Generative Pre-trained Transformer

Jiachang Liu

Duke University

February 5, 2021
Table of Contents

1. GPT-2
2. GPT-3
3. DALL·E
Table of Contents

1 GPT-2

2 GPT-3

3 DALL·E
**Motivation**

- **Problem**: Current NLP systems are brittle and sensitive to slight changes in data distributions. They are better characterized as narrow experts.
- **Goal1**: Move towards more general systems which can perform many tasks
- **Goal2**: No need for manually labeled dataset for each task.
Multi-task learning: potential solution towards building general NLP systems for many tasks:

Training objective:

$$\max_\theta \mathbb{E}_{input, output, task \sim D} \log p(output | input, task)$$ (1)

Drawback: still requires hundreds of thousands of manually curated examples
Generative Pre-Training

- Factorization of conditional probabilities:

\[ p(x) = \prod_{i=1}^{n} p(s_i | s_1, s_2, \ldots, s_{i-1}) \]  \hspace{1cm} (2)

where \((s_i | s_1, s_2, \ldots, s_n)\) is a sequence of symbols

- **Key insight**: language provides a flexible way to specify tasks, inputs, and outputs all as a sequence of symbols
  - (translate to french, english text, french text)
  - (answer the question, document, question, answer).

- “Language modeling is able to in principle, to learn the task without the need for explicit supervision of which symbols are the outputs to be predicted. The supervised objective is the same as the unsupervised objective but only evaluated on a subset of the sequence.”
Examples of Naturally Occurring Tasks

"I’m not the cleverest man in the world, but like they say in French: Je ne suis pas un imbécile [I’m not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "Mentez mentez, il en restera toujours quelque chose," which translates as, "Lie lie and something will always remain."

“I hate the word ‘perfume,’” Burr says. ‘It’s somewhat better in French: ‘parfum.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “-Comment on fait pour aller de l’autre coté? -Quel autre coté?”, which means “- How do you get to the other side? - What side?”.

If this sounds like a bit of a stretch, consider this question in French: As-tu aller au cinéma?, or Did you go to the movies?, which literally translates as Have-you to go to movies/theater?

“Brevet Sans Garantie Du Gouvernement”, translated to English: “Patented without government warranty”.

Figure: Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.
Engineering Effort

- Bigger model size: transformer decoder

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Layers</th>
<th>$d_{model}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>117M</td>
<td>12</td>
<td>768</td>
</tr>
<tr>
<td>345M</td>
<td>24</td>
<td>1024</td>
</tr>
<tr>
<td>762M</td>
<td>36</td>
<td>1280</td>
</tr>
<tr>
<td>1542M</td>
<td>48</td>
<td>1600</td>
</tr>
</tbody>
</table>

- Byte Pair Encoding (BPE)
Experimental Results I

<table>
<thead>
<tr>
<th></th>
<th>LAMBADA (PPL)</th>
<th>LAMBADA (ACC)</th>
<th>CBT-CN (ACC)</th>
<th>CBT-NE (ACC)</th>
<th>WikiText2 (PPL)</th>
<th>PTB (PPL)</th>
<th>enwik8 (BPB)</th>
<th>text8 (BPC)</th>
<th>WikiText103 (PPL)</th>
<th>1BW (PPL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA</td>
<td>99.8</td>
<td>59.23</td>
<td>85.7</td>
<td>82.3</td>
<td>39.14</td>
<td>46.54</td>
<td>0.99</td>
<td>1.08</td>
<td>18.3</td>
<td>21.8</td>
</tr>
<tr>
<td>117M</td>
<td>35.13</td>
<td>45.99</td>
<td>87.65</td>
<td>83.4</td>
<td>29.41</td>
<td>65.85</td>
<td>1.16</td>
<td>1.17</td>
<td>37.50</td>
<td>75.20</td>
</tr>
<tr>
<td>345M</td>
<td>15.60</td>
<td>55.48</td>
<td>92.35</td>
<td>87.1</td>
<td>22.76</td>
<td>47.33</td>
<td>1.01</td>
<td>1.06</td>
<td>26.37</td>
<td>55.72</td>
</tr>
<tr>
<td>762M</td>
<td>10.87</td>
<td>60.12</td>
<td>93.45</td>
<td>88.0</td>
<td>19.93</td>
<td>40.31</td>
<td>0.97</td>
<td>1.02</td>
<td>22.05</td>
<td>44.575</td>
</tr>
<tr>
<td>1542M</td>
<td>8.63</td>
<td>63.24</td>
<td>93.30</td>
<td>89.05</td>
<td>18.34</td>
<td>35.76</td>
<td>0.93</td>
<td>0.98</td>
<td>17.48</td>
<td>42.16</td>
</tr>
</tbody>
</table>

**Figure:** Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. **LAMBADA:** The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict. **CBT:** predict which of 10 possible choices for an omitted word is correct.
Experimental Results II - Winograd Schema Challenge

**Figure:** Performance on the Winograd Schema Challenge as a function of model capacity.

**Example:** The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.  

a) If the word is “feared”, then “they” presumably refers to the city council;  
b) if it is “advocated” then “they” presumably refers to the demonstrators.
Tasks Which GPT-2 Doesn’t Do Well

- Reading Comprehension (Answer questions based on a given dialogue).
- Summarization
- Machine Translation
- Question Answering
Table of Contents

1. GPT-2
2. GPT-3
3. DALL·E
GPT-3, an extension of GPT-2

- What if we use more data and bigger model?

<table>
<thead>
<tr>
<th>Model Name</th>
<th>$n_{\text{params}}$</th>
<th>$n_{\text{layers}}$</th>
<th>$d_{\text{model}}$</th>
<th>$n_{\text{heads}}$</th>
<th>$d_{\text{head}}$</th>
<th>Batch Size</th>
<th>Learning Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3 Small</td>
<td>125M</td>
<td>12</td>
<td>768</td>
<td>12</td>
<td>64</td>
<td>0.5M</td>
<td>$6.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 Medium</td>
<td>350M</td>
<td>24</td>
<td>1024</td>
<td>16</td>
<td>64</td>
<td>0.5M</td>
<td>$3.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 Large</td>
<td>760M</td>
<td>24</td>
<td>1536</td>
<td>16</td>
<td>96</td>
<td>0.5M</td>
<td>$2.5 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 XL</td>
<td>1.3B</td>
<td>24</td>
<td>2048</td>
<td>24</td>
<td>128</td>
<td>1M</td>
<td>$2.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 2.7B</td>
<td>2.7B</td>
<td>32</td>
<td>2560</td>
<td>32</td>
<td>80</td>
<td>1M</td>
<td>$1.6 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 6.7B</td>
<td>6.7B</td>
<td>32</td>
<td>4096</td>
<td>32</td>
<td>128</td>
<td>2M</td>
<td>$1.2 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 13B</td>
<td>13.0B</td>
<td>40</td>
<td>5140</td>
<td>40</td>
<td>128</td>
<td>2M</td>
<td>$1.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 175B or “GPT-3”</td>
<td>175.0B</td>
<td>96</td>
<td>12288</td>
<td>96</td>
<td>128</td>
<td>3.2M</td>
<td>$0.6 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Quantity (tokens)</th>
<th>Weight in training mix</th>
<th>Epochs elapsed when training for 300B tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Crawl (filtered)</td>
<td>410 billion</td>
<td>60%</td>
<td>0.44</td>
</tr>
<tr>
<td>WebText2</td>
<td>19 billion</td>
<td>22%</td>
<td>2.9</td>
</tr>
<tr>
<td>Books1</td>
<td>12 billion</td>
<td>8%</td>
<td>1.9</td>
</tr>
<tr>
<td>Books2</td>
<td>55 billion</td>
<td>8%</td>
<td>0.43</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3 billion</td>
<td>3%</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Review: GPT-2 Zero-shot

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1. Translate English to French:

2. cheese => ........................................
```

- task description
- prompt
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1. Translate English to French:
2. sea otter => loutre de mer
3. cheese => ........................................
```

task description
element
prompt

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1. Translate English to French:
2. sea otter => loutre de mer
3. peppermint => menthe poivrée
4. plush giraffe => girafe peluche
5. cheese => ........................................
```

task description
examples
prompt
Figure: Alice was friends with Bob. Alice went to visit her friend [BLANK]. → Bob. George bought some baseball equipment, a ball, a glove, and a [BLANK]. → ?
### GPT-3 Results on Question Answering

<table>
<thead>
<tr>
<th>Setting</th>
<th>NaturalQS</th>
<th>WebQS</th>
<th>TriviaQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAG (Fine-tuned, Open-Domain) [LPP⁺ 20]</td>
<td>44.5</td>
<td>45.5</td>
<td>68.0</td>
</tr>
<tr>
<td>T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]</td>
<td>36.6</td>
<td>44.7</td>
<td>60.5</td>
</tr>
<tr>
<td>T5-11B (Fine-tuned, Closed-Book)</td>
<td>34.5</td>
<td>37.4</td>
<td>50.1</td>
</tr>
<tr>
<td>GPT-3 Zero-Shot</td>
<td>14.6</td>
<td>14.4</td>
<td>64.3</td>
</tr>
<tr>
<td>GPT-3 One-Shot</td>
<td>23.0</td>
<td>25.3</td>
<td>68.0</td>
</tr>
<tr>
<td>GPT-3 Few-Shot</td>
<td>29.9</td>
<td>41.5</td>
<td>71.2</td>
</tr>
</tbody>
</table>

#### TriviaQA

![Graph showing accuracy vs. parameters in billions for different training scenarios](chart.png)

- **Zero-Shot**
- **One-Shot**
- **Few-Shot (K=64)**

- **Fine-tuned SOTA**
# GPT-3 Results on Machine Translation

<table>
<thead>
<tr>
<th>Setting</th>
<th>En→Fr</th>
<th>Fr→En</th>
<th>En→De</th>
<th>De→En</th>
<th>En→Ro</th>
<th>Ro→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA (Supervised)</td>
<td>45.6</td>
<td>35.0</td>
<td>41.2</td>
<td>40.2</td>
<td>38.5</td>
<td>39.9</td>
</tr>
<tr>
<td>XLM [LC19]</td>
<td>33.4</td>
<td>33.3</td>
<td>26.4</td>
<td>34.3</td>
<td>33.3</td>
<td>31.8</td>
</tr>
<tr>
<td>MASS [STQ+19]</td>
<td>37.5</td>
<td>34.9</td>
<td>28.3</td>
<td>35.2</td>
<td>35.2</td>
<td>33.1</td>
</tr>
<tr>
<td>mBART [LGG+20]</td>
<td>-</td>
<td>29.8</td>
<td>34.0</td>
<td>35.0</td>
<td>30.5</td>
<td></td>
</tr>
<tr>
<td>GPT-3 Zero-Shot</td>
<td>25.2</td>
<td>21.2</td>
<td>24.6</td>
<td>27.2</td>
<td>14.1</td>
<td>19.9</td>
</tr>
<tr>
<td>GPT-3 One-Shot</td>
<td>28.3</td>
<td>33.7</td>
<td>26.2</td>
<td>30.4</td>
<td>20.6</td>
<td>38.6</td>
</tr>
<tr>
<td>GPT-3 Few-Shot</td>
<td>32.6</td>
<td>39.2</td>
<td>29.7</td>
<td>40.6</td>
<td>21.0</td>
<td>39.5</td>
</tr>
</tbody>
</table>

![Translation (Multi-BLEU)](chart.png)

- French -> English
- English -> French
- German -> English
- English -> German
- Romanian -> English
- English -> Romanian

Parameters in LM (Billions) vs BLEU Score.
Figure: BERT++ was first fine-tuned on MultiNLI (392K examples) and SWAG (113K examples) before further fine-tuning on the SuperGLUE training set.
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination

After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church’s annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church’s history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.
Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.

Poor English input: Thank you for picking me as your designer. I’d appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.

Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.

Poor English input: I’d be more than happy to work with you in another project.
Good English output: I’d be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you’re looking for and that’d be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you’re looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.
Poor English input:  We think that Leslie likes ourselves.
Good English output:  We think that Leslie likes us.

Poor English input:  Janet broke Bill on the finger.
Good English output:  Janet broke Bill’s finger.

Poor English input:  Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output:  Mary arranged for John to rent a house in St. Louis.

Poor English input:  Today I have went to the store to to buys some many bottle of water.
Good English output:  Today I went to the store to buy some bottles of water.

Poor English input:  I have tried to hit ball with bat, but my swing is has miss.
Good English output:  I tried to hit the ball with the bat, but my swing missed.

**Figure:** In the third example, the model not only adjusts grammar, but also removes the word ?cheap? in a way that alters meaning.
GPT-3 Arithmetic

![Graph showing the accuracy of arithmetic operations across different parameter sizes.](image)
Some Reflection on GPT-3

- From the example of doing arithmetic, GPT-3 is definitely memorizing training data it has seen.
- Nonetheless, GPT-3 is good at learning from in-context examples across a wide range of tasks.
- This in-context few-shot learning ability is new to the ML and NLP community. The model’s parameters do not only contain factual data but also many language patterns.
Table of Contents

1. GPT-2
2. GPT-3
3. DALL·E
DALL·E is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text-image pairs. It has a diverse set of capabilities, including:

- creating anthropomorphized versions of animals and objects
- combining unrelated concepts in plausible ways
- rendering text
- applying transformations to existing images

DALL·E is a simple decoder-only transformer that receives both the text and the image as a single stream of 1280 tokens—256 for the text and 1024 for the image—and models all of them autoregressively.
an illustration of a baby daikon radish in a tutu walking a dog
a snail made of harp, a snail with the texture of a harp.
a store front that has the word 'pytorch' written on it. pytorch store front. 'pytorch'. pytorch typography.
WALL·E Image Transformation

**TEXT PROMPT**

the exact same cat on the top as a sketch on the bottom

**IMAGE PROMPT**

![Image Grid]

**AI-GENERATED IMAGES**

![Image Grid]