AI - Large Language Models

CS 246

Objectives

- Students will be able to:
  - decipher the acronym LLM
  - explain how LLMs work, in general terms
  - appreciate the A in AI

Terminology

- **Artificial Narrow Intelligence**
  - a form of AI, based on machine learning, that is designed to solve a specific task
  - e.g., facial recognition, self-driving cars, LLMs (like ChatGPT and Claude)
- **Artificial General Intelligence**
  - a system that could reason, plan, and learn any intellectual task, and adapt to new situations
  - Does not exist yet, although LLMs sometimes appear to exhibit AGI through emergent behavior

Terminology

- **Emergent behavior**
    - actions that are not explicitly programmed in or expected, but emerge (once the LLMs have been trained on a sufficiently large training set)
        - solve complicated math and logic problems
        - debug and write code in numerous languages
        - generate an appiconset 😍

Overview of How LLMs Work

- LLMs are ANI, and a particular application of machine learning, using neurons, backpropagation and gradient descent, just like the models we have studied previously
- An LLM begins by looking at all the **tokens** -- words, parts of words, or punctuation -- in a prompt
- Using patterns learned during training, it predicts the next token, then the next token after that, etc., until it has completed a response
- Training is based on *massive* amounts of text gleaned from books, websites, articles and other sources*
- It learns statistical patterns about how language works -- which words tend to follow each other, how sentences are structured, what makes coherent arguments, and factual relationships between concepts
- The model is shown billions of text examples, and adjusting weights to improve its ability to predict the next token
- Cutting edge models can take 2-6 months**, and the model can have hundreds of billions or trillions of parameters

*not all of whose authors are thrilled by this - exercise for next class, dig up and present some references

**that's *continuous* computation on thousands of high-end GPUs or AI chips; just the electricity bill runs into the millions of dollars!

LLM Architecture

- Modern LLMs use a **transformer** architecture, first spelled out in 2017 by researchers at Google, in a paper, "Attention is All You Need"

- Before transformers, the best models of the time read text one word at a time

- Transformers look at all the words at a time, and decide which ones are most important to understanding the whole

- It is called attention

Attention - an Example

- "The bird saw the man with the binoculars"
- Does "with the binoculars" refer to the man or the bird?
- Humans, from context, know to pay attention to "man" and "binoculars", not "bird" and "binoculars"
- Transformers use attention scores -- numbers which indicate how much weight to give to other words, i.e., how much they are related
- That's the *basic* idea and, because this is just an *intro*, is as far as we need to go

Interacting with LLMs Using an API

https://platform.openai.com/docs/libraries