Language Models are Unsupervised Multitask Learners

Written by: Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever

From OpenAI

Presented by: **Ehsan Amjadian from RBC**

Summary

- Many NLP tasks often treated as supervised (explicit supervision)
  - Summarization
  - Question Answering
  - Reading Comprehension
  - Machine Translation
- Can language modeling be utilized for these tasks
  - Zero shot learning
  - Implicit supervision
  - Utilizing WebText
- Capacity of the model:
  - has a log-linear relationship with its performance on the tasks
- Largest model:
  - 1.5 Billion Parameters Transformer
- Generalization
  - Past task specific methods lack generalization

Brief Advantages of the Model

- State of the art LM
- Underfits WebText
- Coherent generated text
- Zero-shot learning
- Transfer learning

Language Model

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

Ordinary Language Models: p(output|input)
Multitask Learner Language Model p(output|input, task)

MT Example:
(translate to french, english text, french text)
Reading Comprehension Example:
(answer the question, document, question, answer)

Previous work showed we can optimize the unsupervised objective to converge

WebText

- A high quality web scrape
- From outbound reddit links which more than 3 karma points
- 45 million links from which the content extractred
- Links dated before Dec 2017
- 8M documents, 40 GB of text

Encoding

- Byte level LM’s disadvantageous
- BPE (byte pair encoding) were used as a middle ground (Unicode code points)

Model Modifications

- Layer Normalization (Ba et al., 2016) moved to the input of each sub block
- Additional layer norm added after the final self attention block
- Modified initialization (accounting for the accumulation of res. Path with model depth) was used. Weights of the res layer factored by 1/sqr(N) where N = # res layers
- Vocab 50K
- Context size 1024 (from 512)
- Batch size 512

Model

- Smallest model = GPT
- Second smallest = largest BERT
- GPT2, order of magnitude more parameters than GPT

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

Summarizer

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

Overlap (train test)

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

Performance on Multiple Tasks

Language Models are Unsupervised Multitask Learners

	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 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Performance Summary

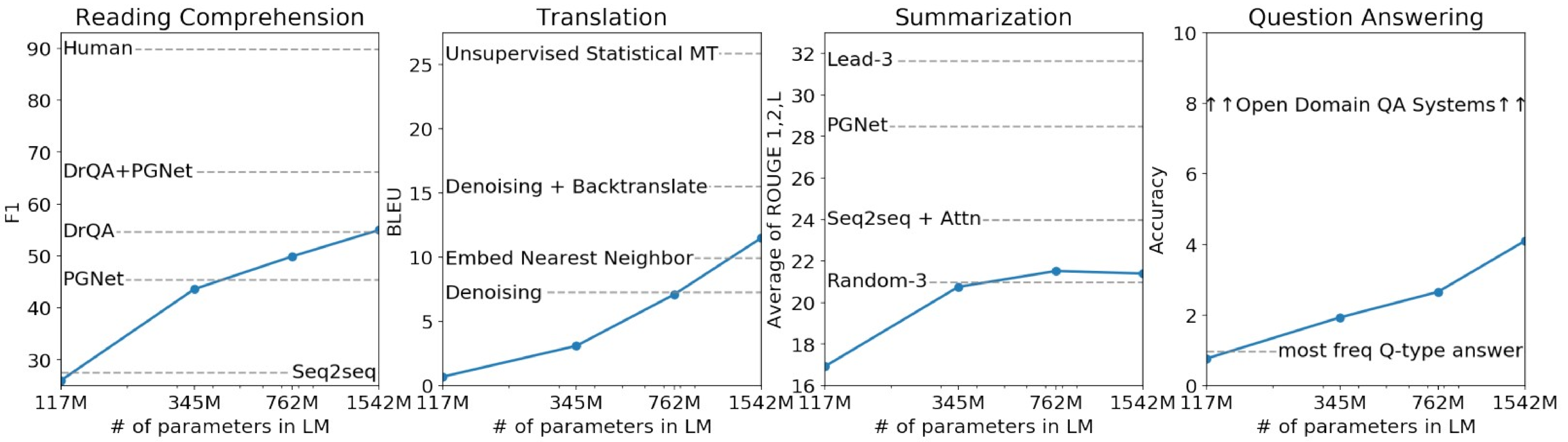

Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

Discussion points

- Is it in fact unsupervised or zero-shot?
- How do you think it compares with:
  - Bert
  - Microsoft’s Multi-Task Deep Neural Networks for Natural Language Understanding?
  - The reason behind certain cut-offs 8-grams
  - The reason behind certain modifications as compared with GPT
  - What are advantages and disadvantages of this approach over training specific network per challenge