Data 188 Introduction to Deep Learning
Spring 2026 Eric Kim

Discussion 12

This discussion covers large-scale NLP pretraining, GPT-style language modeling, instruction/chat fine-tuning, training at scale, context-window limits, and societal impacts of LLMs.

1. GPT Pretraining

During pretraining, what is the primary task that GPT models train on? What is the training dataset like?

2. GPT Finetuning

Recall that many chatbot companies, such as OpenAI for ChatGPT, perform a second-stage fine-tuning step after pretraining (such as SFT or RLHF). What is this fine-tuning step? What could happen if they did not do this step and released the pretrained model to the public?

3. GPT Training at Scale

(a) Recall that GPT models are extremely large. For instance, GPT-4 is rumored to have 1.8T model parameters, which, if stored in float16, would require 3600 GB of GPU memory just to store the model weights.

Suppose we want to train GPT-4, and want to use the Adam optimizer. How much additional GPU memory would we need for Adam? Only consider additional optimizer state, ignore intermediate activations. Assume a naive, unoptimized implementation for Adam.

(b) As of 2026, it is unrealistic to have access to GPU machines that have more than tens of thousands of GB GPU memory on a single node. For example, AWS EC2 offers the p5en.48xlarge, which has 1128 GB GPU memory.

Suppose we only have access to nodes that each have 1128 GB GPU memory. How would we train a model like GPT-4?

4. Article Summarization: Exceeding Context Windows

A common task for Large Language Models (LLMs) is to ask the LLM to generate a brief and accurate summary of an input document, such as a textbook chapter, conference paper, or lecture slides. This is often called article summarization.

Recall that there is a maximum context-window length. For example, GPT-4 has a limit of 128k tokens (and the newly released DeepSeek V4 supports 1M context window). Suppose we have an input text document that, after tokenization, exceeds the LLM's context-window maximum length. What would go wrong if we silently ignored this? What are some possible techniques to overcome this limitation?

5. LLMs: Impacts on Society

As of 2026, it is undeniable that LLMs have made a significant impact on the world in nearly every industry, ranging from technical crafts such as software engineering and mathematics research to creative fields such as visual arts, design, music, and writing. At this point, the "cat is out of the bag"!

Spend some time reflecting on how this will impact our world and society, both in terms of your professional career and your own personal life. Are there dangers of over-relying on LLMs? Or do the pros outweigh the risks?

Contributors:

- Eric Kim, Zekai Wang.

