Lecture 5: How does GPT-3 really work?

1. The Evolutionary Timeline of GPT

The lecture outlines the progression from the original Transformer paper to modern LLMs:

- **2017: The Transformer:** Introduced in the paper *"Attention Is All You Need"*. It featured both an **Encoder** and a **Decoder** and introduced the self-attention mechanism.
- **2018: GPT-1 (Generative Pre-trained Transformer):** Introduced the concept of "Generative Pre-training".
    - **Key Shift:** Unlike the original Transformer, the GPT architecture removed the Encoder and used **only the Decoder**.
    - **Method:** It utilized unsupervised learning on unlabeled text.
- **2019: GPT-2:** Described as an "unsupervised multitask learner".
    - **Scale:** The largest version had ~1.5 billion parameters.
- **2020: GPT-3:** A massive leap in scale with **175 billion parameters**.
- **Present:** The evolution continued to GPT-3.5 and currently GPT-4 (which powers tools like ChatGPT).

2. Zero-Shot vs. Few-Shot Learning

A major focus of the lecture is understanding how models learn from prompts.

- **Zero-Shot Learning:** The model is given a task description with **no supporting examples** (e.g., "Translate cheese into French").
- **One-Shot Learning:** The model is given the task description plus **one single example**.
- **Few-Shot Learning:** The model is provided with **a few examples** of the task to guide its output.
    - **GPT-3's Classification:** The original GPT-3 paper is titled *"Language Models are Few-Shot Learners"*, claiming that while it can do zero-shot tasks, performance significantly improves with examples.
    - **GPT-4's Capability:** Even modern models like GPT-4 admit to being "few-shot learners" essentially—while they have excellent zero-shot capabilities, providing examples yields more accurate responses.

3. Data Scale and Training Costs

The capabilities of GPT-3 are driven by the massive scope of its training data.

- **Dataset Composition:**
    - **Common Crawl:** 60% of the data (410 billion words).
    - **WebText2:** 22% of the data (19 billion words from Reddit links).
    - **Books & Wikipedia:** ~18-19% combined.
- **Total Size:** The model was trained on approximately **300 billion tokens** (where a token is roughly a word).

- **Cost:** The computational cost to pre-train GPT-3 was estimated at **$4.6 million** due to the massive GPU resources required.

4. Architecture: Decoder-Only & Parameters

- **Decoder-Only:** Unlike the original 2017 Transformer (Encoder + Decoder) or BERT (Encoder only), GPT models use **only the Decoder** block.
- **Structure:** GPT-3 consists of 96 Transformer layers stacked on top of each other.
- **Parameter Count:** The defining feature of "Large" Language Models is the parameter count; GPT-3 has **175 billion parameters**, which are essentially the weights of the neural network optimized during training.

5. The Training Mechanism: Auto-Regression

GPT models are trained using a specific process called **Auto-Regression** via unsupervised learning.

- **The Task:** The model is trained *solely* to **predict the next word** in a sequence.
- **Unsupervised / Self-Supervised:** It does not require human-labeled data. The model uses the structure of the data itself:
    - The sentence is split into input and output.
    - The "label" (correct answer) is simply the next word in the sentence.
- **The Auto-Regressive Loop:**
    1. **Iteration 1:** Input "This" $\rightarrow$ Output "is".
    2. **Iteration 2:** The output "is" becomes part of the input. Input "This is" $\rightarrow$ Output "an".
    3. **Iteration 3:** Input "This is an" $\rightarrow$ Output "example".
    - The previous output is continuously fed back as the input for the next prediction.

6. Emergent Behavior

One of the most fascinating concepts covered is "Emergent Behavior."

- **Definition:** The ability of a model to perform tasks it was **not explicitly trained to perform**.
- **The Surprise:** GPT was *only* trained to predict the next word. However, as a byproduct of this training on massive data, it spontaneously learned how to:
    - Translate languages.
    - Solve math problems.
    - Write code.
    - Perform sentiment analysis.
- **Significance:** Researchers did not program these capabilities; they emerged naturally as the model improved at language understanding.