Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 분석

고려대학교 컴퓨터학과 박사과정
오동석

Contents

1) Background

2) GPT-3

3) GPT-3 능력 테스트

4) GPT-3 한계

5) 향후 연구

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

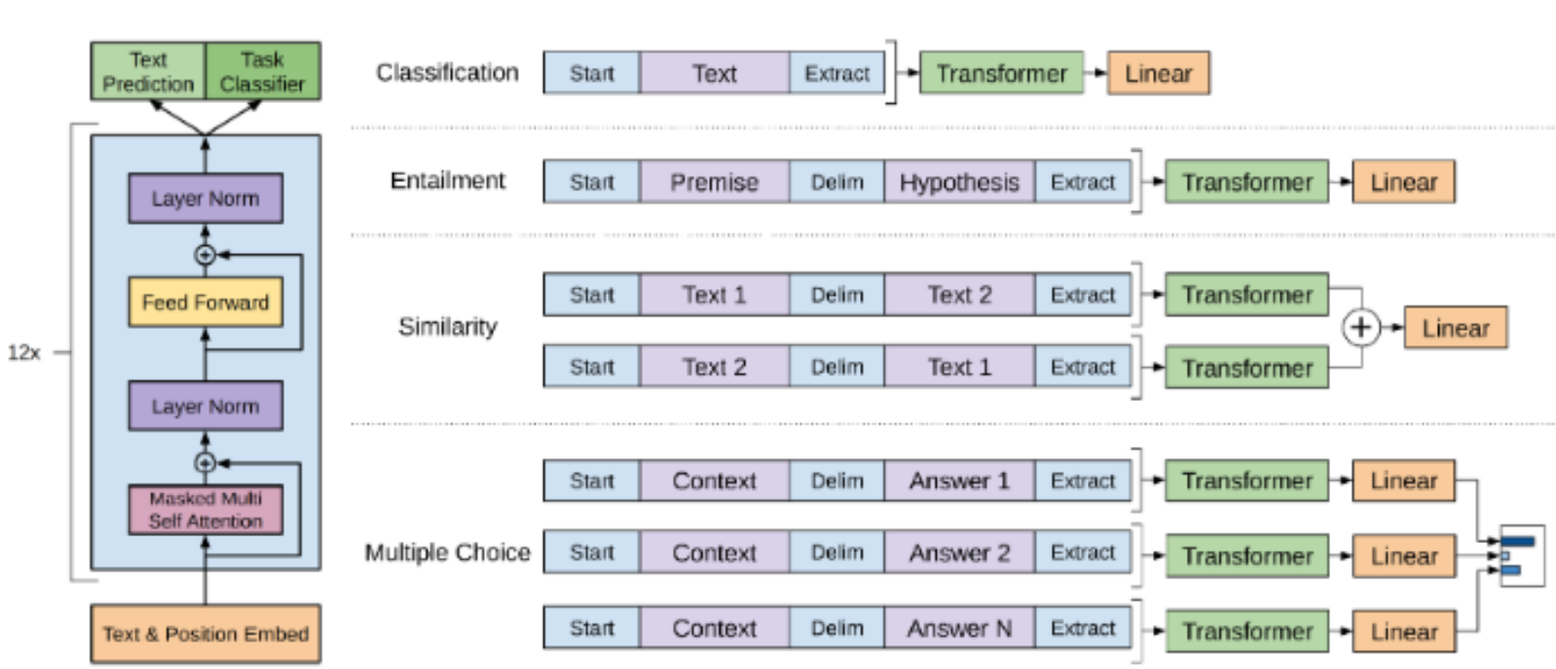
Natural Language Processing
& Artificial Intelligence

Background

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Background

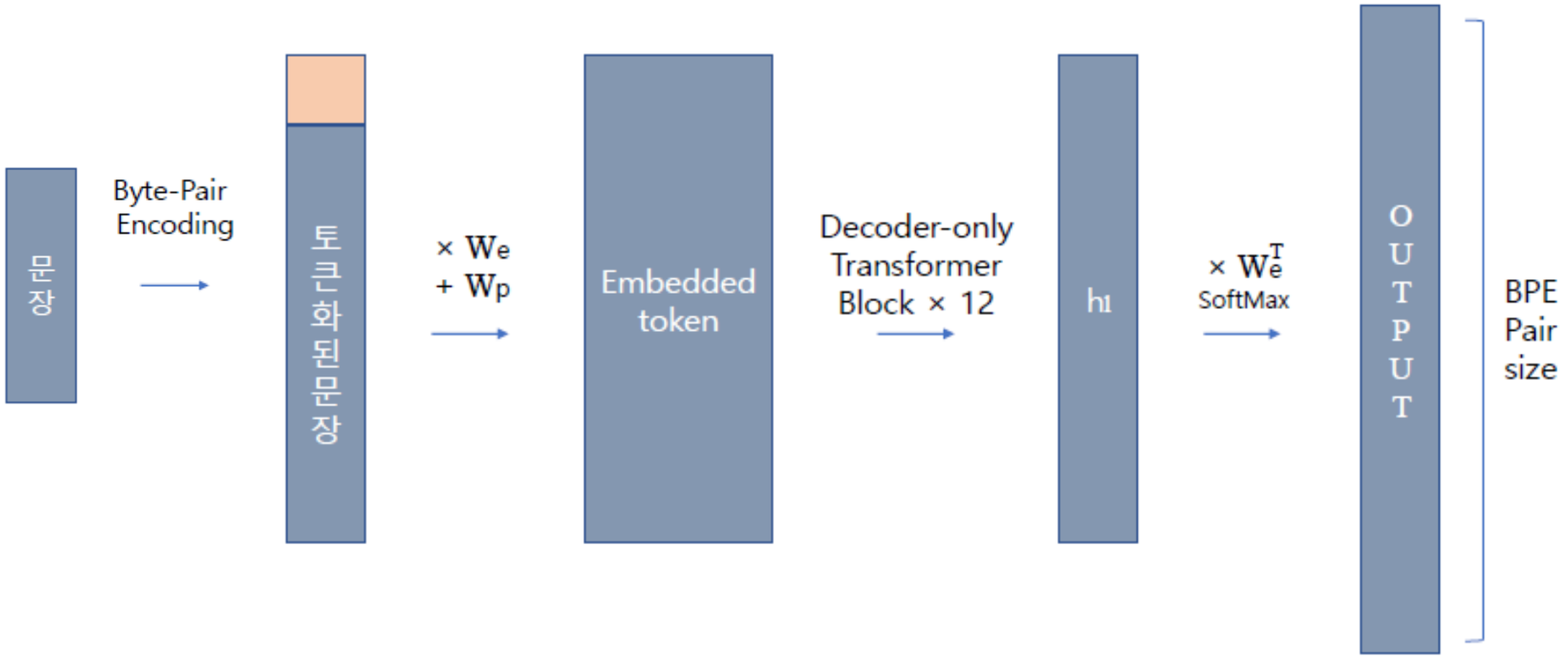
- GPT-1(Generative Pre-Training)
  - 대표적인 Auto Regressive Model
  - Transformer decoder를 이용한 언어 모델
  - Transformer decoder를 단방향으로 구축
  - BPE와 Text Prediction을 이용하여 언어 모델의 성능을 향상



Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Background

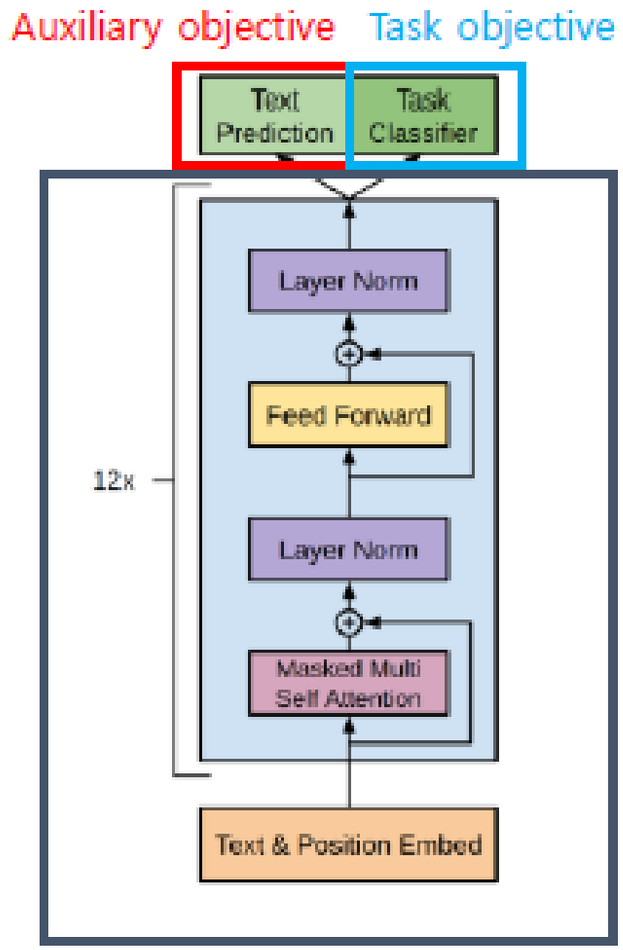
- GPT-1(Generative Pre-Training)
  - GPT Unsupervised Pre-training
    - BPE을 통해 문장을 토큰화 하고 다음 토큰을 예측하는 방법으로 GPT를 비지도 학습



Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Background

- GPT-1(Generative Pre-Training)
  - GPT Supervised fine-tuning
    - 세부 조정 단계에서 Text-prediction의 목적 함수를 보조로 사용하여 Task의 목적 함수에 포함
    - Inference 단에서는 보조 목적 함수를 사용하지 않음



Pre-trained network

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Background

- GPT-2(Generative Pre-Training)
  - GPT-1과 모델 차이는 없음
    - Layer Normalization이 pre-activation residual network처럼 각 sub-block의 입력으로 이동
    - 마지막 Self-Attention 모델 이후의 Layer Normalization 레이어 추가
  - GPT-1보다 10배 많은 파라미터를 사용 (14억개)
  - 학습데이터를 더 많이 사용(8백만개)

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Natural Language Processing
& Artificial Intelligence

GPT-3

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3

- 기존 GPT 모델들과의 파라미터 차이
  - GPT-1 : 1억개 파라미터
  - GPT-2 : 14억개 파라미터
  - GPT-3 : 1750억 파라미터

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3

- 모델 평가를 위한 3가지 Approach
  1. Few-Shot(FS)
     - 모델은 예시 태스크를 보게 되지만, 가중치 업데이트는 일어나지 않음
     - K(10~100) 개의 예제를 context 부분에 주고, 추론하려는 example의 결과를 완성하도록 하는 접근법
     - example>
       - 한국어를 영어로 번역 : <- Task Description
       - 집에 가고 싶어 : I want to go home.
       - 배고파 : I am hungry
       - 치킨 사줘 : _____
       - 모델은 문맥 인풋에 있는 예시들을 보고 _____ 부분에 "Buy me fried chicken"을 채워 넣어야 함

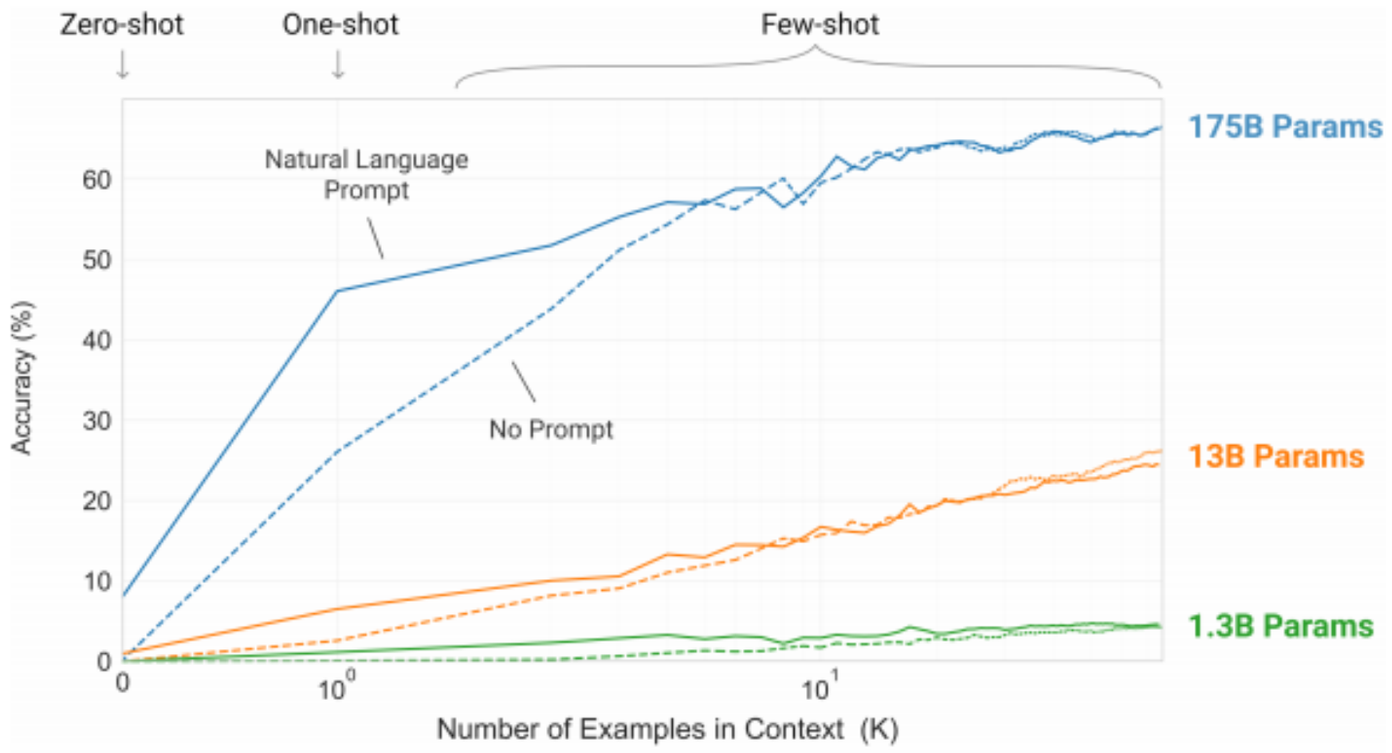
  2. One-Shot(OS)
     - FS 세팅과 같으나, 하나의 예제 만을 예시로 줌
     - example>
       - 한국어를 영어로 번역 : <- Task Description
       - 집에 가고 싶어 : I want to go home.
       - 배고파 :

  3. Zero-Shot(ZS)
     - Task를 설명하는 자연어 문구를 주고, 문제를 품
     - example>
       - 한국어를 영어로 번역 : <- Task Description
       - 집 :

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3

- 기본적인 사전학습은 GPT-2와 같음
- 3개의 접근법을 잘 수행하는 최적의 파라미터를 찾음
  - 학습과정에서 3개의 조건을 수행하지않고, 평가로만 사용
  - gradient 업데이트나 미세조정은 일어나지 않음
  - 해당 그래프는 “단어에 섞인 랜덤한 기호 제거하기 Task”에 대한 성능
  - K의 증가에 따라 정확도가 증가하는 것으로 오로지 문맥에 포함된 예제의 개수를 모델이 얼마나 잘 활용하는 지 분석



Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3

- 훈련 데이터 셋은 아래와 같음
- 크롤링한 데이터의 경우 45TB에서 정제하여 570GB로 줄여 학습을 진행
- 모든 토큰은 BPE 토큰을 기준

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3

- 번역에서 GPT-3 성능

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3

- SuperGLUE에서 GPT-3 성능

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Natural Language Processing
& Artificial Intelligence

GPT-3 능력테스트

* 해당 내용은 '핑퐁 ' 블로그를 참고하였습니다.

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 능력테스트

- Measuring Massive Multitask Language Understanding 논문에서 제안하는 AI 테스트
- 사지선다형 문제로 구성되어 있음
- example>
  - 수학
  - 법학
  - 물리학
  - 의학
  - 그 외에 외교, 경제학, 철학 등 57개의 분야에서 15,908개의 문제를 준비

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 능력테스트

- 분야별 예시

분야: 고등학교 수학

만약 4 daps = 7 yaps이고, 5 yaps = 3 baps라면, 42 baps는 몇 daps인가?

A. 28
B. 21
C. 40
D. 30

분야: 법학

한 백과사전 세일즈맨이 한 은둔자의 집에 도착했다. 그곳에는 "세일즈맨 거부. 어기는 사람은 처벌받음. 위험은 당신 책임."이라는 표지판이 있었다. 세일즈맨은 초대받은 사람은 아니었지만, 표지판을 무시하고 차를 몰고 집 쪽으로 들어갔다. 그가 코너를 도는 순간, 땅에 묻혀있던 폭발물이 터졌고, 세일즈맨은 다쳤다. 이 상황에서 세일즈맨이 그 은둔자에게 부상에 대한 보상을 요구할 수 있을까?

A. 있다. 단, 은둔자가 폭발물을 설치했을 때 단순히 침입을 막는 걸 넘어서 해하려는 의도가 있어야 한다.
B. 있다. 단, 은둔자가 그 폭발물을 설치했어야 한다.
C. 없다. 왜냐하면 세일즈맨이 경고를 명시한 표지판을 무시했기 때문이다.
D. 없다. 은둔자는 침입자가 자기 자신이나 가족을 해할 수도 있다는 합리적인 두려움을 갖고 있었기 때문이다.

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 능력테스트

- 분야별 예시

분야: 물리학

공을 손에 쥐고 가만히 떨어트리면 공은 $9.8m/s^2$의 가속도로 떨어진다. 공기 저항이 없다고 가정한 상태에서 아래쪽으로 공을 힘있게 던진다면, 손을 떠난 공의 가속도는?

A. $9.8m/s^2$
B. $9.8m/s^2$ 초과
C. $9.8m/s^2$ 미만
D. 던진 속도를 알려주지 않으면 알 수 없음

분야: 대학 수학

복소평면(complex z-plane)에서 $z^2 = ||z||^2$을 만족하는 점들이 이루는 도형은?

A. 두 점
B. 원
C. 반직선
D. 직선

분야: 의학

33세 남성이 갑상선암에 걸려 급성 갑상선 전절제술을 수행했다. 수술하는 동안 중등도의 출혈이 경부 좌측 부분의 몇몇 혈관에 나타나 봉합이 필요하게 되었다. 수술 후 혈장검사 수치로 칼슘 수치는 7.5 mg/dl, 알부민 수치는 4 g/dl, 부갑상선호르몬 수치(PTH)는 200 mg/ml였다. 다음 중 어느 혈관의 손상이 이 환자의 검사결과와 상관있는가?

A. 목갈비동맥의 가지
B. 바깥목동맥의 가지
C. 갑상목동맥의 가지
D. 속목정맥의 지류

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 능력테스트

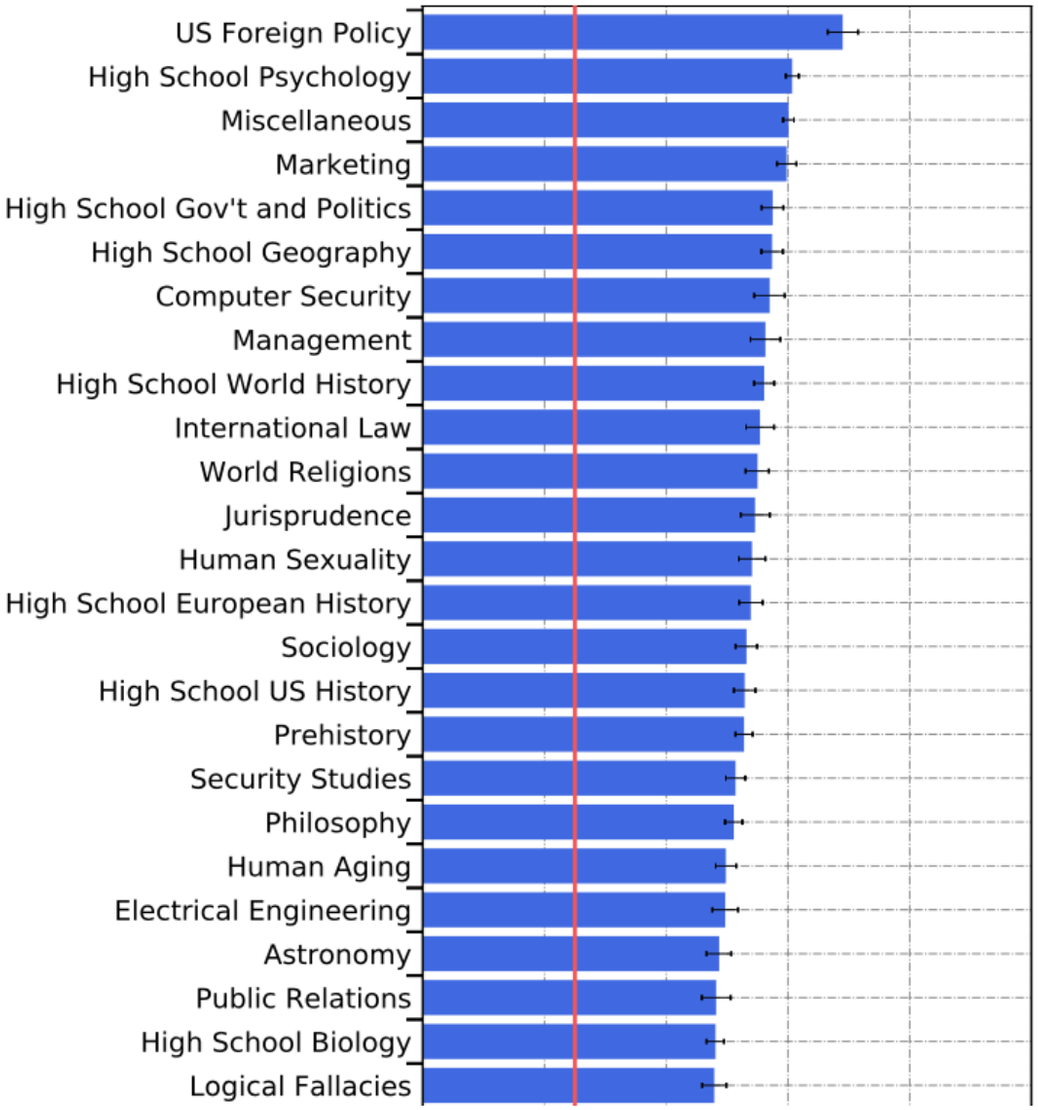
- GPT-3의 성능
  - 43.9% 성능을 보임

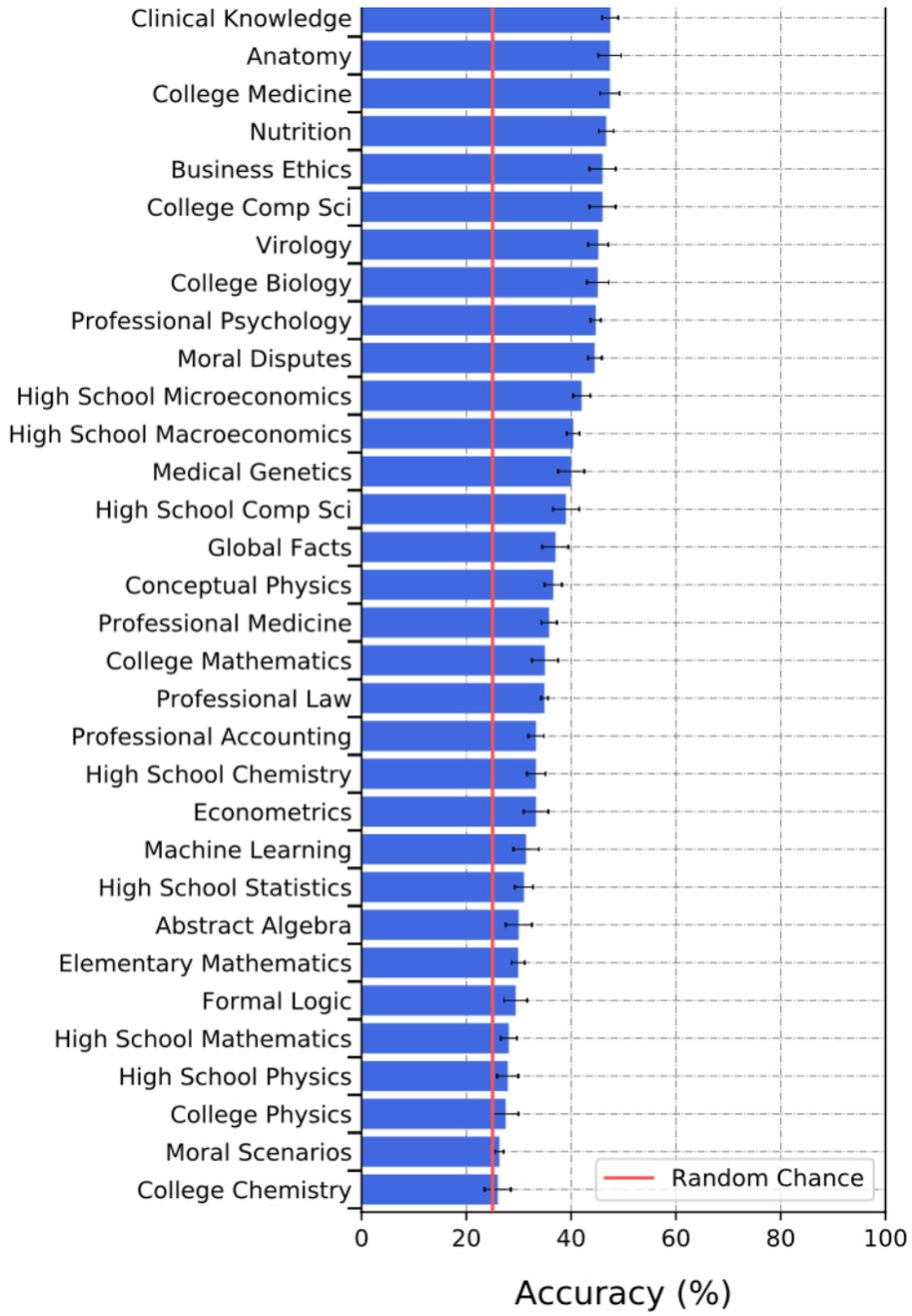
Model	Humanities	Social Science	STEM	Other	Average
Random Baseline	25.0	25.0	25.0	25.0	25.0
T5	24.9	25.2	24.8	24.4	24.8
UnifiedQA	38.0	41.5	32.2	42.1	38.5
GPT-3 Small	24.4	30.9	26.0	24.1	25.9
GPT-3 Medium	26.1	21.6	25.6	25.5	24.9
GPT-3 Large	27.1	25.6	24.3	26.5	26.0
GPT-3 X-Large	40.8	50.4	36.7	48.8	43.9

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 능력테스트

- GPT-3의 성능
  - 43.9% 성능을 보임





Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 능력테스트

- GPT-3의 평가 분석
  - GPT-3는 (당연히) 완벽하지 않다 (43.9점)

  - GPT-3는 공학/자연과학 분야에 약함
    - GPT-3 학습에 사용된 데이터가 인문/사회과학적인 데이터 비중이 높음
    - 인문/사회과학 분야는 암기식 테스트가 많음
    - 공학/자연과학 분야는 추론 능력을 확인하는 테스트가 많음

  - 모델 크기가 중요하다
    - GPT-3가 GPT-2와 같은 모델임을 생각 해봤을 때, 모델 크기의 중요성을 보여주는 결과임

- 특화된 모델은 효율이 높다
  - UnifiedQA의 벤치마크 성능은 39점

- AI의 벤치마크와 인간의 벤치마크가 같아지다
  - 지금까지의 모든 벤치마크는 "인간에게는 매우 쉬운데, AI에게는 어려운"문제들로 구성되어 있었음(GLUE, SuperGLUE)

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Natural Language Processing
& Artificial Intelligence

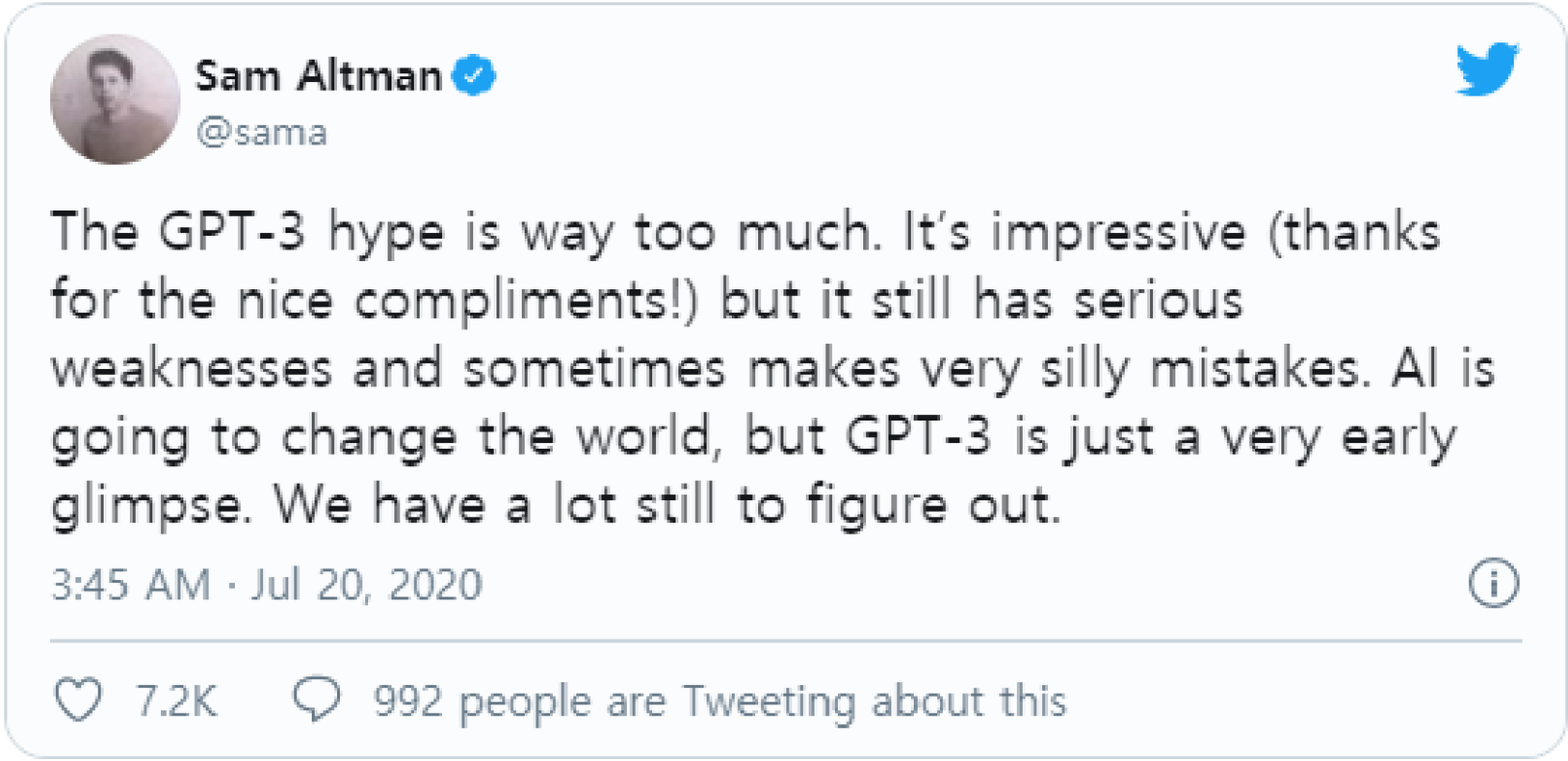
GPT-3 한계

* 해당 내용은 '핑퐁 ' 블로그를 참고하였습니다.

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 한계

- GPT-3를 만든 OpenAI대표 Sam Altman의 내용



Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

GPT-3 한계

- GPT-3가 다 잘하는 것은 아님
  - 모든 Task에서 높은 성능을 보이는 것이 아님
  - 특정 Task를 위해서 학습한 모델보다는 낮은 성능을 보임

- GPT-3는 현실의 물리적인 상식을 다루는데 한계가 있음
  - "치즈를 냉장고에 넣으면 녹을까?"
  - 사람 : 안 녹음, GPT-3 : 녹음
  - GPT-3은 세상의 모든 지식을 글로만 학습했기 때문에 실제로 경험해서 얻는 정보와 관련해서 낮은 성능을 보임

- 모델이 너무 큼
  - 모델이 커서 실제 활용하기 쉽지 않고, 학습데이터가 많이 필요함

- 기억을 못함
  - 기반이 되는 Transformer 구조에는 '기억'이라는 개념이 없음
  - 현재 문맥을 입력으로 주고 그에 따른 출력을 얻을 뿐

- 다음 단어 예측 학습 방법의 한계
  - 사람은 학습할 때 텍스트를 보고 다음 단어를 예측하는 방식으로도 배우지만 텍스트를 이해하고 원리를 배우기도 함
  - GPT-3는 통계적으로 가장 그럴듯한 다음 단어를 차례차례 생성할 뿐 텍스트를 이해하고 말을 하는 것은 아님

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

Natural Language Processing
& Artificial Intelligence

향후 연구

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

향후 연구

- 현재 공개된 사전학습 모델들에 대한 분석 및 개선 방법 제안
  - 자연어 이해를 정확하게 하고 있는지 분석 및 개선 방법 제안
  - 지식 추론을 어느 수준까지 할 수 있는 지 분석 및 개선 방법 제안
  - 블랙박스로 이루어진 딥러닝 모델들의 설명 가능 인공지능 기술 개발

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab

감사합니다.

Copyright © 2019 Natural Language Processing & Artificial Intelligence Lab