Session 3: NLP and Transformers

Youssef Mohamed
What is NLP?

“How computers can be used to understand and manipulate natural language text or speech to do useful things”[1]

HOW?

• Tokenization
• Part Of Speech (POS) Tagging
• Chunking

Applications

• Sentiment Analysis
• Speech Recognition
• Translation

[3] https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a075d52e2ba
Improving Language Understanding by Generative Pre-Training

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever
OpenAI
Pretraining

- BooksCorpus (800M words)
  - List of unpublished books

\[
L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)
\]

(1)

\[
h_0 = UW_e + W_p
\]

\[
h_i = \text{transformer\_block}(h_{i-1}) \forall i \in [1, n]
\]

(2)

Masked Input

(mask the words appearing later so the attention network can't use them)

Le  Chat  Est  Noir

Le  Chat  Est  Noir

Le  Chat  Est  Noir

Le  Chat  Est  Noir

Le  Chat  Est  Noir

Le  Chat  Est  Noir
Fine-tuning

\[ L_2(C) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m). \quad (4) \]

\[ P(y|x^1, \ldots, x^m) = \text{softmax}(h^m_l W_y). \quad (3) \]

\[ L_3(C) = L_2(C) + \lambda * L_1(C) \quad (5) \]
Performance

- NLI: showing the relationship between sentences (entailment, contradiction or neutral)
- QA: a passage then MCQ based on it
- SS: if two sentences are semantically similar or not
- C: grammatically correct or not,

“GPT-1 performed better than specifically trained supervised state-of-the-art models in 9 out of 12 tasks the models were compared on”
Task-Specific Input Transformations

- Start and end tokens
- Delimiter added so input could be sent as ordered sequence
- Minimal changes to the model
Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei
Overview

• GPT-1 vs GPT-3
  – Context window size was increased from 512 for GPT-1 to 2048 tokens for GPT-3
  – 96 layers with each layer having 96 attention heads compared to 12 in GPT-1
  – 117M parameter in GPT-1 and 175B in GPT-3

• The usual pretraining and fine tuning (or is it ?)
Few shot learning

- Model does not need fine tuning
- only examples of the task
Data-sets

- 3.14E+23 flops in total
- 355 years!! to train GPT-3 on a V100 [1]
- $4,600,000!! to train GPT-3 using the lowest cost GPU cloud provider [1]

<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>

Tasks and Results
Question Answering

- Performs better in factual question answering
- Better than Open-Domain model

Example 1
Question: what color was john wilkes booth's hair
Long answer: Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.
Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.
Short answer: BOOLEAN:NO

<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>
Translation

- To English performs better
Reading Comprehension (Reasoning)

- Performs worse for reasoning tasks.

<table>
<thead>
<tr>
<th></th>
<th>SuperGLUE Average</th>
<th>BoolQ Accuracy</th>
<th>CB Accuracy</th>
<th>CB F1</th>
<th>COPA Accuracy</th>
<th>RTE Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tuned SOTA</td>
<td>89.0</td>
<td>91.0</td>
<td>96.9</td>
<td>93.9</td>
<td>94.8</td>
<td>92.5</td>
</tr>
<tr>
<td>Fine-tuned BERT-Large</td>
<td>69.0</td>
<td>77.4</td>
<td>83.6</td>
<td>75.7</td>
<td>70.6</td>
<td>71.7</td>
</tr>
<tr>
<td>GPT-3 Few-Shot</td>
<td>71.8</td>
<td>76.4</td>
<td>75.6</td>
<td>52.0</td>
<td>92.0</td>
<td>69.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>WiC Accuracy</th>
<th>WSC Accuracy</th>
<th>MultiRC Accuracy</th>
<th>MultiRC F1a</th>
<th>ReCoRD Accuracy</th>
<th>ReCoRD F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tuned SOTA</td>
<td>76.1</td>
<td>93.8</td>
<td>62.3</td>
<td>88.2</td>
<td>92.5</td>
<td>93.3</td>
</tr>
<tr>
<td>Fine-tuned BERT-Large</td>
<td>69.6</td>
<td>64.6</td>
<td>24.1</td>
<td>70.0</td>
<td>71.3</td>
<td>72.0</td>
</tr>
<tr>
<td>GPT-3 Few-Shot</td>
<td>49.4</td>
<td>80.1</td>
<td>30.5</td>
<td>75.4</td>
<td>90.2</td>
<td>91.1</td>
</tr>
</tbody>
</table>

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Premise: I knocked on my neighbor’s door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.
**Made up tasks**

**Arithmetic**
- Gap between three digit addition and four digit addition?

**Article Generation**
- Performs worse the bigger the model is
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
BERT
Pre-training

Masked LM
- Random masked words
- Each embedding is a word integrated with segment and position
- T is a word vector

Next Sentence Prediction (NSP)
- If two sentences follow each other
- Binary classification (C)
Data sets

- BooksCorpus (800M words)
  - List of unpublished books

- English Wikipedia (2,500M words)
  - Wikipedia articles
Fine-tuning

- Supervised training based on the task
- Replacing the output layer
- Modifying the input layer
- Start and end words