Language Models are Unsupervised Multitask Learners

Alec Radford et al. (2019, openai)

김유리

참고!

본 논문의 내용을 보기 전에 아래 내용 먼저 보시는 것을 추천 드립니다!

Season#7 - 05. Attention Is All You Need

Season#5 - 09. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

openAI GPT paper- Improving Language Understanding by Generative Pre-Training

Index

Introduction

Introduction

NLP Task는 보통 task-specific한 dataset을 이용한 supervised learning이 일반적

→ **supervised learning**은 **데이터의 분포에 민감** (narrow expert)

Introduction

저자의 의도 : 새로운 dataset을 만들고 **labeling하는 작업 없이**

많은 task들에서 성능을 낼 수 있는 **general system**을 지향하자!

Introduction

general system을 위해서는?

single domain dataset에 대한 single domain training에 원인이 있지 않을까

Introduction

WebText라는(이 논문에서 만든 dataset) **수백만 dataset**으로 학습할 때

어떤 라벨링이나 explict supervision없이 위 LM을 학습시키고 싶음

→ **GPT-2** 제안

Methods

Methods
Approach

본 논문 접근법의 핵심은 **Language modeling방법으로 접근** 했다는 것

Language modeling : 문장의 확률을 계산하거나, 또는 이전 단어들이 주어졌을 때 다음 단어가 나올 확률을 계산하는 것

Methods
Approach

보통 Language modeling은 **조건부 확률을 이용해 sequential하게 단어를 예측**(conditional probabilities)

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

즉 $p(output|input)$하는 것인데 general system에서 여러 개의 task를 진행해야 한다면?

$\rightarrow p(output|input, task)$

즉, **어떤 task를 해야 할 지 model의 조건으로 넣어준다는 것**

Methods
Approach

어떤 task를 해야 할 지 model의 조건으로 넣어준다?

How? 선행 연구 중 **Multi-task learning** 논문을 래퍼런스로 소개(**McCann**)

Methods
Approach

McCann?

(translate to French, English text, French text), (answer the question, document, question, answer) 처럼

처음에 무슨 task를 진행하는지 언급해 주고

그 다음에 input들이 들어가서 원하는 결과(French text or answer) 를 출력하는 방식

(아래 예시 그림을 보면, 원하는 task를 question형식으로 줌)

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Methods
Approach

McCann과의 차이점

McCann은 **multi-task learning으로 실제 여러 개의 dataset**을 가져와서 **학습**한 것

GPT-2는 LM을 **supervised-learning**으로 한 것

즉, GPT-2는 **fine-tunning의 과정이 없다**고 볼 수 있으며, **어떤 task에도 적용 가능**

→ 실제로 학습 해 보면, Multi-task learning에 비해 **매우 느림**

Methods
Training Dataset

이전의 LM에서는 **single domain text**를 사용 (ex. News articles, Wikipedia, Fiction blocks 등)

다양한 도메인의 dataset도 있지만(Common crawl), 문제점 존재

→ 데이터의 많은 부분이 이해 할 수 없는 내용이며, 실제로 openAI가 해당 데이터로 연구를 시작했는데 비슷한 문제점 발견

Methods
Training Dataset

Web scrape을 해서 데이터(**WebText**)를 제작! 제작방법을 간단히 정리하면,

- 사람에 의해 필터링 되거나 큐레이팅 된 웹페이지만 스크랩
- Reddit의 link 사용
- 이중 Karma 3개 이상 받은 것만 사용(Karma : 페북의 좋아요)
- 모아진 텍스트인 WebText는 4500만개 링크의 텍스트 서브세트를 포함, HTML에서 텍스트를 추출하기 위해 우리는 Dragnet과 Newspaper를 조합해 사용
- 2017년 12월 이전 post만 가져왔음
- 40GB text, 총 8백만 문서 생성
- Wikipedia 문서는 부분은 제거 (다수의 링크가 위키피디아로 향하고 있어서 중복될 우려가 있었기 때문)

Methods
Input Representation

byte-level 접근이 generality 측면에서 좋고 Unicode string은 pre-processing, tokenization,
vocab size에 대한 걱정없이 LM에 적용 가능
→ **GPT-2** 모델에서는 **byte단위로 작동**

Methods
Model

openAI **GPT와 구조 거의 비슷함, 세부 사항 수정이 있음** (Transformer기반 아키텍쳐 사용)

- Moving normalization layer to the input of each sub-block
- Adding normalization layer after final self-attention model

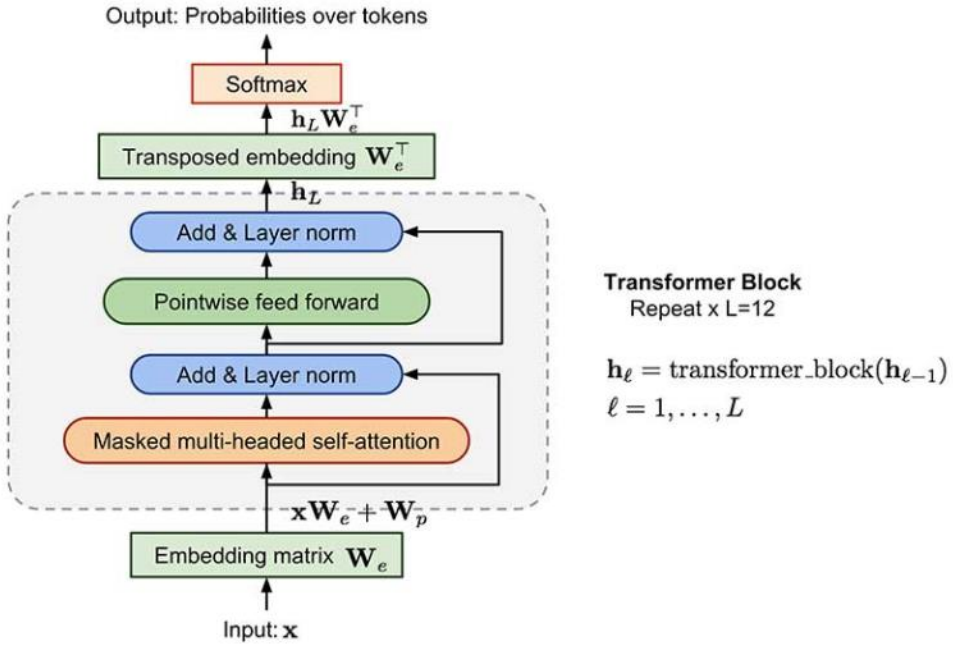

Fig. 2. The transformer decoder model architecture in OpenAI GPT.

Experiments

Experiments

실험결과를 한 줄로 요약하면, WebText만 가지고 실험 하니까 **underfitting**됨

Experiments

본 논문에서는 총 8가지 실험 진행

- Language Modeling
- Children's Book Test
- LANBADA
- Winograd Schema Challenge
- Reading Comprehension
- Summarization
- Translation
- Question Answering

Experiments

본 논문에서는 총 8가지 실험 진행

- **Language Modeling**
- Children's Book Test
- LANBADA
- Winograd Schema Challenge
- **Reading Comprehension**
- **Summarization**
- **Translation**
- Question Answering

Experiments
Language Modeling

4가지의 다른 시나리오로 학습 (다른 parameter를 가지는 4가지 모델을 train)

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.

Experiments
Language Modeling

8개 항목 중 **7개에서 SOTA**

Peen Treebank, WikiText-2같은 **소규모(1~2백만 개) 데이터셋**에서 눈에 띄는 **개선**

LANBADA, Children's Book 같은 **long-term dependencies 해결을 위한 dataset**에서도 눈에 띄는 **개선**

1BW에서 성능이 안 좋음 → 1BW's sentence level shuffling removes all long-range structure

	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).

Experiments
Reading Comprehension

CoQA (Conversation Question Answering dataset - 7개 domain으로 구성)로 test

Input : [Paragraph, QA history, final Q, token A]

token A : Answer하라는 의미의 token

Unsupervised-learning인데도 불구하고 dev셋에서 base-line system 3개 보다 좋다고 하는데,

CoQA 리더 보드에 있는 **BERT기반 모델에 비해서는 성능 떨어짐**

Experiments
Summarization

CNN and Daily Mail dataset으로 실험

Input : [Paragraph , TL, DR] (TL : Too Long, DR : didn't read)

LM으로 생성할 때, Top-2확률 word 중에서 random sampling 100개의 token 생성

생성된 token 중 처음 3개의 문장을 요약된 결과라고 지정

classic neural baseline 보다 살짝 좋은 정도의 성능 (Summarization만을 위한 모델보다는 성능 떨어짐)

Experiments
Translation

앞서 언급한 McCann처럼 무슨 task인지 처음에 알려줌

영어 → 프랑스어 task할 때 **Input : [Example sample pair, English sentence]** (반대의 경우도 마찬가지)

English → French : 성능 매우 안 좋음

French → English : SOTA는 아니지만 **다른 unsupervised learning보다 괜찮음**

openAI가 데이터를 만들 때, 영어가 아닌 언어로 된 문서들은 제거 했는데 성능 나옴

why? 10MB의 French language가 남아있었음

→ 이것은 더 많은 프랑스어 데이터가 있었다면 더 좋은 성능이 나왔을 수도 있다는 것을 나타냄

Generalization vs Memorization

Generalization vs Memorization

최근의 연구에서 **dataset 문제가 존재**

→ **CIFAR-10**에는 train/test 간 **3.3%의 overlap**이 있었음(2019년 어떤 논문에서 발견)

→ 그래서 새롭게 나온 데이터 셋이 **CIFAIR**

위의 문제**(overlap) 때문에 generalization performance가 over-reporting**될 수 있음

Generalization vs Memorization

본 논문에서는 WebText라는 dataset을 구축했기 때문에 **overlap이 일어나지 않도록 고려**해야 함

- Bloon filter 사용
- 8-gram 겹치는 정도를 데이터 간 비교 후 측정
- 많은 데이터에서 overlap 문제점 존재
- CoQA는 document(paragraph)는 15% 겹치지만 QA는 겹치는게 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%

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

overlap, similar text가 학습에 끼치는 영향을 알아내는 것이 중요

Generalization vs Memorization

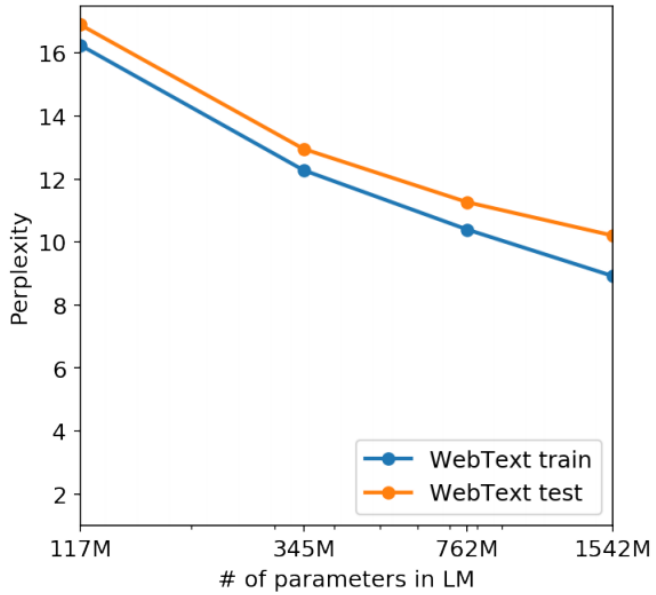

Figure 4. The performance of LMs trained on WebText as a function of model size.

위 그림을 보면 LM 파라미터가 늘어날수록 train, test 모두 perplexity가 떨어짐

즉, GPT-2 또한 아직 **underfitting** 됨

Discussion & Conclusion

Discussion & Conclusion

Discussion

- Supervision없이 task를 배우는 pre-training기술도 가능성이 있음
- 현재 상태는 충분한 capacity가 있을 때, 몇개의 baseline보다 성능이 좋은 정도
- GPT-2로 많은 zero-shot task에 성능을 측정 해 보았을 때, 가능성은 있지만 아직 이것의 fine-tunning ceiling은 명확하지 않음
- BERT에서 언급했듯 uni-directional representation은 비효율적이라고 했는데 GPT-2에서 이것을 어떻게 극복할지 불분명

Discussion & Conclusion
Conclusion

- 크고 다양한 dataset을 이용해 언어 모델을 학습하면 많은 dataset과 domain에 대해 좋은 성능이 나올 수 있음
- 본 논문에서 제안한 GPT-2는 8개의 dataset 중 7개에서 SOTA

Summary

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

Lasgest model

- 1.5 million Parameters Transformer

Generalization

- Past task specific methods lack generalization

Thank you

Appendix

GPT

GPT

Unsupervised training을 한 후 supervised training을 한다!

GPT

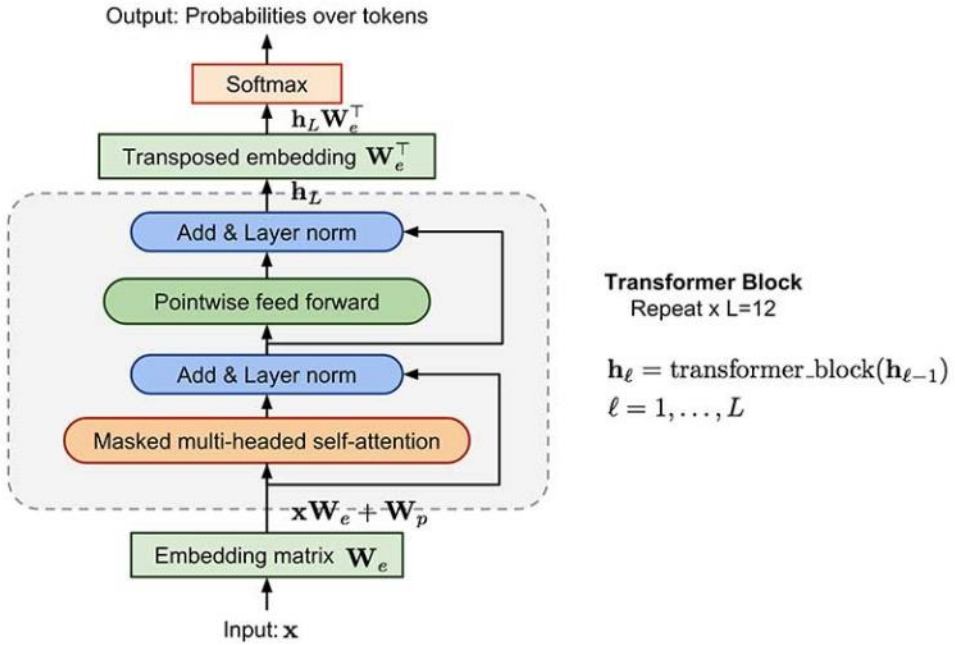

Fig. 2. The transformer decoder model architecture in OpenAI GPT.

토큰화 된 문장을 Token Embedding matrix로 구성하는 과정에서 Transformer의 decoder부분만 사용

기존의 Transformer가 Encoder/Decoder 6쌍으로 구성되었다고 하면 여기서는 Decoder만 12개(Multi-head)로 구성

간단히 요약하면, 문장단위로 Encoding(BPE)하고 Transformer Decoder를 거쳐 Context-level Embedding을 하는 과정을 통해 Unsupervised Learning을 사용한 LM을 학습