Large Language Models & ChatGPT

Introduction
Fall 2023 | Aug 28

Why this class and why now?

- WWII 1940’s
- Imitation Game/Turing Test 1950
- Hodgkin/Huxley model of the brain 1952
- Perceptron (single layer Neural Network) 1957
- ELIZA (conversational, boolean chatbot) 1964
- N-gram models on computers 1990’s
- Recurrent Neural Networks (RNN) with Long Short Term Memory (LSTM) 1997
- Siri 2011

Why this class and why now?

Winograd Schema Challenge

The trophy doesn't fit in the brown suitcase because it is too large. What is too large?

- The trophy
- The suitcase

2016:

- 58% of NLP programs can get this right

2017:

- Google publishes "Attention is all you need". Transformers (the T in GPT) are invented
- 75% get this right (BERT)

Why this class and why now?

2018

- Google publishes BERT - SOTA by 2020. Bidirectional model. 110 - 340 million parameters
- OpenAI introduces GPT-1, a unidirectional conversational model. 117 million params

2019

- GPT-2 available to researchers (too toxic to make public), 1.5 billion params

2022 (Nov)

- GPT-3, 117 billion params, safety guards, a GUI, and an explosion in NLP

Introduction

Welcome

Course overview

Name, Program, P(Doom)

P(Doom)

- Shorthand for a philosophical stance
- *P(Doom) ≅ P(existential demise)*

To answer this question

- What is existential demise?
- What else are we assuming?

APS Systems (Carlsmith, 2021)

- “It will become possible and financially feasible to build AI systems with the following properties:
  - **Advanced capability:** they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today’s world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).
  - **Agentic planning:** they make and execute plans, in pursuit of objectives, on the basis of models of the world.
  - **Strategic awareness**: the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.

What not to do (Tegmark, 2023)

- **Don’t teach it to code:** this facilitates recursive self-improvement
- **Don’t connect it to the internet:** let it learn only the minimum needed to help us, not how to manipulate us or gain power
- **Don’t give it a public API**: prevent nefarious actors from using it within their code
- **Don’t start an arms race: t**his incentivizes everyone to prioritize development speed over safety

The False Promise of ChatGPT (Chomsky et al., 2023)

- The primary fears revolve around superintelligence
- Superintelligence involves rational thought
- LLM’s are stochastic parrots
- Stochastic parrots cannot think and cannot reason
  - ✓ Description: “The apple falls.”
  - ✓ Prediction: “The apple will fall if I open my hand.”
  - ✗ Causal “Any such object would fall, because of the force of gravity” or “because of the curvature of space-time”
  - ✗ Thinking: “The apple would not have fallen but for the force of gravity.”

Constitutional A.I. (Anthropic)

- Start training a model (Claude) with a “Constitution” - or list of rules for how to behave
- Train Claude as usual
- Train another model only to enforce the Constitution and censor Claude whenever Claude violates the constitution.
- Never deploy Claude without its enforcer.

Some say Claude is very reticent to respond.

How much of the fear is marketing? (Merchant, 2023)

- “One of the biggest harms of large language models is caused by claiming that LLMs have ‘human-competitive intelligence,’” -Timnit Gebru
- Everyone wants to try a technology that promises to wipe out humanity
- If everyone is doing it, FOMO hits companies hard
    - If company A drives labor costs down, company B has to, too
- Discriminatory, racist, sexist material will be written, misinformation will be spread. The question is how much stake do we put into it?

What is your P(Doom)?

What is a language model anyway?

Words and probabilities

- Imagine you are writing an email.
- You start a sentence "Thanks for the update..."
- What do you predict the next word should be?

- Language models complete this task by assigning a probability to each possible next word.
- N-gram language models are the simplest.

N-gram language model (Jurafsky & Martin, 2023)

- **Goal**: Predict the probability of a sentence.
- **Why**?
  - Machine Translation
    - P(we must vote) > P(our must vote)
  - Spelling correction
    - P(it's getting late) > P(it's geting laet)
  - Speech Recognition
    - P(We went to catch up) > P(We went to ketchup) [for a laugh on Buzzfeed]
  - Predictive Text
    - Emails
  - Question Answering, etc.

N-gram language models

Goal: Predict probability of a sentence.

$$P(W) = P(w1,w2,w3,w4,w5...wn)$$

Subgoal: Predict the probability of of a word, given all the other words

$$P(w5|w1,w2,w3,w4)$$

Conditional Probabilities & The Chain Rule

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$ Rewrites as $$P(A \cap B) = P(A) \cdot P(B \mid A)$$

The probability of A and B (intersection) is the probability of A times the probability of B given A

P(“it was the best of times”) =

P(it) × P(waslit) × P(thelit was) × P(bestlit was the)

× P(oflit was the best) × P(timeslit was the best of)

How do we come up with those probabilities?

- Big database with all the sentences possible in the world stored?
  - No! Too many sentences
- Average raw word counts across the language?
  - No! Not enough word pairs
- Magic?

The Markov Assumption

Simplifying assumption: The future depends only on the present.

P(times | it was the best of) ≅ P(times | of) ← Bigram

P(times | it was the best of) ≅ P(times | best of) ← Trigram

Maximum Likelihood Estimation MLE

- All and only the words in the corpus
- Calculate the number of times each word appears in a context
- Return a probability for that word in that context
- The probability is the MLE

Calculating Probabilities

Let’s assume that our entire corpus is the first 4 clauses of *A Tale of Two Cities*

<s>It was the best of times</s>

<s>it was the worst of times</s>

<s>it was the age of wisdom</s>

<s>it was the age of foolishness</s>

The Probabilities

P(it | <s>) = 4/4 = 1.0

P(was | it) = 4/4 = 1.0

P(the | was) = 4/4 = 1.0

The Probabilities

P(it | <s>) = 4/4 = 1.0

P(was | it) = 4/4 = 1.0

P(the | was) = 4/4 = 1.0

P(best | the) = 1/4 = .25

P(worst | the) = 1/4 = .25

P(age | the) = 2/4 = .5

P(of | best) = 1/4 = .25

P(of | worst) = 1/4 = .25

P(of | age) = 2/4 = .5

P(times | of) = 2/4 =.5

P(wisdom | of) = 1/4 =.25

P(foolishness | of) = 1/4 = .25

P(<s> it was the best of times</s>)

P(it | <s>) = 4/4 = 1.0

P(was | it) = 4/4 = 1.0

P(the | was) = 4/4 = 1.0

P(best | the) = 1/4 = .25

P(of | best) = 1/4 = .25

P(times | of) = 2/4 =.5

= .03125