Lecture 4: What are Transformers?

1. The Origin Story

- **The "Secret Sauce":** The Transformer architecture is identified as the fundamental breakthrough behind modern Large Language Models (LLMs).
- **The Paper:** Introduced in the 2017 Google Brain paper titled *"Attention Is All You Need"*.
  - **Impact:** The paper has received over 100,000 citations in just a few years.
- **Original Purpose:** The Transformer was not originally designed for text completion (like ChatGPT); it was built for **Machine Translation** tasks, specifically translating English to German and French.

2. How a Transformer Works (The 8-Step Process)

Dr. Dander explains the architecture using a simplified schematic of translating English to German.

- **Step 1: Input Text:** The model receives the raw source text (e.g., "This is an example").
- **Step 2: Tokenization (Pre-processing):**
  - The input sentences are broken down into smaller chunks called **tokens** (words or sub-words).
  - Each token is assigned a unique numerical ID,.
- **Step 3: The Encoder (Vector Embeddings):**
  - The Token IDs are passed into the **Encoder** block.
  - **Vector Embedding:** The encoder converts tokens into vectors in a high-dimensional space to capture **semantic meaning**.
  - *Example:* In this vector space, related words cluster together. "Apple," "Banana," and "Orange" form one cluster; "King," "Man," and "Woman" form another.
- **Step 4: Passing Information:** The Encoder sends these semantic vector embeddings to the **Decoder**,.
- **Step 5: The Decoder:**
  - The Decoder receives the embeddings from the Encoder *and* the **partial output text** (the words translated so far).
- **Steps 6 & 7: Generation:**
  - The Decoder generates the translation **one word at a time**.
  - It uses the context of what has already been translated plus the meaning from the Encoder to predict the next word.
- **Step 8: Final Output:** The full translated sentence is produced (e.g., "Das ist ein Beispiel").

3. The "Secret Sauce": Self-Attention

The lecture explains why the paper is titled *"Attention Is All You Need."*

- **Definition:** The **Self-Attention Mechanism** allows the model to weigh the importance of different words relative to one another.
- **The Problem it Solves:** In long text, the context for a current word might be hidden in a sentence that occurred much earlier.
- **Long-Range Dependencies:** Attention allows the model to look at the entire sequence at once and capture dependencies regardless of how far apart words are in the text.
    - *Example:* If predicting a word in the 4th sentence, the model can "pay attention" to a specific context provided in the 1st sentence to ensure accuracy.

4. Variations: BERT vs. GPT

The original Transformer had both an Encoder and a Decoder. Later models specialized by using only one part of this architecture.

Feature	**BERT**	**GPT**
Full Name	Bidirectional Encoder Representations from Transformers	Generative Pre-trained Transformer
Architecture	**Encoder Only**	**Decoder Only**
Direction	**Bidirectional** (Looks at context from left and right)	**Left-to-Right** (Predicts next word based on previous history)
Training Task	Predicts "masked" (hidden) words in the middle of a sentence	Predicts the **next word** in a sequence
Best Use	Sentiment Analysis, understanding nuance	Text Generation, Text Completion

- **Why BERT for Sentiment?** Because it looks at words from both directions, it can distinguish between "Bank" (finance) and "Bank" (river) based on surrounding context better than a unidirectional model.

5. Important Distinctions (Myth-Busting)

The lecture clarifies that "Transformer" and "LLM" are not interchangeable terms.

- **Not all Transformers are LLMs:** Transformers are used in **Computer Vision** (Vision Transformers or ViT) to detect potholes, classify tumors, or recognize images.
- **Not all LLMs are Transformers:** Before 2017, Large Language Models existed but were built on older architectures like **RNNs** (Recurrent Neural Networks) and **LSTMs** (Long Short-Term Memory networks).