Language Models are Unsupervised Multitask Learners

Alec Radford * 1 Jeffrey Wu * 1 Rewon Child 1 David Luan 1 Dario Amodei ** 1 Ilya Sutskever ** 1

핵심 아이디어

1. Fine-tuning 없이 진행
2. Zero-shot learning

→ 범용적인 language model을 만드는 것을 목표로 함

Introduction

- Machine learning system은 대규모 데이터셋, high-capacity model, 지도 학습의 조합으로 주어진 task에 대해 좋은 성능을 보임
- 하지만 이런 시스템들은 data distribution이나 task specification이 조금만 바뀌면 망가지기 쉬움
- 현재의 시스템은 generalist가 아닌 특정 task에 대해서만 특화된 narrow expert임

→ 본 논문에서는 매번 task마다 별도로 데이터를 만들고 labeling하지 않아도 다양한 작업을 수행할 수 있는 **general system**을 지향함

Introduction

- 본 논문에서는 down-stream task에 대해 zero-shot setting에서 진행

  (parameter나 architecture의 변경 없이)

- 일부 task에서는 SOTA를 달성하기도 함 → zero-shot에서의 가능성을 보여줌

Approach

- 가장 핵심적인 것은 **"language modeling"**임

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

- Single task는 conditional distribution $p(output|input)$를 추정
- General system에서는 같은 input이더라도 다른 작업을 수행할 수 있어야 하기 때문에 $p(output|input, task)$로 표현

  ex) "Translate to French: I am happy", "TL;DR: I am happy"

Training dataset

- 이전 연구들은 news article, wikipedia와 같은 single domain text만을 사용함
- 본 논문에서는 다양한 분야와 맥락에서의 task를 위해 크고 다양한 데이터셋이 필요함
- 기존에 Common Crawl이 존재했지만 데이터 품질에 문제가 많음

→ 그래서 직접 **WebText**를 만들어 사용함

WebText

- Reddit에서 3 krama 이상의 데이터만 수집(heuristic indicator로 생각 가능)
- 추후 중복 제거 및 Wikipeida document 삭제(다른 데이터셋에서도 자주 사용되기 때문에 train, test task간의 중복 발생 가능성 염두) → 총 40GB, 800만 개 이상의 문서

Input Representation

- Language model은 모든 문자열에 대해 확률을 계산하고 생성할 수 있어야 함
- 하지만 기존 large scale LM은 lower-casing, tokenization, 사전에 없는 단어는 <unk>로 처리하는 것들로 인해 모든 문자열을 처리하는 제약이 존재함
- 텍스트를 UTF-8 byte로 표현하면 전부 다룰 수 있다는 장점이 존재하지만 world-level LM보다 성능이 낮음

Input Representation

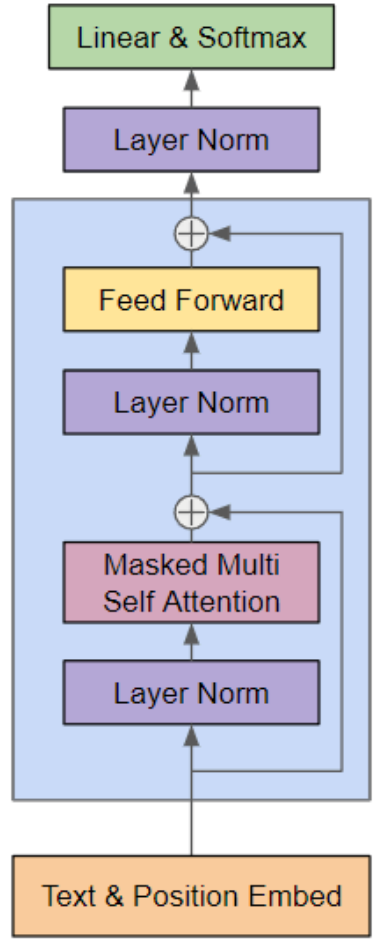
- 본 논문에서는 Byte Pair Encoding(BPE)를 사용하고자 함
  - 기존 BPE는 자주 나타나는 문자 조합을 묶어서 하나의 토큰으로 만드는 방식을 말함.
  - 자주 쓰이는 단어는 world-level, 쓰이지 않는 것은 character-level로 처리가 가능해 효율적
  - World-level 성능을 어느정도 유지를 하며 모든 문자열에 대해 대응이 가능함
- 기존 BPE를 일부 수정하고자 함
  - 기존 BPE는 문자 범주를 구분하지 않고 토큰을 합침

    ex) dog, dog., dog! 등을 각각 다른 토큰으로 만듦 → 중복으로 인한 낭비가 발생함
  - 그래서 같은 문자 범주끼리만 합침(문자+문자, 기호+기호), 하지만 space는 예외로 합침

    ex) [‘dog’], [‘dog’, ‘.’], [‘dog’, ‘!’]으로 분리
- World-level의 장점과 byte-level의 generality를 결합 가능

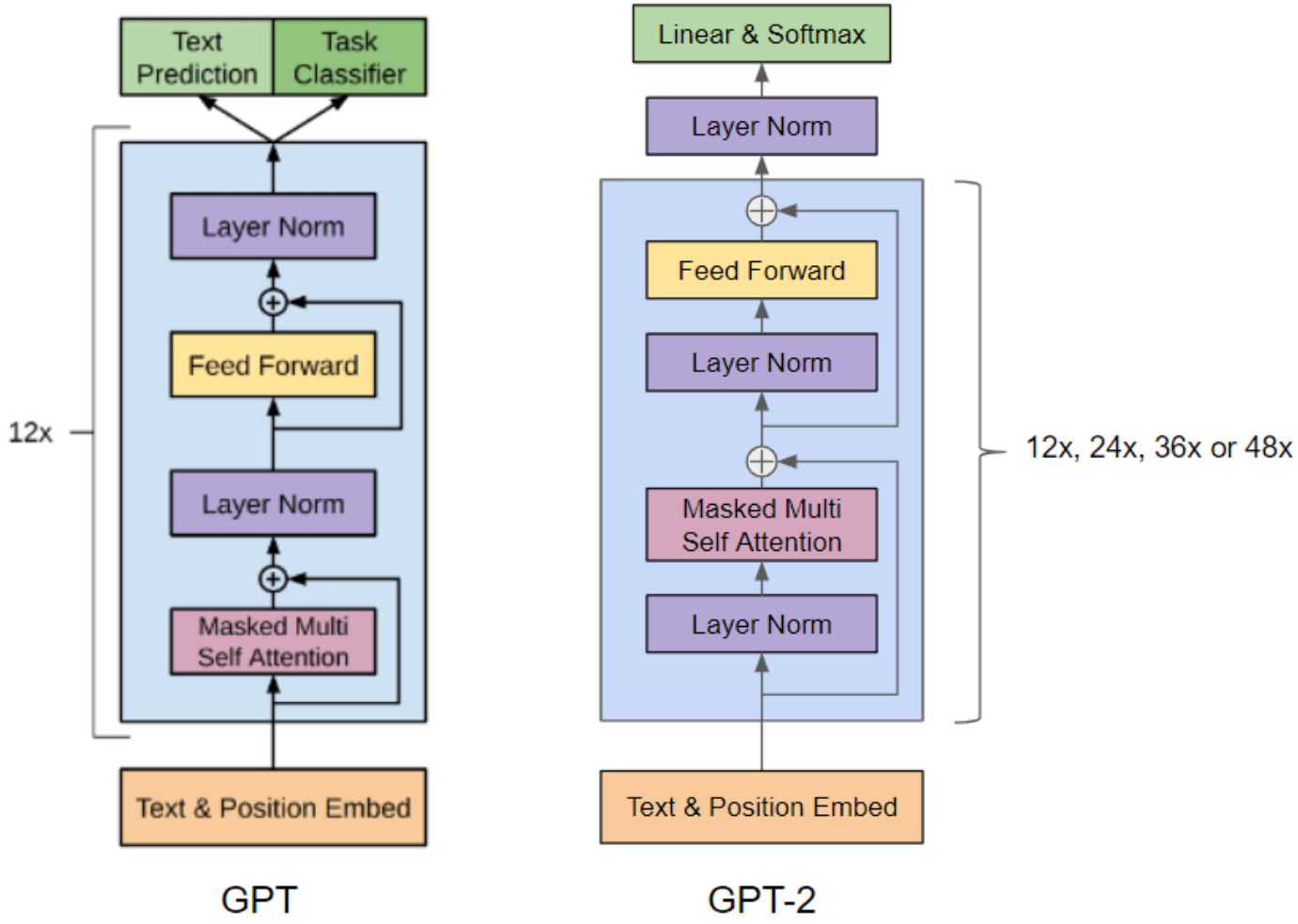
Model

- Transformer decoder 활용
- 대부분 GPT1을 따름(일부 수정 사항 존재)
  - Layer Normalization을 각 sub block의 input으로 옮겨짐(pre-norm 방식으로 학습에 더 안정적. 추후 대부분의 모델에서 이 방식을 채택하여 사용함)
  - 마지막 self-attention block 이후에 layer normalization 추가
  - 모델 깊이에 따른 residual path의 누적 때문에 가중치 초기화 방법 변경 (factor of $1/\sqrt{N}$)
  - 50,257개로 vocab size 확장
  - Context size 512→1024 token으로, batchsize는 512로 키움



GPT-2

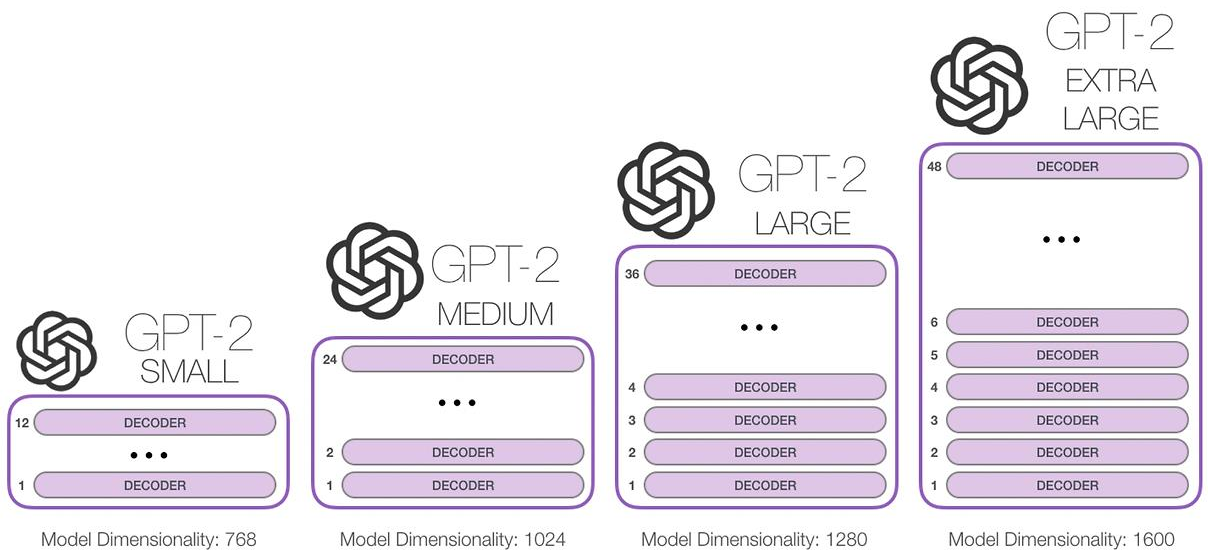
Model



https://velog.io/@tm011899/GPT-2-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0Language-Models-are-Unsupervised-Multitask-Learners%EC%9E%91%EC%84%B1-%EC%98%88%EC%A0%95

Experiments

총 4가지 size의 모델을 구성



Parameters	Layers	d_{model}	
117M	12	768	GPT1
345M	24	1024	BERT Large
762M	36	1280	
1542M	48	1600	

Table 2. Architecture hyperparameters for the 4 model sizes.

Language Modeling

- 다양한 task에 zero-shot learning의 성능을 보여주기 전 기본적인 language modeling에서 다른 도메인의 데이터셋에서도 잘 작동을 하는지 보여주고자 함
  - GPT2는 byte level로 작동을 하기 때문에 pre-processing, tokenization이 필요 없음 → 다른 데이터셋으로 성능 평가 가능함
- 총 8개 중에 7개에서 SOTA를 달성함

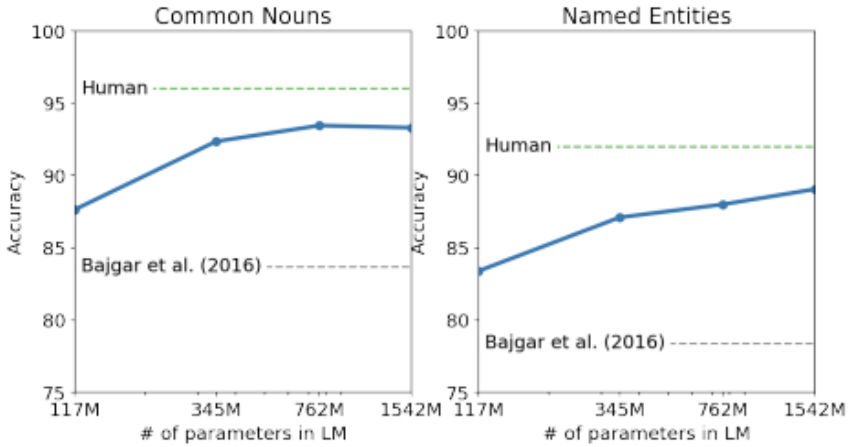
Sentence level shuffling으로 의미적 연결성을 파괴함

	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

Children's Book Test

She turned the corner and saw ____ standing there.
선택지: John, table, running, quickly, door, Peter, happiness, city, girl, house

- 다양한 단어 유형(고유명사, 명사, 동사, 전치사)에 대한 성능 측정
- Cloze test: 10개의 후보 중 빈칸에 들어갈 것을 고르는 방식
- 논문과 동일한 방식(각 후보 단어로 완성된 문장의 확률을 계산해 가장 확률이 높은 것을 선택)
- 모델이 커질수록 성능은 지속적으로 향상되어 인간 성능에 가까워지면 gap을 줄여나감
- 데이터셋 중 일부 WebText에 포함되는 것을 확인해 중복 데이터를 validation set으로 구성



LAMBADA

Laura walked into the old bookstore she used to visit as a child. The smell of aged paper and dust instantly brought back memories. She wandered through the narrow aisles, running her fingers across the spines of books. Then she stopped, smiling as her eyes fell on the one book she had loved the most ____

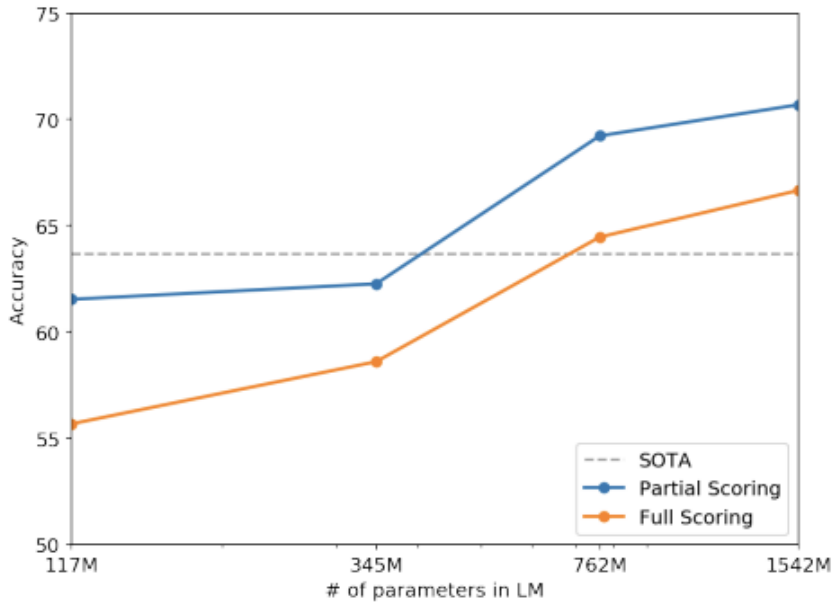
- 긴 문맥(50 단어 이상)을 바탕으로 문장의 마지막 단어를 예측하는 task
- Perplexity: 99.8 → 8.6 / Accuracy: 19% → 52.66%
- 마지막 단어라는 제약을 활용하지 못한 오류 존재 → 마지막에는 stop-word가 등장할 확률이 낮기 때문에 stop-word는 제외함 → 성능이 63.24%로 향상됨

Winograd Schema Challenge

- Commonsense reasoning 능력을 측정
- Full scoring(전체 문장에 대입하고 전체 확률을 계산), Partial scoring(대명사 이후 부분만 확률 계산)
- 기존 SOTA보다 7% 높은 70.70% 달성

문장:
The trophy doesn't fit into the brown suitcase because **it** is too small.

질문:
"it"이 가리키는 것은?
→ the suitcase



Reading Comprehension

- Conversation Question Answering dataset(CoQA)을 활용(7개의 도메인 문서에 대해 question asker, question answerer의 대화 쌍으로 구성)
- 55 F1 score을 보임
- 89 F1 score을 보인 BERT보다는 낮지만 zero-shot 기반 의미 있는 결과값
- 하지만 오류를 확인해보니 “who?”를 묻는 질문에 문서의 이름을 가져오는 등 이해를 하기 보다는 깊은 추론 없이 이전 단어를 복사하는 retrieval based heuristics 방법에 그침

Summarization

- CNN, Daily Mail dataset을 활용
- 요약 작업이라는 진행하기 위해 “TL:DR”을 텍스트에 추가해 유도함
- Top-k random sampling(k=2)으로 100개의 token 생성하고 첫 3개의 문장을 사용

  → 중복 ↓ / 문장 다양성 ↑
- 표면상으로는 그럴듯해 보이지만 세부적인 사항(숫자, 대상)은 오류 자주 발생
- ROUGE 1/2/L 점수는 classic neural baseline에 근접하지만 random하게 3개의 문장을 선택한 것보다 아주 조금 나은 수준 → 좋은 성능을 보이지는 않음
- “TL:DR”을 제거하면 6.4 하락 → 자연어 프롬프트만으로도 작업 유도 가능함을 의미

Translation

- WMT-14 English-French dataset 활용
- “English sentence=french sentence”의 형태로 몇개의 예시를 제공한 후, English sentence= 뒤 문장을 생성하도록 함
- French를 english로 번역하는 것은 반대의 경우보다는 상대적으로 좋은 성능을 보였는데 이는 GPT2가 학습한 언어에 French가 극히 소량만 존재함
- 하지만 이럼에도 불구하고 번역 기능을 수행한 것에 가능성을 보임

Question Answering

- Natural Question dataset 활용
- Question과 answer을 쌍으로 몇개의 예시 제공
- 가장 작은 모델은 1%를 넘지 못하고 가장 큰 모델은 4% 정도 성능이 나옴
- 기존 QA task는 30~50%의 성능으로 이에 많이 못 미치는 수준
- 자신 있는 1% 질문에 대해서는 63.1%의 정확도를 보임

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%

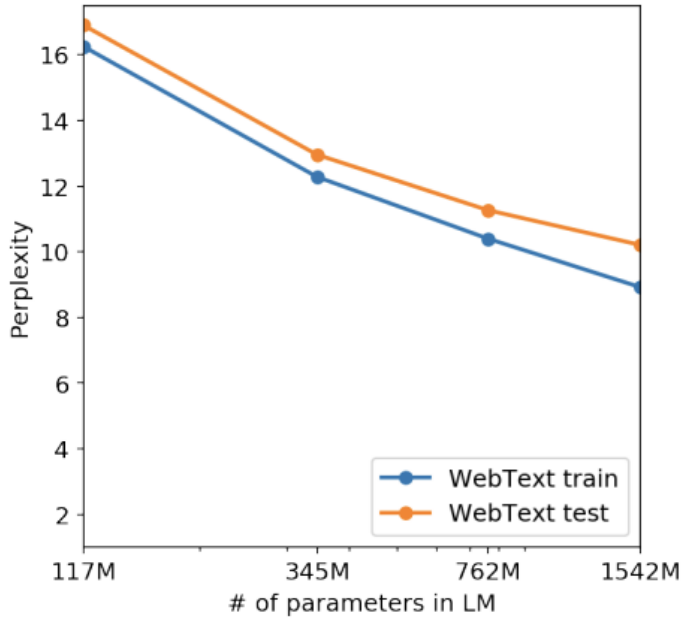
Generalization vs Memorization

- Train dataset과 test dataset 사이 데이터가 겹치는 경우가 존재함
  - 모델이 본 적 있는 문장(test 중복) 을 맞히면, 진짜 일반화 능력이 아니라 memorization일 수 있음
- 본 논문에서는 8-gram을 사용해 bloom filter를 만들어 test함
- 일부 중복이 존재하기는 하지만 대부분 성능 향상에 큰 영향이 주지 않았음
- 조금 향상이 된 부분도 있지만 이는 기존 train/test dataset에서도 이정도 수준의 중복은 존재하기 때문에 신뢰성에 큰 영향 없음

	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%

Generalization vs Memorization

- 추가로 memorization을 하는것인지에 대한 검증을 위해 WebText를 train/test dataset으로 나누어 성능을 평가함
- Model size가 커짐에 따라 train/test의 성능은 계속 좋아지는 것을 통해 memorization을 하지 않으며 underfitting되어 있음을 의미함



Discussion

- Unsupervised learning은 아직 연구할 것이 남아 있음
- Reading comprehension의 경우 견줄만한 성능이 나오지만 summarization은 아직 성능이 사용하기에는 부족함
- Question Answering, translation도 sufficient capacity인 경우 baseline보다 성능이 좋음

→ Zero-shot learning이라는 점에서 혁신적인 방향을 열었지만 모든 task에서 아직 실용적인 성능을 내는 것은 아님

- BERT에 의해 증명된 uni-directional representation의 비효율성을 additional training data와 capacity로 극복할 수 있을지 불분명함

Conclusion

- 크고 다양한 데이터셋으로 학습된 large language model은 다양한 도메인과 데이터에서 좋은 성능을 보임
- 8개의 benchmark 중 7개에서 zero-shot으로 SOTA 달성
- 충분히 크고 범용적인 모델 + 다양하고 풍부한 학습 데이터의 조합으로 일반적인 언어 능력을 학습할 수 있음을 보여줌