

Language Models are Unsupervised Multitask Learners

Author: Alec Radford

OpenAI

Presenter: Faizan Ahmad

<https://qdata.github.io/deep2Read>

Outline

- 1 Introduction
- 2 Related Work
- 3 OpenAI Generative Pre-Training (GPT) 2
- 4 Evaluation and Results
- 5 Discussion

- **Pre-training** and **fine tuning** - current trend in NLP
- Can we do without fine tuning?
- Can a single model learn multiple tasks?
- Language models to the rescue

- Task Specific Architectures
 - Last 7-10 years
- Single Model Finetuned on Different Tasks
 - BERT by Google
 - OpenAI GPT
- Single Model for Multiple Tasks without Finetuning
 - Reading Comprehension

OpenAI Generative Pre-Training (GPT) 2

- Language modeling at the core
- Train a large language model and solve multiple tasks with it

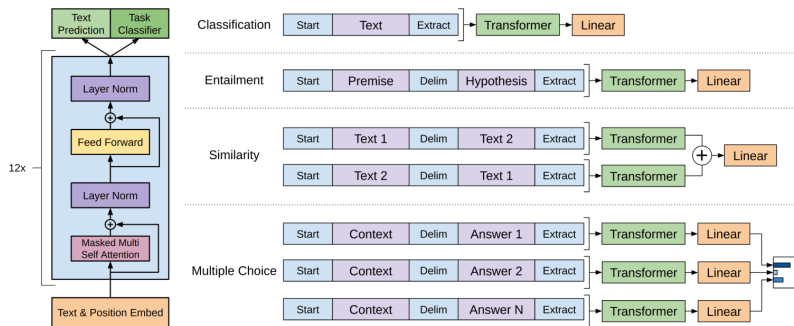


Figure: GPT-1 Model

OpenAI Generative Pre-Training (GPT) 2

- Which dataset to use?
- Character based input or word based?
- How many parameters?

OpenAI Generative Pre-Training (GPT) 2

Dataset (WebText)

- Current large data sets have bad quality
- Need new quality data sets
- How?
 - Scrap links from Reddit
 - Keep the ones with more than 3 likes
 - 45 Million Links, 8 Million Documents

OpenAI Generative Pre-Training (GPT) 2

Input

- Word level input converges slowly
- Character level input does not work very well
- Need a middle ground
- Byte Pair Encoding to the rescue (Sub Words)

r ·	→	r·
l o	→	lo
lo w	→	low
e r·	→	er·

Figure 1: BPE merge operations learned from dictionary {‘low’, ‘lowest’, ‘newer’, ‘wider’}.

Figure: BPE

OpenAI Generative Pre-Training (GPT) 2

Parameters

- Bigger is better

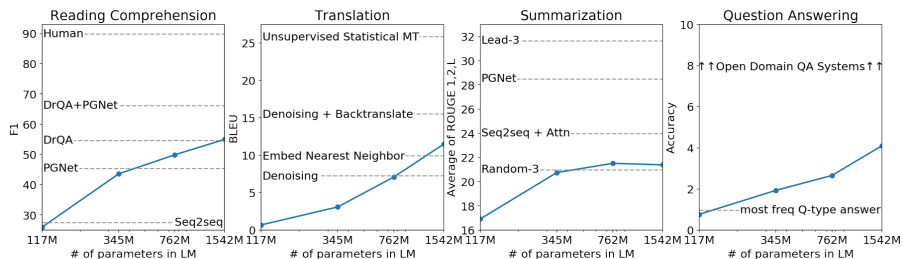


Figure: Effect of Parameters

Results

Zero-shot Setting

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

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2

Figure: Summary Results

Results

Children's Book Test

- Fill in the blanks
- 10 Choices
- New state of the art results

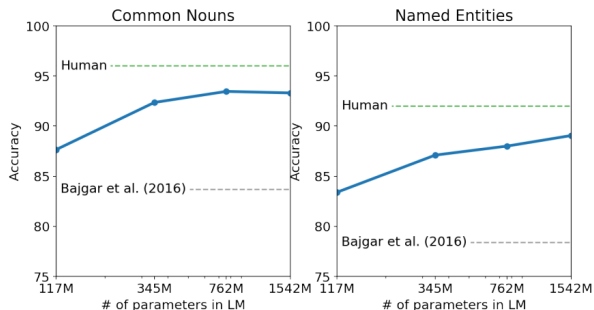


Figure: CBT Results

- Fill the last word of a sentence with at least 50 tokens
- New state of the art results

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

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2

Figure: Summary Results

Results

Qualitative Results

Context (WebText test)

Pinky

The pink ghost's AI is designed to "feel" opposite of the red ghost's behavior. Pinky actually attempts to get out in front of Pac-Man. This is accomplished by setting the target 4 tiles ahead of Pac-Man's current location in the direction that Pac-Man is travelling. One exception to this is when Pac-Man is traveling up. Due to an overflow bug in the code, the calculation includes a left offset equal to the expected up offset.

The use of Pac-Man's expected position allows the player to exploit the navigation code for Pinky. This was most often seen in situations where Pac-Man and Pinky are in close proximity to each other. For example, if they were headed toward one another, once they are less than 4 tiles apart, the target tile is now behind Pinky. If he encounters an intersection while in this configuration, he will attempt to double-back (by turning) in order to reach the target tile.

Player's would often perform what was referred to as a "head fake" by changing directions rapidly just as Pinky was approaching an intersection. By displacing the target tile just as Pinky

Small model completion

was coming to the left, they would get the "correct" direction.

If the location is actually farther from what Pac-Man was at, the player's will not respond to the ghost's movement, and the ghost will instead return to the player's current location. This is because Pac-Man does not move by turning.

The ghost will have to go to another location of some kind to return where he began, and this is achieved in many ways. It will often simply leave the location, as

GPT-2 completion

was about to cross the intersection, Pac-Man would be able to dodge the ghosts' projectiles and return to the safe location of the safe house.

Buster

The yellow ghost's AI is specifically designed to try to avoid Pac-Man from leaving the safe house. The player must press the SPACE bar to reveal the safe house's doors, then go into them to find the yellow ghost hiding in the next room.

- New trend in deep learning for NLP
- High quality text generation
- **Adversarial attacks on these models?**