Language Models are Unsupervised Multitask Learners (GPT-2)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019).

2019. 08. 10.

김영 (0kimy@yonsei.ac.kr)

1. Introduction

- 데이터 분포의 변화에 강건한, 좀 더 일반적인 기능을 할 수 있는 시스템을 구축하고자 하는 것이 GPT-2 개발의 기본적인 목적
  : 특정한 태스크와 제한적인 영역의 데이터셋에 특화된 시스템보다는 여러 가지 작업을 고르게 잘 수행할 수 있는 시스템이 필요하다!
- Language modeling을 이용하여 zero-shot setting으로도 자연어 처리의 여러 가지 과제를 해결할 수 있음을 보이고자 함

2. Approach

Language Modeling

- 'Unsupervised distribution estimation'으로 볼 수 있음
- 특정 시퀀스의 출현 확률을 각 심볼이 출현하는 조건부 확률들의 곱으로 factorize하여 나타냄
  $p(x) = \prod_{i=1}^{n} p(s_n|s_1, ..., s_{n-1})$

Task Conditioning

- General systems은 다양한 작업을 수행하기 때문에
  $p(output|input, task)$와 같은 형태로 모델링되어야 함
  : task가 output의 조건으로서 주어져야 함
- McCann et al. (2018). "The Natural Language Decathlon: Multitask Learning as Question Answering" 연구에서는 task의 내용을 input 앞에 언급함으로써 이 문제를 해결

2.1. Training Dataset

- 다양한 도메인과 방대한 크기의 텍스트로 모델을 훈련하고자 함
- 이를 위해 Common Crawl과 같은 web scrape 언어 자원을 사용하고자 하였음
    - Trinh & Le (2018)의 연구와 저자의 초기 실험 결과에 따르면 Common Crawl은 데이터 퀄리티에 문제가 있는 것으로 판명
- GPT-2는 자체 제작 web scrape 리소스인 'WebText'를 사용함
    - Reddit 문서 중 최소 3 karma를 받은 문서에 한하여 수집
    - Reddit links (약 450만 개) + Dragnet + Newspaper를 결합하여 40GB의 데이터셋 완성
    - Wikipedia 문서는 데이터셋에서 제외되었음

2.2. Input Representation

Byte Pair Encoding

- BPE는 character-level과 word-level language modeling 간의 절충이 될 수 있음
- GPT-2에서는 UTF-8 byte-level에서 인코딩을 하되 일반적인 단어가 특수 문자와 결합하여 vocab을 확장하는 경우를 제한하기 위하여 문자의 종류를 분류하였음
- Byte-level 인코딩은 유니코드 텍스트의 전처리와 토큰화, vocab size 등에 구애를 받지 않는다는 강점이 있음

2.3. Model

- 모델 구조는 GPT-1과 대동소이 (Transformer-based)
- 차이점
    i. Layer normalization
        - LN이 각 sub-block의 앞단으로 이동
        - 마지막 self-attention block 다음에 LN이 추가됨
    ii. A modified initialization
        - residual layers의 weight를 $1/\sqrt{N}$로 설정
          (N은 residual layers의 총 개수)
    iii. Vocabulary size($\rightarrow$ 50,257), context size(512 $\rightarrow$ 1024), batch size($\rightarrow$ 512)의 확장

3. Experiments

3.1. Language Modeling

	Parameters	Layers	d_{model}
GPT-1	117M	12	768
BERT ≈	345M	24	1024
	762M	36	1280
GPT-2	1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

	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).

3.2. Children's Book Test

- Hill et al. (2015) "The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations"
- 개체명, 명사, 동사, 전치사 등 여러 카테고리의 단어에 대한 Language Model의 판별력을 측정
- 10개의 선택지 중 하나를 골라 빈 칸을 채우는 문제

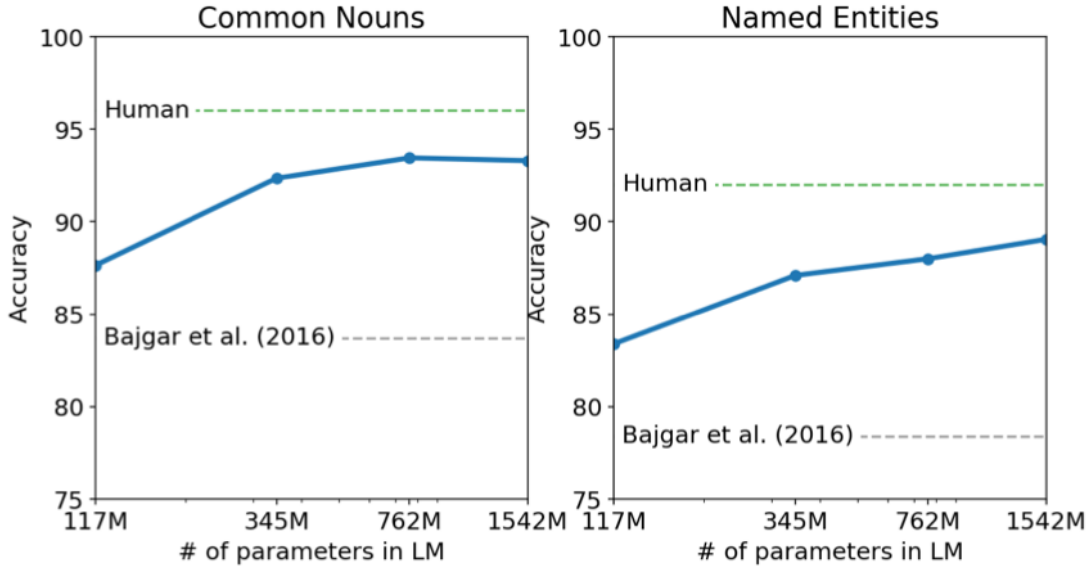

Figure 2. Performance on the Children's Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

3.3. LAMBADA

- Paperno et al. (2016). "The LAMBADA Dataset: Word Prediction Requiring a Broad Discourse Context"
- 최소 50개 가량의 이전 토큰을 보아야 정답을 맞출 수 있는 과제
- Perplexity 99.8 (Grave et al., 2016) $\rightarrow$ 8.6
  Accuracy 19% (Dehghani et al., 2018) $\rightarrow$ 52.66%
  의 성능을 보이며 SOTA 기록

3.4. Winograd Schema Challenge

- 텍스트의 ambiguity resolution을 바탕으로 하는 commonsense reasoning 과제
- Partial scoring과 full scoring으로(Trinh & Le, 2018)으로 계산했을 때 partial score가 더 나음
- Accuracy 70.70%를 기록, 기존의 SOTA를 7% 개선

3.5. Reading Comprehension

- CoQA 데이터셋을 대상으로 측정: 7개 도메인에 대한 글과 이에 대한 질의 응답 대화 쌍으로 구성
- SOTA는 BERT-based supervised model(Devlin et al., 2018)로, 사람이 기록한 F1-score 89에 근접
- GPT-2는 F1 55로 SOTA를 넘지는 못했으나, 4개 baseline 모델 중 3개와 유사하거나 더 나은 성능을 보임

3.6. Summarization

- CNN & Daily Mail 데이터셋을 대상으로 요약 성능 측정

	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)

3.7. Translation

- 시스템에 task를 알려주기 위하여 번역 대상이 되는 문장과 번역된 문장 사이에 심볼 ' = '를 삽입하여 구별
- WMT14 English $\rightarrow$ French 테스트 셋에서 5 BLEU 기록
- WMT14 French $\rightarrow$ English 테스트 셋에서는 11.5 BLEU 기록
- 당시 unsupervised MT(Artetxe et al., 2019)의 SOTA는 33.5 BLEU로 GPT-2의 번역 성능은 좋지 못한 편
  : WebText에서 미처 제거되지 않은 10MB의 프랑스어 데이터가 그나마 영향을 준 것으로 분석됨 $\rightarrow$ French Corpus를 보여줬을 때 결과가 달라질 가능성이 있음

3.8. Question Answering

- exact match metric 기준 정확도는 6% 이하이지만, calibration이 잘 되는 것으로 평가: 모델의 확신이 높은 상위 1%의 답에 한해서는 63.1%의 정확도 기록

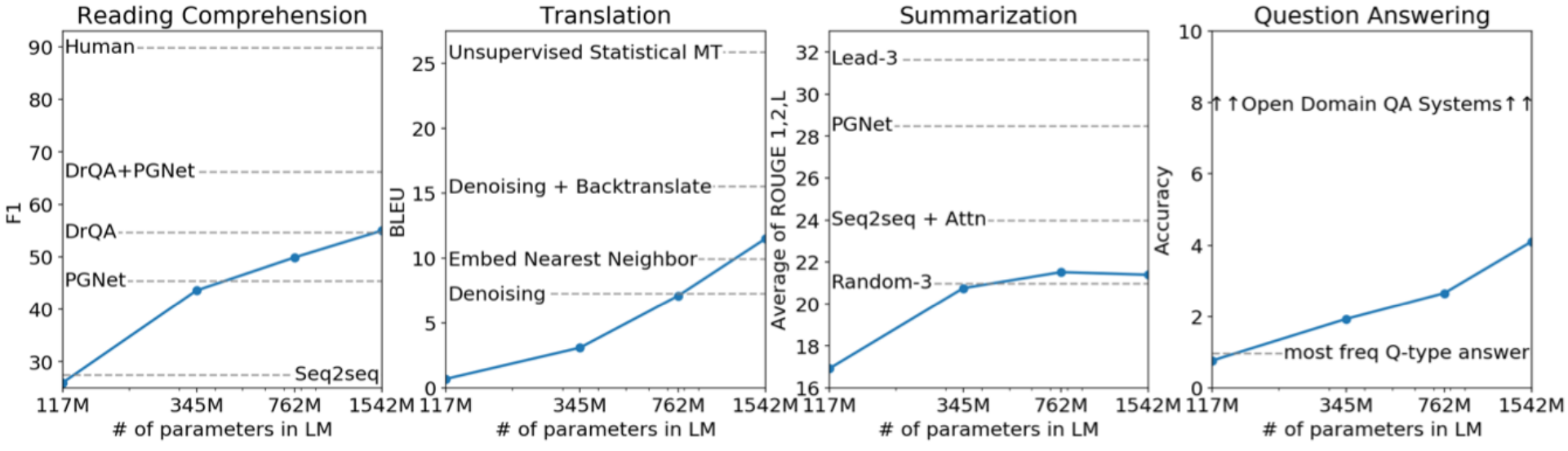

Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural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

4. Generalization & Memorization

- 모델의 generalization 성능을 평가하는 데에 있어 데이터셋의 overlapping 정도를 체크하는 것이 중요한 이슈가 되었음
  (예) CIFAR-10은 train-test images 간에 3.3%의 overlap이 있었던 것으로 밝혀짐
- 그래서 WebText 트레이닝 셋의 토큰을 대상으로 8-gram 단위의 Bloom filters를 만들어 데이터셋의 overlap 정도를 규명하였음

	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.

5. Discussion and Conclusion

- GPT-2는 zero-shot setting에서도 다양한 과제를 준수한 성능으로 해결할 수 있었음
  : unsupervised model의 가능성을 보여줬다는 점이 GPT-2의 의의
- 그러나 summarization task 등 특정 과제에서는 만족할 만한 성능을 보여주지 못했음
- GPT-1처럼 fine-tuning을 적용했을 때 성능이 얼마나 더 향상될 수 있을지를 후속으로 연구하고자 함
- BERT에서 지적했던 uni-directional representations의 단점도 보완되어야 함