conceptual)	*Deep Generative Modelling: A Comparative Review...* (Search for this title on arXiv)

Reading for Module 1

جامعة الحسين التقنية
Al Hussein Technical University

Module Topic	Resource Type	Title / Author / Provider
4. The Pre-training and Scaling Hypothesis	**Seminal Paper**	**"Scaling Laws for Neural Language Models"** by Kaplan, J., et al. (2020)
5. High-Level Look at Transformer Architecture and the Attention Mechanism	**Seminal Paper (Essential)**	**"Attention Is All You Need"** by Vaswani, A., et al. (2017)
	Visual Explanation	**Visualized Transformer Explanations** (search for resources like the "Transformer Explainer" by Polo Club of Data Science)