Language Models are Unsupervised Multitask Learners

이번 게시물에서는 GPT-2를 제안한 Language Models are Unsupervised Multitask Learners 논문에 대해 리뷰해보려고 한다.

원문 링크는 다음과 같다.

https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Introduction

이 논문이 작성될 시기의 machine learning system들은 data의 분포와 model이 수행해야 하는 task의 변화에 대해 매우 민감했었다. 또한, 그 당시 machine learning system은 모든 task에서 general하게 좋은 성능을 보여주는 방향이 아닌, 수행해야하는 특정 task에 대해 특화되어있는 모습을 보였다.(원문에서는 narrow expert rather than competent generalist 라고 표현하였다.)

저자들은 이러한 system에서 탈피하여 각각의 task에 대한 train dataset을 만들고 labeling할 필요가 없는, 많은 task를 수행할 수 있는 general system을 만들고자 하였다.

그 당시의 machine learning system을 구축하는 주요한 접근법은 원하는 task에 대한 correct behavior을 담고 있는 training dataset을 수집하고, 해당 dataset 안에 포함되어있는 task에 대한 correct behavior를 모방하도록 훈련시킨 다음, independent and identically distributed(IID)한 test dataset에서 train된 machine learning system의 성능을 테스트하는 것이었다.

그러나, 가능한 input의 개수가 많은 task에 대한 model들에서의 불규칙한 model의 행동은 기존의 접근법의 한계를 수면 위로 드러나게 하였다. (즉, 가능한 input의 개수가 많다보니 model의 generalization이 어려워지는 것이다.)

저자들은 단일 domain에서, 단일 task에 대한 train을 하는 추세가 현재 machine learning system에서의 부족한 generalization을 불러일으키는 원인이라고 추측하였고, 현재의 architecture에서 generalization이 풍부한, 보다 robust한 system을 구축하기 위해서는 다양한 domain과 task에서 training과 evaluation을 진행해야 한다고 주장한다. 그러면서, 이와 관련된 연구인 GLUE benchmark와 decaNLP를 언급한다.

(GLUE benchmark와 decaNLP의 경우 아래의 게시물에서 다루었다.)

[논문 리뷰] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding - GLUE

[논문 리뷰] The Natural Language Decathlon:Multitask Learning as Question Answering - DecaNLP를 중심으로

이어서, general performance를 향상시키는 방법으로 다양한 domain과 task를 포함하는 Multitask learning을 소개하지만, NLP에서의 mutitask learning은 아직 연구가 많이 진행되지 않은 초기 단계라고 언급한다. 또한, 현재의 machine learning system은 generalization을 갖추기 위해서 많은 양의 dataset이 필요하기 때문에 multitask learning을 위한 많은 양의 dataset의 생성과 그에 맞는 목적 함수의 설계는 어렵기 때문에, multitask learning을 위한 추가적인 setup이 필요하다고 말한다.

그 당시 가장 좋은 성능을 내는 model은 pre-training과 fine-tuning을 같이 활용하는 model들이었는데, 논문에서는 이러한 방법론도 task를 수행하기 위해서는 여전히 supervised training이 필요하다고 지적한다. 그러면서 다른 방향의 연구들을 언급하는데, 해당 연구들은 supervised data가 아예 없거나 극소수일 때 특정한 task들을 수행하기 위한 language model에 관련된 연구이다.

이러한 흐름 속에, 논문에서는 parameter나 architecture의 수정 없이 바로 downstream task를 수행하는 zero-shot learning이 가능한 language model인 GPT-2를 제안한다.

Approach

GPT-2의 핵심은 GPT1과 마찬가지로 language modeling이다. Language modeling에 대해 간략하게 소개해보겠다.

$$\hat{\theta} = \arg\max \sum_{i=1}^{N} \log P(x^i_{1:n}; \theta)$$
$$\text{where } x_{1:n} = \{x_1, ..., x_n\}$$

$x_{1:n} = \{x_1, \ldots, x_n\}$은 문장 x는 n개의 단어로 구성된다는 의미이다.

$\sum_{i=1}^{N} \log P(x^i_{1:n}; \theta)$는 log-likelihood로, ground truth로부터 나온 sample(dataset)을 분포 θ가 얼마나 설명하는지 나타낸다. 즉, sample들을 분포가 얼마나 잘 설명하는지를 나타낸다. 여기서의 sample은 문장으로, N개의 문장을 모아놓은 dataset에 포함된 문장이다.

따라서, 이러한 sample을 분포가 얼마나 잘 설명하는지 나타내는 log-likelihood를 최대화하기 위해 $\arg\max$를 사용한다.

$$\begin{aligned} & P(x_{1:n}) \\ = & P(x_1, \ldots, x_n) \\ = & P(x_n|x_1, \ldots, x_{n-1}) \ldots P(x_2|x_1)P(x_1) \\ = & \prod_{i=1}^{n} P(x_i|x_{<i}) \end{aligned}$$

이와 관련된 자세한 내용은 아래의 게시물에서 다뤘었으니 참고 바란다.

[NLP] 언어 모델(Language Model) 이란?

논문에서는 최근 몇년동안 transformer와 같은 architecture의 등장으로 이러한 language model의 조건부확률을 계산하는 model의 표현력이 많이 향상되었다고 언급한다.

이후, 저자들은 single task를 수행할 때의 probabilistic framework는 위와 같이 $P(output|input)$으로 표현될 수 있지만, 다양한 task를 수행해야하는 general system의 경우 input은 같지만, input뿐만 아니라 수행되어지는 task도 같이 condition 되어야 한다고 언급한다. 즉, $P(output|input, task)$와 같이 input 뿐만 아니라 task도 주어진 상태에서 output의 probability를 계산해야 하는 것이다. 이러한 task conditioning은 architectural level, 혹은 algorithmic level에서 구현되어왔지만, 저자들은 이러한 방식 이외에 decaNLP에서 제안된 방법처럼 **언어가 task, input, output을 sequence of symbol로 다양하게 제공할 수 있음을 주목한다. (decaNLP에서 제안된 방법은 다양한 task에 대한 input을 (Question, Context, Answer)의 triplet으로 변환하여 여러 task를 question answering task로 다루는 것이었다.** 자세한 내용은 아래의 링크를 참고 바란다.)

[논문 리뷰] The Natural Language Decathlon:Multitask Learning as Question Answering - DecaNLP를 중심으로

예를 들어, 영어-프랑스어 번역 task에 대해서는 (translate to french, english text, french text)의 sequence로 표현할 수 있고, reading comprehension task에서는 (answer the question, document, question, answer)의 sequence로 표현할 수 있는 것이다.

저자들은 decaNLP가 제안된 연구에서 이러한 형태의 example을 통해 MQAN model에서 single model로 multitask learning이 가능했던 것에 대해 주목한다.

기본적으로, language modeling은 별도의 supervision 없이도 이러한 decaNLP의 task들을 학습할 수 있다. Supervised objective의 경우 sequence의 subset에서 evaluation되는 것을 제외하고는 unsupervised objective와 동일하기 때문에, **unsupervised objective의 전역 최솟값과 supervised objective이 같다. 단, unsupervised objective의 convergence가 문제가 되곤 한다.**

저자들의 예비 연구에서, 위와 같은 unsupervised approach setting의 충분히 큰 language model은 multitask learning을 수행할 수 있음을 확인하였지만 supervised approach보다 학습 속도가 상당히 느렸다고 한다.

Training dataset

논문에서는 가능한 다양한 domain과 context에서의 language demonstration을 수집하기 위해 가능한 크고 다양한 dataset을 구축하고자 하였다.

그 당시, Common Crawl과 같은 web scraping은 다양하고 거의 무한하다는 특성덕분에 주목받고 있었던 text source이고, 현재의 language model을 modeling할 때 사용하는 dataset의 크기보다 훨씬 크다는 특징이 있었지만 data quality issue도 존재하였다. 실제로 Common Crawl을 사용했던 연구들에서도 이러한 data의 신뢰성 부분을 지적하였으며, 저자들의 초기 연구에서도 이와 비슷한 data issue가 관찰되었다고 한다.

이러한 문제를 해결하기 위해 한 연구에서는 Common Crawl dataset에서 target dataset과 비슷한 document만을 포함한 Common Crawl의 small subsample을 활용하는 방안을 제시하였지만, 저자들은 이 방법은 실용적이나, 본 연구에서는 task가 실행되기 전에는 task에 대한 가정을 제거하고 싶었다고 말한다.

따라서, 본 연구에서는 **reddit에서 karma 3개 이상을 받은 모든 outbound link를 scraping하였다고 한다. 이러한 방식은 사용자가 해당 link를 유익하게 생각하는가에 대한 heuristic indicator가 될 수 있다고 언급한다.**

이렇게 만들어진 dataset을 WebText라고 명명하였으며, 총 4천 5백만개의 link에 관한 text를 담고있다고 한다. 또한 2017년 12월 이후의 link들은 포함하지 않았으며, de-duplication과 heuristic based cleaning 과정을 거친 이후에는 8백만개의 document, 40GB의 text로 구성되었음을 밝힌다.

또한, Wikipedia document의 경우 다른 dataset에서도 많이 보이는 data source이고, 이는 data overlapping 문제를 야기할 수 있기에 WebText에서 Wikipedia document는 제외하였다.

Input representation

당시 language modeling을 위해서는 lower-casting, tokenization, oov에 대한 대처 등, preprocessing 단계를 거쳐야만 했다.

논문에서는 이러한 preprocessing을 충족시키는 방법으로 unicode 문자열을 UTF-8로 변환하여 byte-level에서의 처리를 제시한다. 다만, 저자들은 byte-level language model이 word-level language model보다 large scale의 dataset에서 성능이 뒤떨어진다는 문제점을 주목한다.

이러한 문제를 해결하기 위해, 논문에서는 BPE 알고리즘을 byte-level에서 동작시키는 Byte-level BPE(BBPE)를 제안한다. 이러한 **BBPE는 기존 BPE는 unicode 문자열이기 때문에 매우 큰 base vocabulary size를 가지는 것에 비해, 기본 vocabulary의 size를 256으로 줄일 수 있다는 장점이 있다.**

그러나 이렇게 byte-sequence에 BPE를 바로 적용하게 되면 BPE의 greedy한 특성으로 인해 sub-optimal한 merge를 야기할 수 있다는 단점이 발생한다. 저자들은 이러한 문제를 **character category가 다르면 merge하지 않게끔 하여 해결**했다고 한다.

이러한 BBPE를 이용한 input representation은 word-level language model의 경험적 이점과 byte-level approach의 일반성을 결합할 수 있게 하는 장점이 있다. **해당 approach가 어떠한 unicode 문자열을 대상으로도 language model로 하여금 확률을 계산할 수 있게끔 하기 때문에 preprocessing, tokenization, vocab size와 관련 없이 어떠한 dataset에서도 evaluate할 수 있게 하는 장점이 존재하는 것이다.**

Model

논문에서는 GPT-2의 구조는 몇 가지의 수정사항을 제외하고는 GPT1의 구조를 따랐다고 언급한다. GPT1에 관련한 내용은 아래에 링크로 남겨두도록 하겠다.

[논문 리뷰] Improving Language Understanding by Generative Pre-Training - GPT

GPT-2에서의 수정 사항을 하나씩 살펴보자.

먼저 layer normalization이 pre-activation residual network와 비슷하게 각 sub-block의 input 부분으로 이동하였다. 또한, 추가적인 layer normalization이 마지막 self-attention block에 적용되었다.

이어서 model의 깊이에 따라 residual path의 누적에 관한 initialization이 변경되었다. Residual layer의 weight에 $1/\sqrt{N}$을 곱해주면서 scaling한다.(N은 residual layer의 개수이다.)

마지막으로, GPT1에 비해 vocabulary size가 50,257로 증가하엿으며 context vector의 크기도 512에서 1024로, batch는 512로 증가하였다.

Experiments

저자들은 다음과 같은 4개의 model을 생성하여 실험을 진행하였다.

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.

위에서부터 아래 순서대로 model의 크기가 커지며, 가장 작은 model은 GPT1의 크기와 동일하고, 2번째로 큰 model은 BERT-LARGE와 같은 크기이다. 저자들은 해당 model들 중에서 가장 큰 model을 GPT-2라고 소개한다.

Learning rate의 경우 Webtext의 5%에 해당하는 held-out sample을 통해 수동으로 조정하였으며, 4개의 model 모두 WebText에 underfitting되었기 때문에 더 많은 시간을 training에 투자하면 더 좋은 성능이 나올 것이라 주장한다.

Language Modeling

저자들은 model을 zero-shot task에 적용하는 초기 단계로 WebText language model이 zero-shot domain의 language modeling에서 어떻게 작동하는지에 대해 관심을 가졌다.

input representation 부분에서 언급했듯이, 본 연구의 model은 byte-level에서 동작하기 때문에 어떠한 language model benchmark를 적용할 수 있었고, language modeling dataset의 경우 표준 예측 단위(character, byte, word 등 token의 기준을 의미한다.)당 scaled NLL loss 혹은 exponentiated nll loss를 이용하여 evaluate하기에 WebText LM에도 이러한 방식을 적용하여 evaluate하였다.

이와 관련된 결과는 다음과 같다.

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).

GPT-2가 zero-shot setting에서 8개의 dataset중 7개에서 SOTA를 달성함을 확인할 수 있다. 특히, PTB나 WikiText-2와 같은 **small dataset에서 큰 성능 향상**이 있음을 확인할 수 있다. LAMBADA나 CBT와 같이 LM의 **long-term dependency를 측정하는 dataset에서도 비약적인 성능 향상**을 이루어 냈다.

그러나 1BW dataset에서는 오히려 성능이 하락했음을 확인할 수 있는데, 저자들은 이에 대해서 1BW는 가장 큰 dataset이기도 하고, 1BW의 destructive pre-processing(sentence level shuffling)이 long range structure를 제거하기 때문에 발생하는 현상이라고 추측한다.

Children's Book Test

Children's Book Test(CBT) dataset은 named entities, nouns, verb와 같이 품사에 따른 LM의 성능을 측정하기 위해 고안된 dataset이다. CBT dataset은 perplexity를 evaluation metric으로 사용하지 않고, 빈칸과 이에 대한 10개의 선택지가 주어질 때 빈칸에 알맞은 항목을 선택하는 cloze test의 accuracy를 evaluation metric로 사용한다.

각각의 선택의 확률과, 이 선택으로 인한 문장의 나머지 부분에 대한 확률을 계산하고 가장 높은 확률의 선택지를 선택하게 한다.

해당 결과는 다음과 같다.

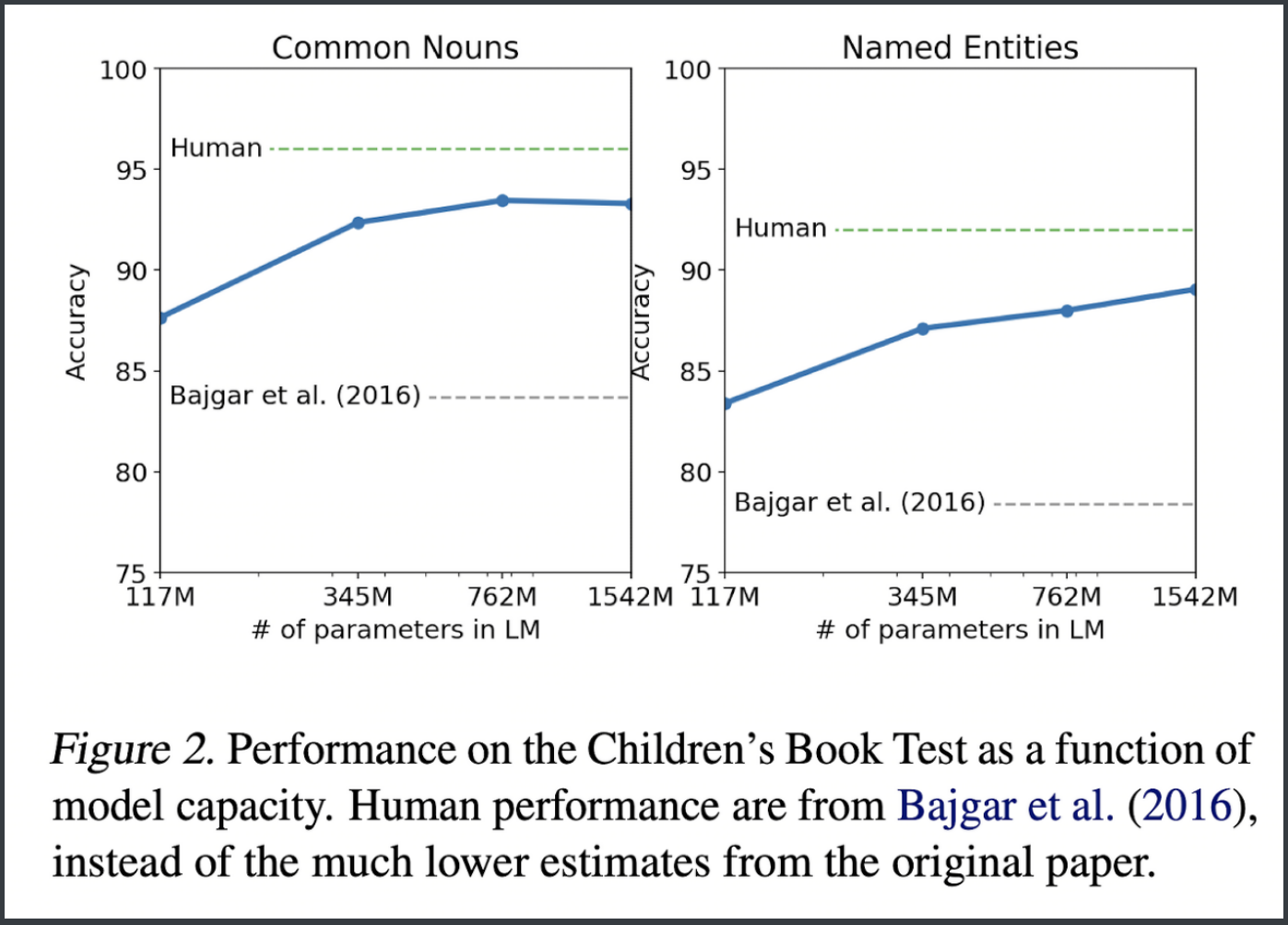

Figure 2. Performance on the Children's Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

위의 표에서도 확인할 수 있었던 것처럼, common noun을 predict할 때에는 93.3%의 accuracy, named entities에서는 89.1%의 SOTA를 달성하였다.

LAMBADA

LAMBADA dataset은 LM의 long-range dependency를 측정하기 위해 고안된 dataset이다. 각각의 단락의 맨 마지막 단어를 예측하는 task이며, GPT-2는 여기서 PPL에 대해 SOTA의 성능을 달성하였다. (PPL : 99.8 -> 8.6, ACC : 19% -> 52.66%)

그런데, 저자들은 GPT-2의 대부분의 prediction이 final word가 아닌 sentence의 valid continuation이라는 것을 알게 되었고, 이것에 대한 대안으로 stop-word filter를 추가하자 accuracy가 63.24%로 상승하게 되었다. 결과적으로 GPT-2는 PPL과 accuracy 모두에서 SOTA를 달성할 수 있었다.

Winograd Schema Challenge

Winograd Schema는 해당 model이 text 속에서 모호함을 얼마나 잘 해결할 수 있는지 측정하는 commonsense reasoning 수행 능력을 측정한다.

저자들의 WebText LM이 해당 task에서 보여준 성능은 다음과 같다.

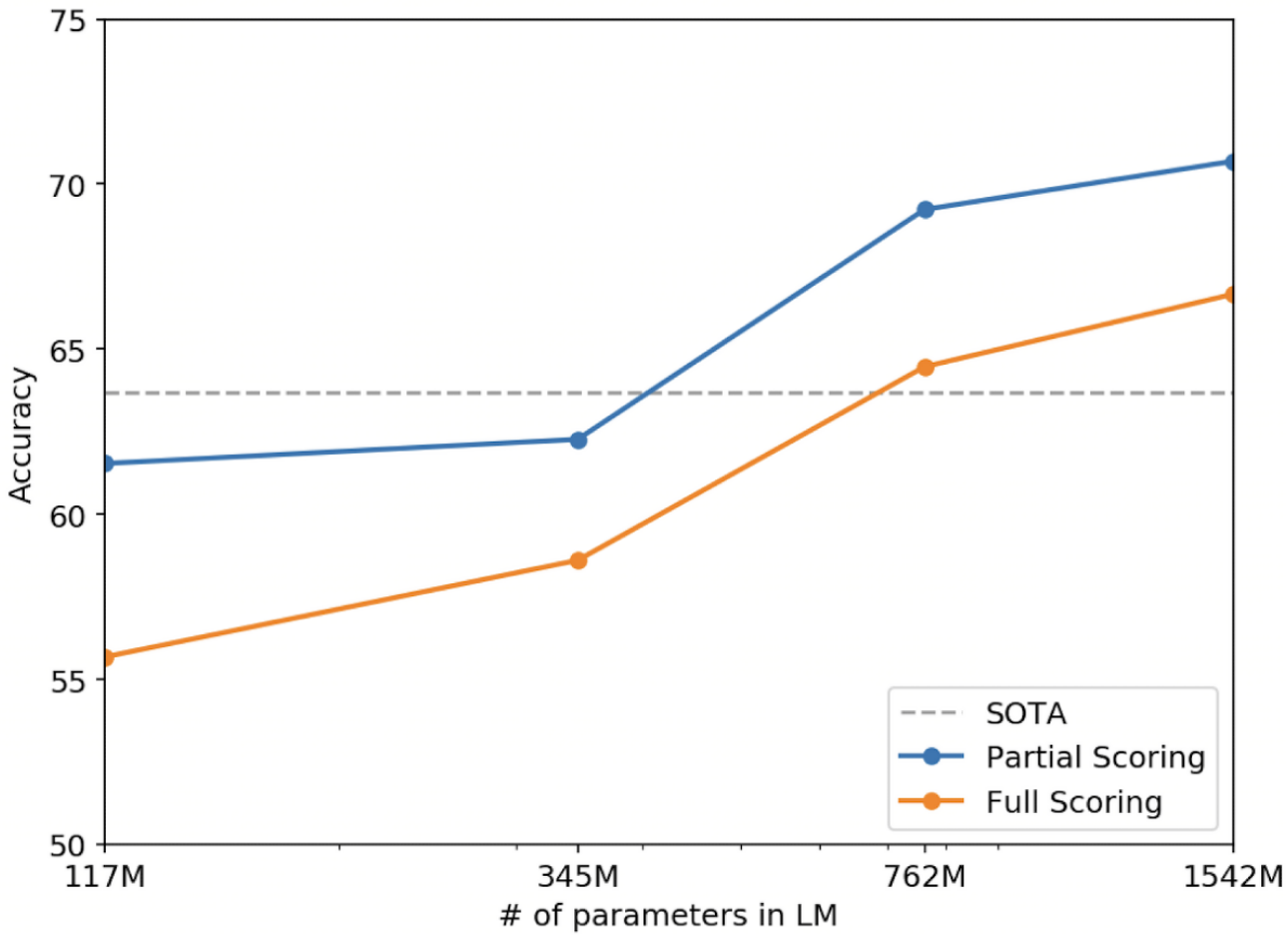

Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

여기서, parameter수가 가장 많은 GPT-2는 기존 SOTA model의 성능에서 7%가 향상된 70.7%의 성능을 보여주었다.

Reading Comprehension

Conversation Question Answering dataset(CoQA)는 7개의 다른 domain으로부터 추출된 document에서 question asker와 question answerer의 대화로 구성된 dataset이다. CoQA는 model의 reading comprehension 능력과 대화 이력에 기반한 답변 능력을 테스트한다.

GPT-2는 55의 F1 score를 달성하였는데, 이는 zero-shot setting으로 얻어진 결과이다. 참고로, fine-tuning을 진행한 BERT의 F1 score가 89였으며, 저자들은 GPT-2의 경우 supervised training(fine tuning)없이 55의 F1 score를 얻어냄을 강조한다.

Summarization

저자들은 GPT-2의 summarization 능력을 평가하기 위해 CNN and Daily Mail dataset을 사용하였다.

결과는 다음과 같다.

	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)

GPT-2가 이러한 summarization task에서는 잘 작동하지 않음을 확인할 수 있다.

Translation

저자들은 GPT-2가 한 언어에서 다른 언어로 번역하는 방법을 학습하기 시작했는지 테스트를 진행한다. WMT-14 English-French test set에서는 5의 BLEU score를 얻었으며, 이는 기존 unsupervised translation 관련 연구보다 낮은 성능이다.

WMT-14 French-English test set에서는, 11.5 BLEU score의 보다 더 좋은 성능을 얻을 수 있었다. 그러나 이 성능 또한 타 model에 비해서는 좋지 않은 성능이다.

Question Answering

GPT-2는 SQuAD와 같은 reading comprehension에서 사용되는 정확하게 일치하는지 아닌지에 대한 metric으로 평가되었을 때 4.1%의 question에 대해 정확하게 대답하였다고 한다. 심지어, WebText LM 중에서 크기가 가장 작은 model의 경우 1%의 정확도도 넘지 못했다고 한다. 저자들은 GPT-2는 5.3배 많은 question에 대해 정확하게 답변하였고, 이는 model의 capacity가 이러한 종류의 task에서 neural system이 좋지 않은 주요한 원인이었음을 시사한다고 주장한다.

GPT-2가 생성된 answer에 대해 할당하는 확률(해당 answer에 대한 확률)은 잘 calibrated 되어있는데, 실제로 GPT-2가 가장 신뢰할 수 있는 답변이라고 한 1%(probability가 높은 1%의 answer)에 대한 accuracy는 63.1%이었다.

아래 표는 해당 answer들의 집합이다.

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

그러나, 아직까지도 GPT-2의 성능은 information retrieval과 extractive document를 혼합한 open domain QA system보다 낮다.

Generalization vs Memorization

그 당시 CV쪽에서 이루어진 연구들 중에서는 dataset에 무시하지 못할 양의 복제된 이미지가 포함되어있다는 것을 밝혀낸 연구들이 있었다. 예를 들어, CIFAR-10 dataset의 경우 train data와 test data 사이에 3.3%의 중복이 있음을 확인되었다.

이러한 현상은 model의 일반화 성능(generalization performance)을 과대해석하게끔 한다.

(즉, 이러한 overlapping이 많을수록 model은 generalization보다는 memorization에 더 치우치게 될 것이고, 이는 다양한 task에 적용할 때 악영향을 미친다.) 또한 dataset의 크기가 커질수록 이 현상은 점점 심해지게 된다.

즉, WebText와 같은 거대한 dataset에서는 이러한 데이터 중복이 얼마나 발생하는지에 대한 분석이 필수적이다.

저자들은 이를 조사하기 위해 8-gram의 Bloom filter를 사용하였다. 해당 Bloom filter는 dataset이 주어졌을 때, 해당 dataset의 8-gram이 WebText의 training set에서도 얼마나 존재하는지를 계산해준다.

이러한 Bloom filter를 통해 얻은 결과는 다음과 같다.

	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.

일반적인 LM dataset들은 WebText training set과는 1~6%, 평균적으로 3.2%의 overlap이 있음을 확인할 수 있었다. 그런데, 각각의 dataset들이 자기 자신의 training set과 test set에서 평균적으로 5.9%의 overlap이 있는 사실 또한 알 수 있었다.

전반적으로, WebText의 training set과 특정 evaluation dataset간의 data overlap은 evaluation result에 작지만 일관된 이점을 가져다준다. (논문에서는 다른 dataset과 WebText간의 data overlap이 performance에 미치는 영향을 조사하였다. 본 게시물에서는 이 부분은 생략하도록 하겠다.) 그러나, WebText의 training dataset은 위에서 살펴본것처럼 각각의 dataset의 training dataset과 test dataset에서 나타나는 overlap보다 더 적은 data overlap을 보여준다.

이어서, 저자들은 WebText LM의 성능이 memorization에 기인하는지 확인하기 위해 WebText의 held-out set에서의 성능을 조사해보았다. 결과는 아래와 같다.

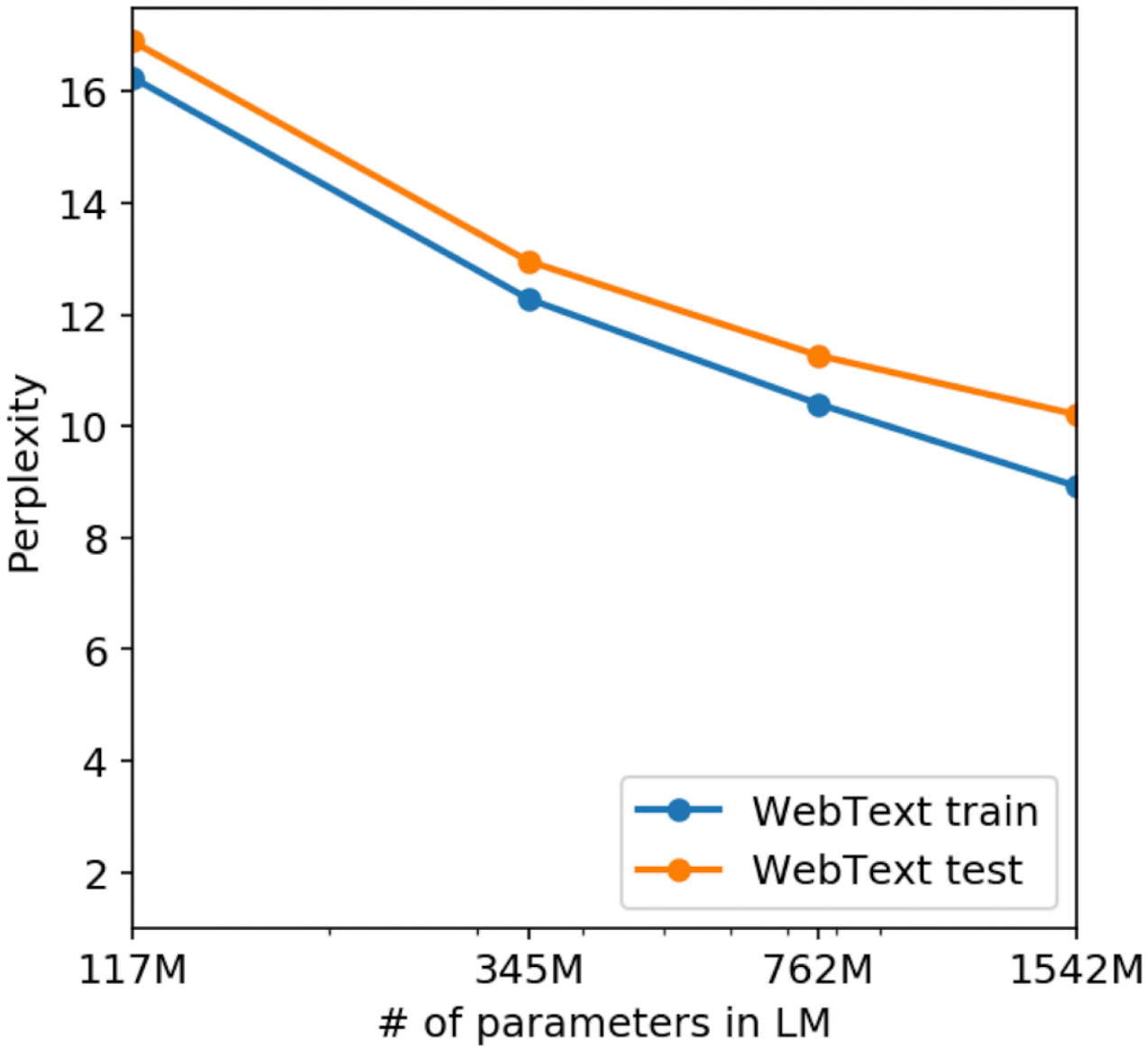

Figure 4. The performance of LMs trained on WebText as a function of model size.

결과를 보면, training set과 test set에서의 결과를 확인할 수 있었는데, train set에서의 성능과 test set에서의 성능이 비슷하며, 모두 model의 size가 올라갈수록 성능도 올라감을 보였다.

저자들은 이러한 결과는 GPT-2가 아직 WebText corpus에 underfitting되었음을 말해주는 결과라고 말한다.

(즉, training에 더 많은 시간을 투자하면 성능이 오를 수 있다는 것이다.)

Conclusion

Large language model이 충분히 크고 다양한 dataset에서 학습된다면 많은 domain과 dataset에서 잘 작동할 수 있다. GPT-2는 zero-shot setting에서 SOTA를 달성하였고,(8개의 dataset 중 7개), 이러한 model이 zero-shot setting에서 잘 작동할 수 있는 task의 다양성은 충분히 다양한 text corpus에서 학습된 high-capacity의 model이 supervision없이도 다양한 task를 수행하는 방법을 학습한다는 것을 시사한다.