Language Models are Unsupervised Multitask Learners

Sunyoung Song

1. Introduction

- 머신러닝 시스템들은 **큰 데이터셋, 큰 용량의 모델, 지도학습을 통해 훌륭한 성과**를 보여줌
- 현재 머신러닝 시스템들은 **‘narrow experts’**
- 일반화가 필요함
- Supervised가 없다면 더 범용적으로 사용할 수 있을 것이라고 생각함
  -> **supervised fine-tuning 없이 적용**하게 됨
- **GPT-1 과의 차이**
  - GPT-1 : unsupervised pre-training 과 supervised fine-tuning
  - GPT-2 : supervised fine-tuning 이 없음

2. Approach

- 원래는 조건부 확률로 sequential하게 단어 예측

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

- **p(output | input)** -> 여러 개의 task에 적용하려면 **p(output | input, task)**

2. Approach

- **Training Dataset**
  - WebText 라는 dataset 직접 구성
    - 40GB의 800만 개가 넘는 문서 data
    - 직접 크롤링
      - Reddit에서 3 karma 이상을 받은 글들만 사용
      - Wikipedia 문서와 중복인 부분은 제거

- **Input Representation**
  - Input 할 때 BPE (Byte Pair Encoding) 을 사용
    - 딕셔너리의 모든 단어를 글자(character) 단위로 분리
    - 가장 빈도수가 높은 unigram 쌍을 하나의 unigram으로 통합
    - 위 작업을 정해진 횟수만큼 반복

2. Approach

- **BPE 동작 원리**

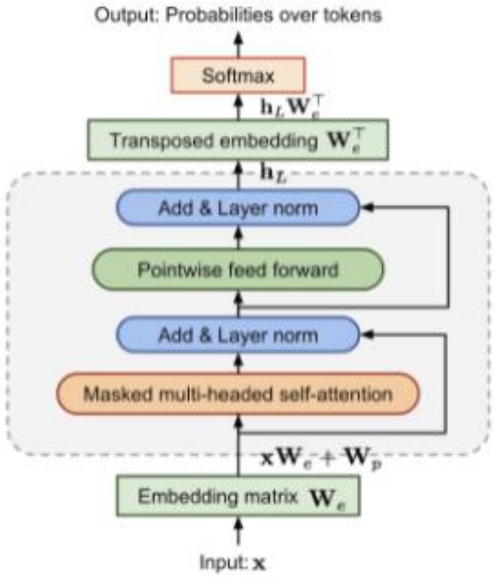
```
#dictionary
h i g h : 5, h i g h e r : 2, n e w e s t : 6, w i d e s t : 3
```

```
#updated dictionary
h i g h : 5, h i g h e r : 2, n e w es t : 6, w i d es t : 3
```

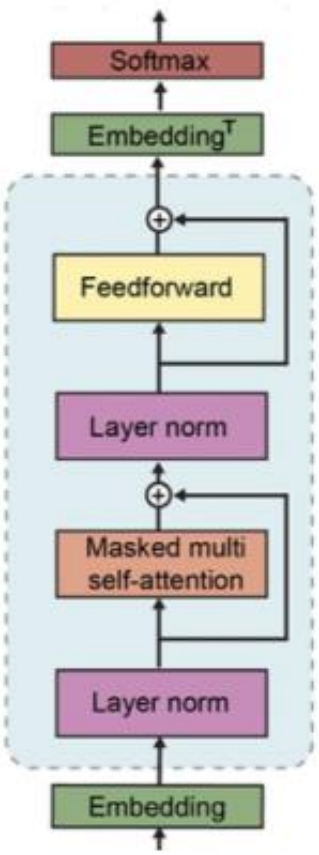
```
#updated dictionary
h i g h : 5, h i g h e r : 2, n e w est : 6, w i d est : 3
```

2. Approach

- **Model**
  - GPT-1 과의 차이점
    - Layer normalization의 위치
    - Residual path의 누적에 관한 부분의 초기화 방법 변경
    - Context size 증가
    - Vocabulary 증가
    - Batch size 증가



GPT-1 decoder block

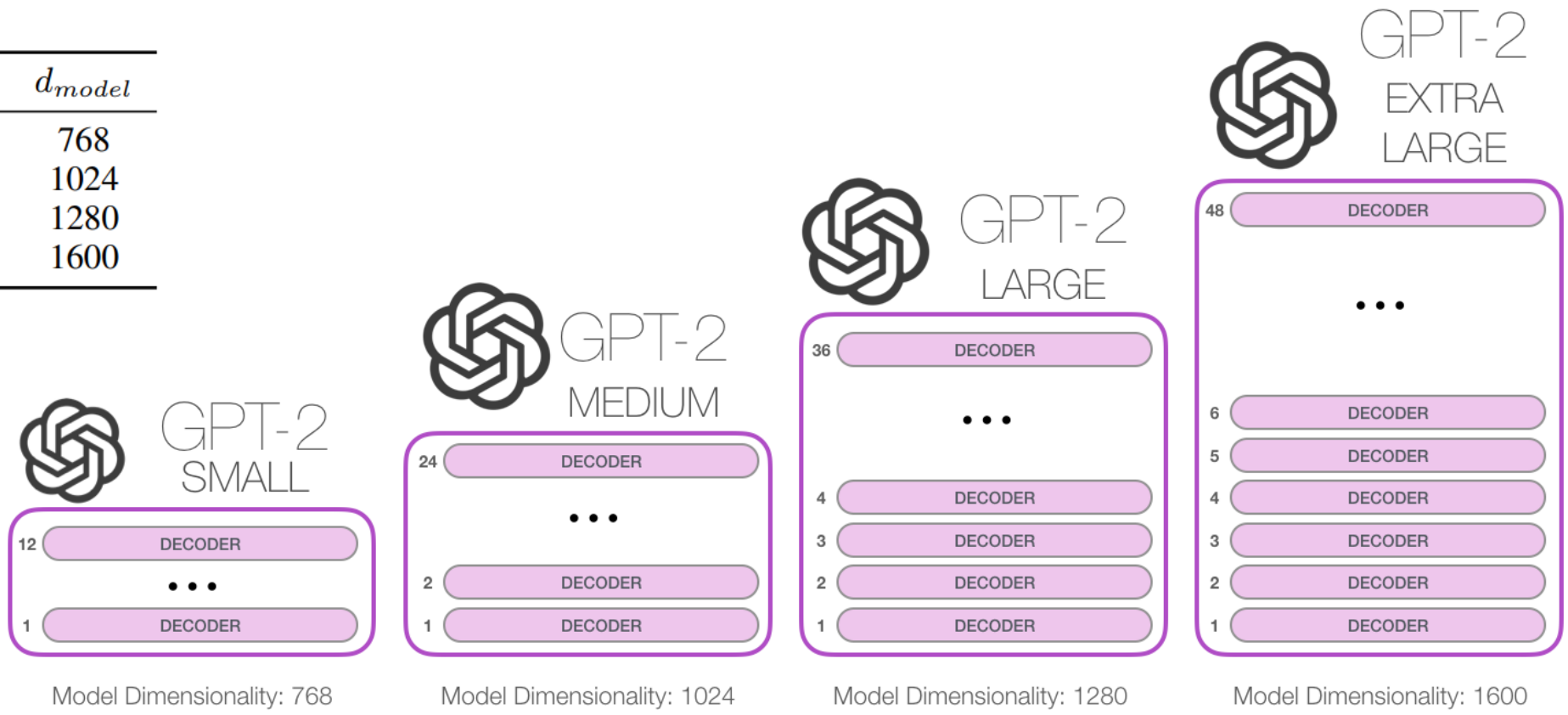


GPT-2 decoder block

2. Approach

- **Model**
  - 대용량 모델

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



3. Experiments

1. **Language Modeling**
   - 다양한 벤치마크 데이터셋에 zero-shot 환경에서 성능 비교를 진행한 결과
   - Fine-tuning을 진행하지 않은 zero-shot 환경임에도 7개에서 SOTA를 달성

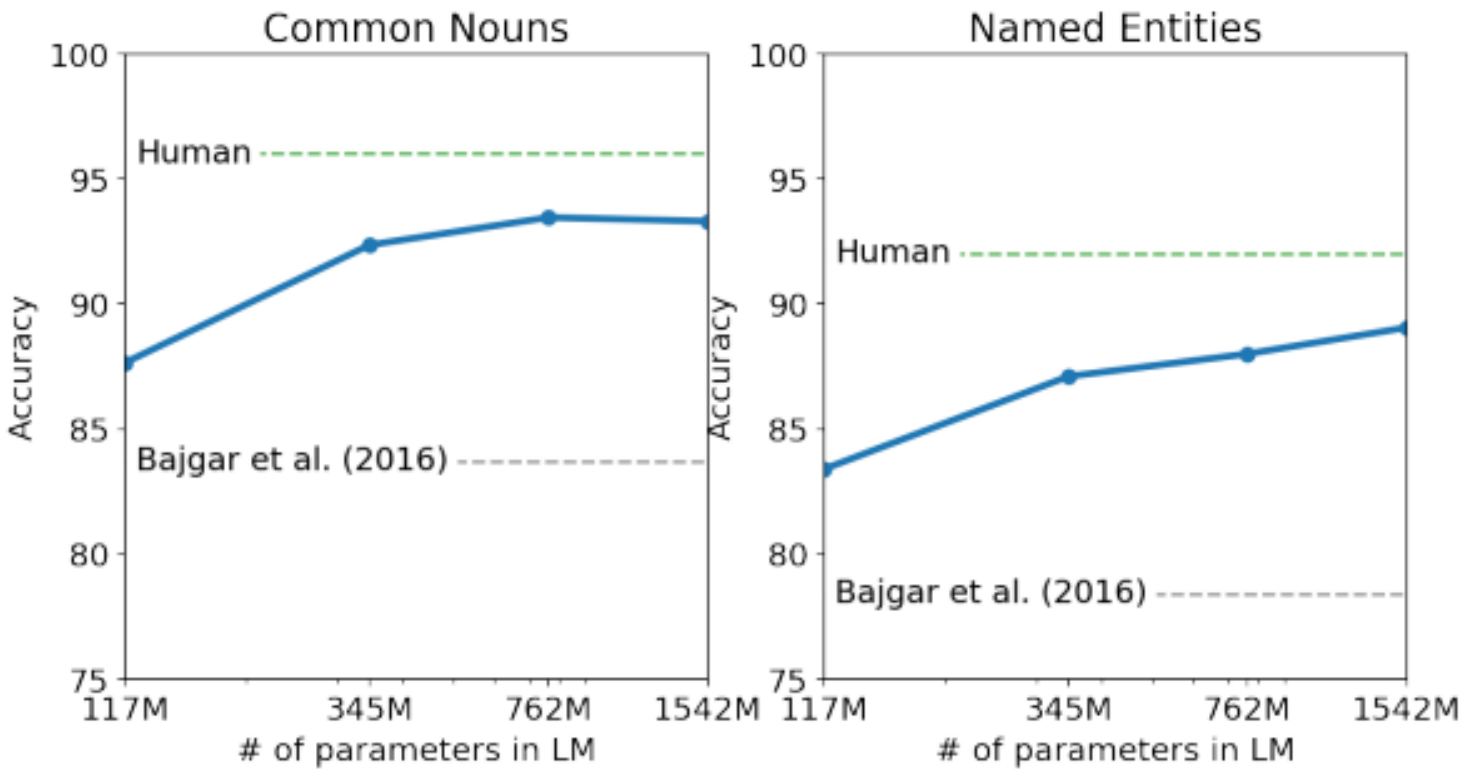
Language Models are Unsupervised Multitask Learners

	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

3. Experiments

2. Children's Book Test

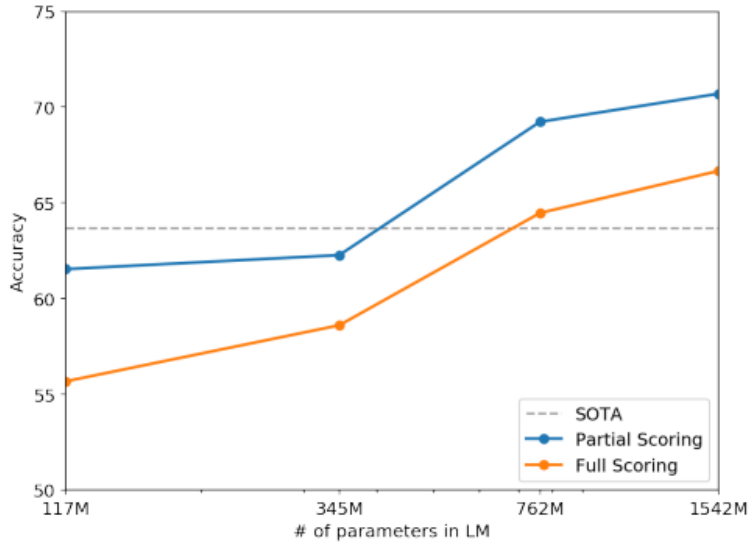
- 품사(고유명사, 명사, 동사, 전치사)에 따른 모델의 성능 비교



3. Experiments

3. **LAMBADA**
   - 텍스트의 장거리 의존성 평가
   - accuracy : 19% -> 52.66%
   - Perplexity : 99.8 -> 8.6

4. **Winograd Schema Challenge**
   - Text의 모호성을 푸는 작업을 통해 언어모델의 추론능력을 평가
   - 기존의 SOTA 모델보다 7% 높은 정확도



3. Experiments

5. Reading Comprehension

- Conversation Question Answering dataset (CoQA) : 7개 domain에 대한 Q,A를 포함하고 있는 dataset
- 이 dataset을 통해 언어모델의 문서 이해능력과 Q,A 능력을 동시에 평가
- SOTA 모델인 BERT에는 미치지 못했지만, GPT-2는 fine-tuning 없이 55의 F1 score 성능을 보여줌

6. Summarization

- Task-specific한 결과를 유도하기 위해 문서 이후에 TL;DR: 토큰을 추가
- TL;DR: 토큰 없이 실험했을 때보다 TL;DR: 토큰을 추가한 실험이 더 좋은 성능을 보임
- TL : Too long
- DR : didn't read

	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

3. Experiments

7. Translation

- WMT-14 English-French dataset 활용
- 영어-불어, 불어-영어 경우에서 비교 진행
- **영어-불어**: 성능이 좋지 못함
- **불어-영어**: 기존 모델보다 좋은 성능을 보여줌

8. Question Answering

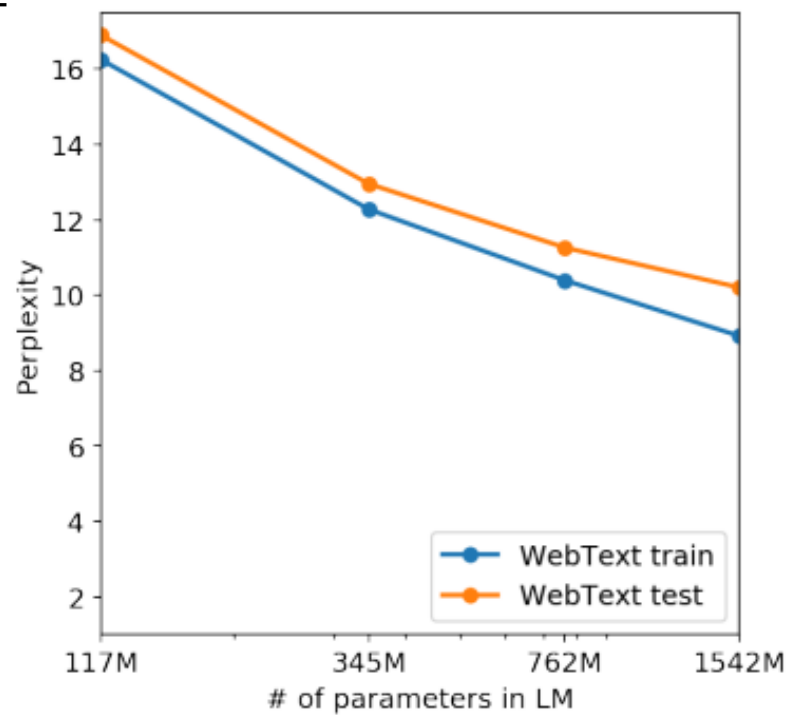
- 기존 모델보다 5.3배 높은 정확도를 보임

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%

4. Generalization vs Memorization

- Train set과 Test set의 과도한 중복(overlap)은 모델의 Memorization을 유도하고 Generalization 성능을 왜곡하여 나타낼 수 있음
- WebText dataset에서는 크지 않은 overlap을 보였지만, 어느정도 영향이 있었음을 확인
  - 기존 dataset이 가지고 있는 overlap보다는 크지 않음
- 모델이 underfitting 되어있음을 확인

	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%



5. Conclusion

- GPT-2 는 GPT-1 모델을 기반으로 하여 Unsupervised pre-training 작업을 극대화시킨 pretrained language model
- GPT-2는 기존의 pre-trained language model과 다르게 fine-tuning을 필요로 하지 않음
- GPT-2의 zero-shot 학습 성능은 독해 등에서는 좋은 성능을 보임
- 하지만 요약과 같은 문제에서는 기본적인 성능만 보여줌

Thank You

감사합니다.