In-context Learning

David Mareček

January 4, 2024
• Current NLP methods use **Large Language Models (LLMs)** based on **Generative Pretrained Transformers (GPT)**.

• These models are trained in a *supervised* way, but then, they are often used in a *semi-supervised* or even *unsupervised* way, which is called *few-shot* or *zero-shot* setting.

**This lecture is about:**

• Generative Pretrained Transformers (GPT) – what it does, how is it trained, GPT evolution through time.

• In-context learning (ICL) – learning from examples given in the prompt.

• Chain of Thought (CoT) – forcing the model to think step-by-step.
Generative Pre-trained Transformer (GPT)

- Based on the given context \((E_1, \ldots E_N)\), it predicts the next token \(T_N\).
- 12 to 96 layers. At each layer, a token state attends to (looks at) the previous layer states of all preceding tokens.
- Trained on huge textual data across different languages (commoncraw, books, Wikipedia).
Generative Pre-trained Transformer (GPT)

- After each **Attention layer**, there is one **Feed-Forward** layer.
- **Residual Connections** – bridges across the attention and feed-forward layers.
- **Positional Encoding** adds the information about the position of the input token.
Evolution of GPT models

- **2018**
  - Jun. 11th
  - GPT-1
    - The first version of GPT was released

- **2019**
  - Feb. 14th
  - GPT-2
    - The second version of GPT was released

- **2020**
  - May 28th
  - GPT-3
    - Initial GPT-3 preprint paper was published at arXiv. API became publicly available on Nov. 18th, 2021

- **2022**
  - Nov. 30th
  - ChatGPT
    - ChatGPT was announced on OpenAI blog. ChatGPT API became available on Mar. 1st, 2023

- **2023**
  - Mar. 14th
  - GPT-4
    - GPT-4 was released via ChatGPT. API will be publicly available soon.
Evolution of GPT models

<table>
<thead>
<tr>
<th></th>
<th>GPT-1</th>
<th>GPT-2</th>
<th>GPT-3</th>
<th>GPT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>117M</td>
<td>1.5B</td>
<td>175B</td>
<td>??? (1T)</td>
</tr>
<tr>
<td>Decoder Layers</td>
<td>12</td>
<td>48</td>
<td>96</td>
<td>???</td>
</tr>
<tr>
<td>Context Token Size</td>
<td>512</td>
<td>1024</td>
<td>2048</td>
<td>32,768</td>
</tr>
<tr>
<td>Hidden Layer</td>
<td>768</td>
<td>1600</td>
<td>12288</td>
<td>???</td>
</tr>
</tbody>
</table>

- **GPT-2** – The model is able to generate locally coherent text. It can be finetuned to solve another tasks.
- **GPT-3** – Harder to recognize whether a text was generated by human or by GPT-3. It can be used to solve many tasks by prompting couple of examples (in-context learning)
- **GPT-4** – Reinforcement learning by human feedback (RLHF). Many tasks can be solved simply by asking a question.
In-Context Learning

• Learning from examples in the context.

In-Context Learning (ICL)

- Examples written in natural language provide an interpretable interface to communicate with LLMs.
- Much easier to incorporate human knowledge into LLMs by changing the examples and templates.
- Simulates the decision process of human beings by learning from analogy.
- Compared with supervised training, ICL is a training-free learning framework.
- It greatly reduces the computation costs for adapting the model to new tasks.
Why does In-Context Learning work?

- After pre-training, LLMs can exhibit intriguing ICL capabilities (emergent capabilities) without being updated.
- While intuitively reasonable, the working mechanism of the ICL remains unclear.
- Studies showed that the ICL ability grows as the parameters of LLMs increase from 0.1 billion to 175 billion.
- It has been shown that the performance of ICL heavily depends on the source of pre-training corpora rather than the scale.
- Some researches showed that ICL behaves similarly to fine-tuning from multiple perspectives.
- Another research showed that LM does in-context learning by using the prompt to “locate” the relevant concept it has learned during pre-training to do the task.
Few-shot, one-shot, zero-shot

- **Few-shot** - The task is defined and a few examples are given to the prompt followed by the question.
  Translate English to German:
  bread => das Brot
  apple => der Apfel
  potatoe => die Kartoffel
  cheese =>

- **One-shot** - Only one example is given.
  Translate English to German:
  bread => das Brot
  cheese =>

- **Zero-shot** - Only the task definition and the question.
  Translate English to German
  cheese =>
Chain-of-Thought prompting

• The standard prompting techniques do not perform well on complex reasoning tasks
  • arithmetic reasoning
  • commonsense reasoning
  • symbolic reasoning

• Chain-of-Thought (CoT) incorporates intermediate reasoning steps that can lead to the final output into the prompts.
Chain-of-Thought prompting

**Standard Prompting**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

**Chain-of-Thought Prompting**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. \(5 + 6 = 11\). The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Chain-of-Thought Examples

Math Word Problems (free response)
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Math Word Problems (multiple choice)
Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788
A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)
Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock
A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA
Q: Yes or no: Would a pear sink in water?
A: The density of a pear is about 0.6 g/cm$^3$, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding
Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?
A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding
Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."
A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (instructing a robot)
Human: How would you bring me something that isn’t a fruit?
Explanation: the user wants something to eat that isn’t a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation
Q: Take the last letters of the words in "Lady Gaga" and concatenate them.
A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?
A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

**Few-shot vs. Zero-shot**

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: 

(Output) The answer is 8. X

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: 

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 X

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Two-step prompting in CoT

Why LLMs Can Perform CoT Reasoning?

• still active area of research

• It is widely hypothesized that it can be attributed to training on code since models trained on it show a strong reasoning ability.

• A recent study showed that both the presence of arithmetic equations for mathematical reasoning and textual description of individual steps are essential for the model performance.
Greedy decode
This means she uses $3 + 4 = 7$ eggs every day. She sells the remainder for $2$ per egg, so in total she sells $7 \times 2 = 14$ per day.
The answer is $14$.

Sample a diverse set of reasoning paths

- She has $16 - 3 - 4 = 9$ eggs left. So she makes $2 \times 9 = 18$ per day.
The answer is $18$.

- This means she sells the remainder for $2 \times (16 - 4 - 3) = 26$ per day.
The answer is $26$.

- She eats 3 for breakfast, so she has $16 - 3 = 13$ left. Then she bakes muffins, so she has $13 - 4 = 9$ eggs left. So she has $9$ eggs * $2 = 18$.
The answer is $18$.

Marginalize out reasoning paths to aggregate final answers