23.07.27 _ 7주차

딥러닝 논문 요약 및 구현 스터디

발표자

정명찬

GPT-3

Language Models are Few-Shot Learners

Abstract

- 수천 수만 개 예제 학습으로 Fine-tuning -> 시간, 비용 문제
- 인간은 몇 가지 예제만으로도 새로운 언어 작업 수행 가능

- Fine-tuning 없이, 언어 모델 자체를 크게 확장
- Few-shot learning으로 SOTA fine-tuning 기법과 경쟁력 가짐
- 이전 LM보다 10배 많은 params
- 다양한 자연어처리 태스크에서 강력한 성능

I. Introduction

Fine-Tuning 방식의 문제점

1. NLP 특정 태스크는 대량의 레이블된 데이터셋을 구하기 어려움
2. 언어모델의 사전 학습에는 일반화 정보 흡수, fine-tuning에는 좁은 분포의 작업 학습
3. 인간의 언어 작업과의 괴리 – 인간은 대규모 레이블된 데이터를 필요로 하지 않음

I. Introduction

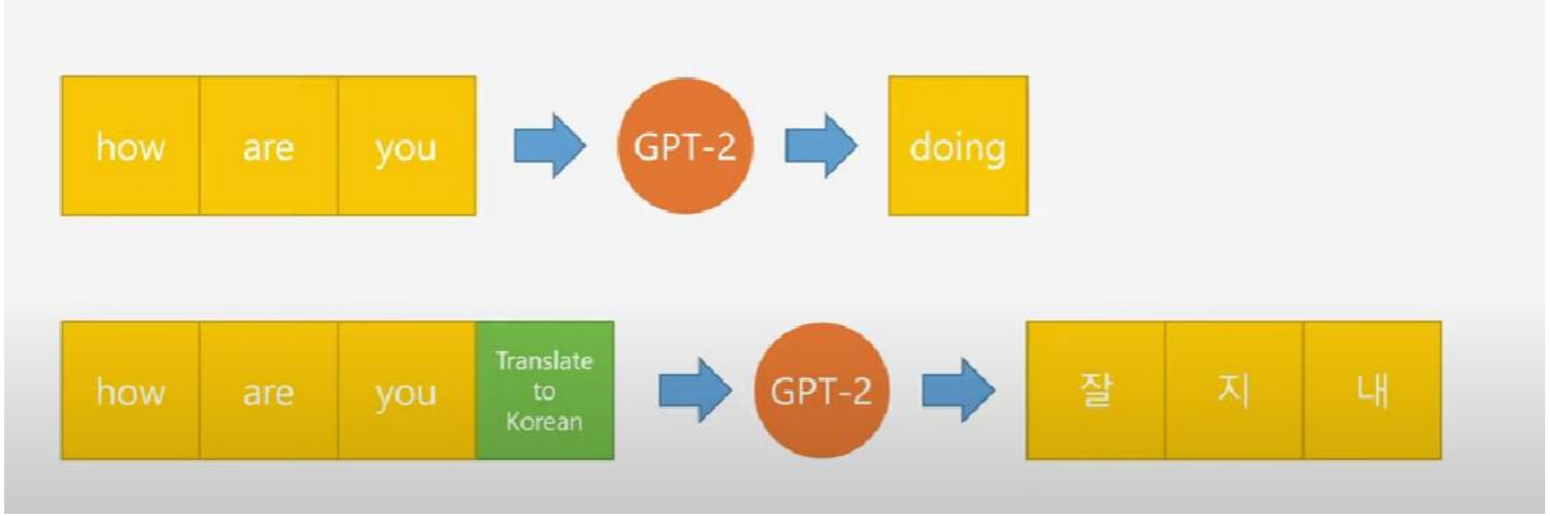
Meta-Learning

- 언어 모델 학습 시, 텍스트 입력을 특정 작업 명세의 형태로
- 그러나, 여전히 fine-tuning 방식보다 성능 떨어짐 -> few-shot learning으로 극복

I. Introduction

Meta-Learning

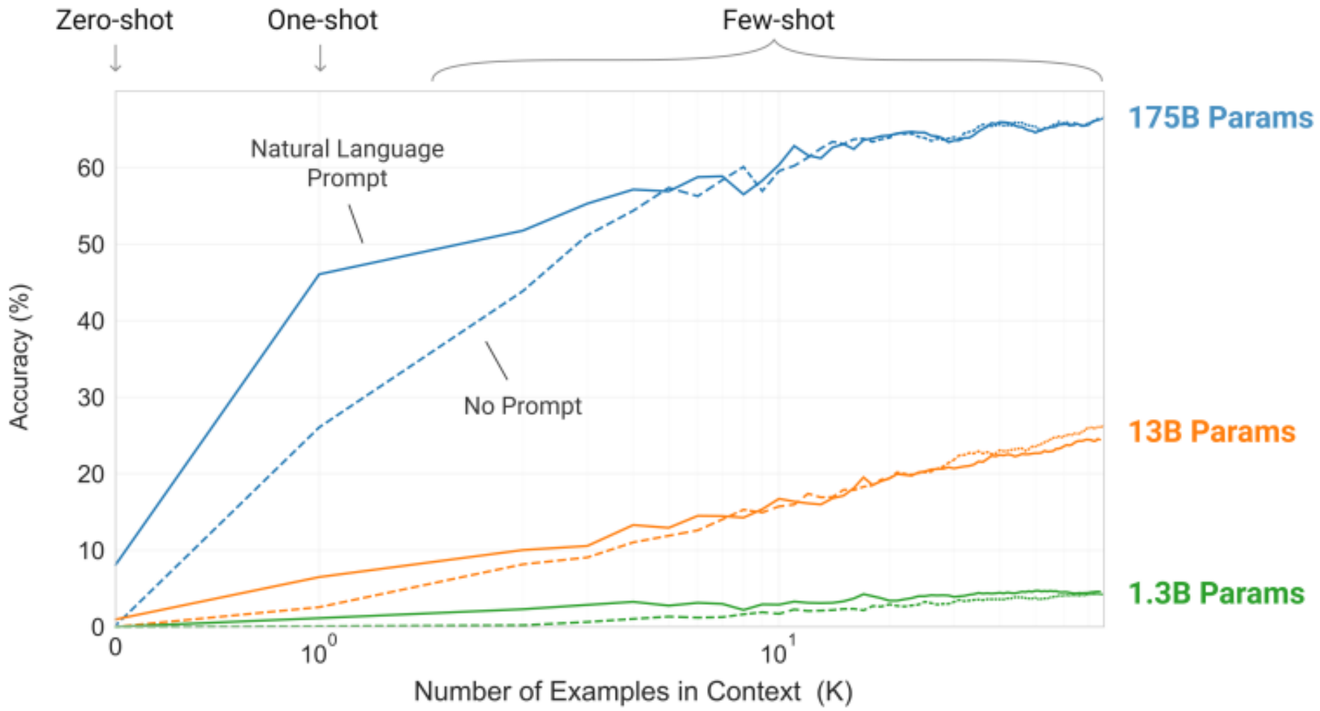
- 언어 모델 학습 시, 텍스트 입력을 특정 작업 명세의 형태로
- 그러나, 여전히 fine-tuning 방식보다 성능 떨어짐 -> few-shot learning으로 극복



I. Introduction

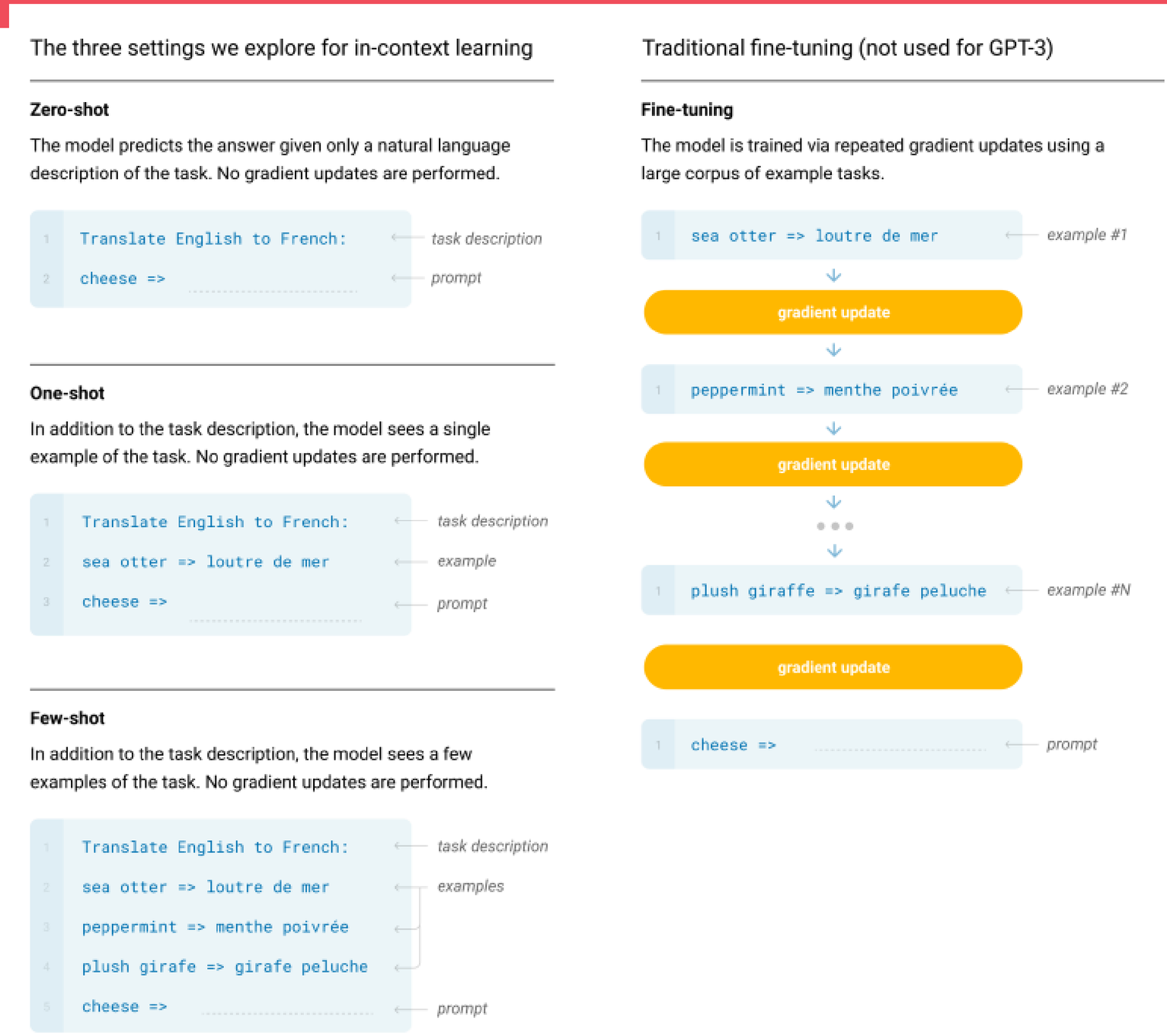
Few-shot (in-context) learning

- 그래디언트 업데이트하는 것이 아닌 단순 예제 시연



II. Approach

- Zero-shot(GPT-2): 무슨 작업을 하는지만 알려주고, 예시를 주지 않음
- One-shot: 작업의 예시 한 개 제공
- Few-shot: 작업의 예시 몇 개 제공



II. Approach

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

II. Approach

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Weight in training mix -> dataset의
  크기에 따라 비율 조정하지 않고,
  학습 진행

II. Approach

- Few-shot 을 위해, 학습 데이터셋에서 무작위로 K개의 예시를 추출하여 조건으로 설정

- 일부 태스크에서는 예시 대신 지시사항(프롬프트)를 입력하기도 함

- Zeroshot, oneshot, fewshot에 대한 결과 보고

II. Approach

- Few-shot 입력 예시

Context →

```
Article:
Informal conversation is an important part of any business
relationship.Before you start a discussion,however,make sure you understand
which topics are suitable and which are considered taboo in a particular
culture.  Latin Americans enjoy sharing information about their local
history, art and customs.You may expect questions about your family,and
be sure to show pictures of your children.You may feel free to ask similar
questions of your Latin American friends.The French think of conversation
as an art form,and they enjoy the value of lively discussions as well as
disagreements.  For them,arguments can be interesting and they can cover
pretty much or any topic ---- as long as they occur in are respectful and
intelligent manner.
In the United States,business people like to discuss a wide range of
topics,including opinions about work,family,hobbies,and politics.  In
Japan,China,and Korea,however,people are much more private.They do not
share much about their thoughts,feelings,or emotions because they feel
that doing so might take away from the harmonious business relationship
they're trying to build.Middle Easterners are also private about their
personal lives and family matters.It is considered rude,for example,to ask
a businessman from Saudi Arabia about his wife or children.
As a general rule,it's best not to talk about politics or religion with
your business friends.This can get you into trouble,even in the United
States,where people hold different religious views.In addition,discussing
one's salary is usually considered unsuitable.Sports is typically a
friendly subject in most parts of the world,although be careful not to
criticize national sport.Instead,be friendly and praise your host's team.

Q: What shouldn't you do when talking about sports with colleagues from
another country?

A: Criticizing the sports of your colleagues' country.

Q: Which is typically a friendly topic in most places according to the
author?

A: Sports.

Q: Why are people from Asia more private in their conversation with others?

A: They don't want to have their good relationship with others harmed by
informal conversation.

Q: The author considers politics and religion _ .

A:
```

III. Results

PTB (The Wallstreet Journal) 데이터셋으로 언어 모델링 평가

- 현대 인터넷에서 구할 수 있는 이전의 것으로, 평가에 유용

-> Zero-shot으로 더 낮은 Perplextity

Setting	PTB
SOTA (Zero-Shot)	35.8[a]
GPT-3 Zero-Shot	**20.5**

III. Results

번역, 대명사의 가리키는 대상 찾기,

상식 추론, 기계 독해 등의 NLP task에서 대부분 SOTA를 능가하진 못하지만,

fine-tuning하지 않고도, few-shot learning으로 성능이 개선되는 걸 확인

SuperGLUE (자연어처리 태스크 모음)

Fine-tuned BERT-Large보다 더 좋은 성능을 보이는 경우도 있음

NLI (두 문장 사이의 관계 이해)

Fine-tuned SOTA보다 뒤쳐진 성능

Few-shot 이용할 경우 성능 개선되지만 여전히 어려운 과제

III. Results

Synthetic and Qualitative Tasks (산술능력, 단어의 글자 재배열, SAT스타일 문제, 영문법 수정, 기사생성**)**

산술 능력

- 일반적인 수 계산이 아닌 “자연어처리” 문제로 연산 해결

Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

III. Results

Synthetic and Qualitative Tasks (산술능력, 단어의 글자 재배열, SAT스타일 문제, 영문법 수정, 기사생성**)**

글자 조작

- lyinevitab -> inevitably
- Criroptuon -> corruption

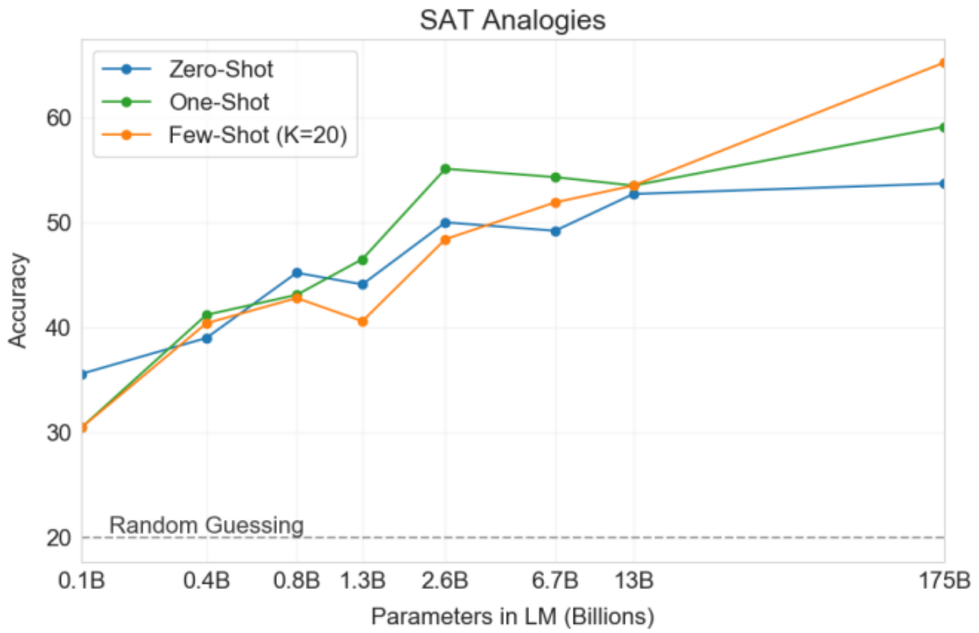
Setting	CL	A1	A2	RI	RW
GPT-3 Zero-shot	3.66	2.28	8.91	8.26	0.09
GPT-3 One-shot	21.7	8.62	25.9	45.4	0.48
GPT-3 Few-shot	37.9	15.1	39.7	67.2	0.44

III. Results

Synthetic and Qualitative Tasks (산술능력, 단어의 글자 재배열, SAT스타일 문제, 영문법 수정, 기사생성**)**

SAT 스타일

- “audacious is to boldness as (a) sanctimonious is tohypocrisy, (b) anonymous is to identity, (c) remorseful is to misdeed, (d) deleterious is to result, (e) impressionable is to temptation”.

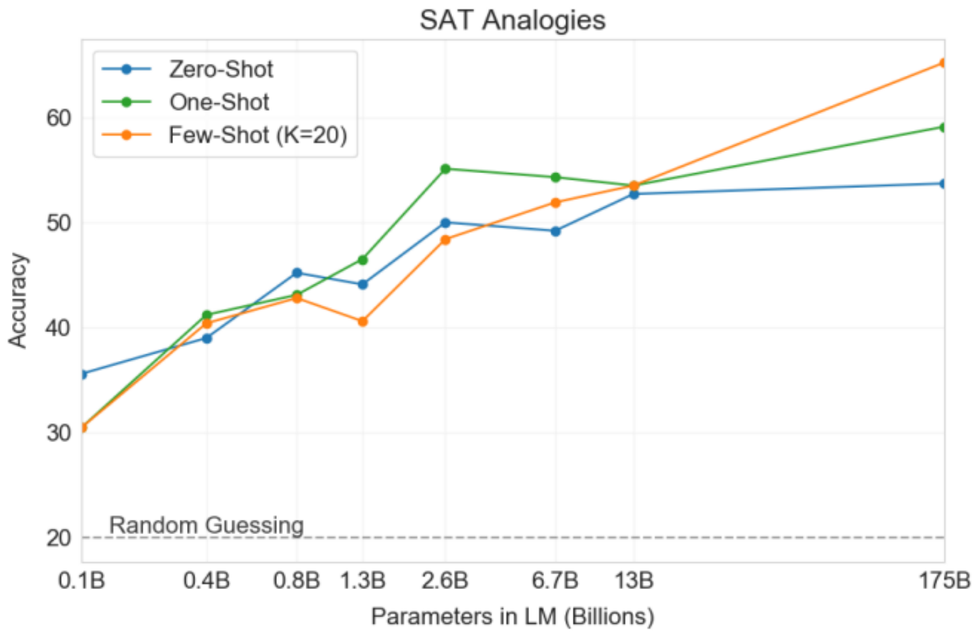


III. Results

Synthetic and Qualitative Tasks (산술능력, 단어의 글자 재배열, SAT스타일 문제, 영문법 수정, 기사생성**)**

SAT 스타일

- “audacious is to boldness as (a) sanctimonious is tohypocrisy, (b) anonymous is to identity, (c) remorseful is to misdeed, (d) deleterious is to result, (e) impressionable is to temptation”.



III. Results

Synthetic and Qualitative Tasks (산술능력, 단어의 글자 재배열, SAT스타일 문제, 영문법 수정, 기사생성**)**

뉴스 기사 생성

- 첫 문장을 주고 짧은 기사를 작성하라고 함

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

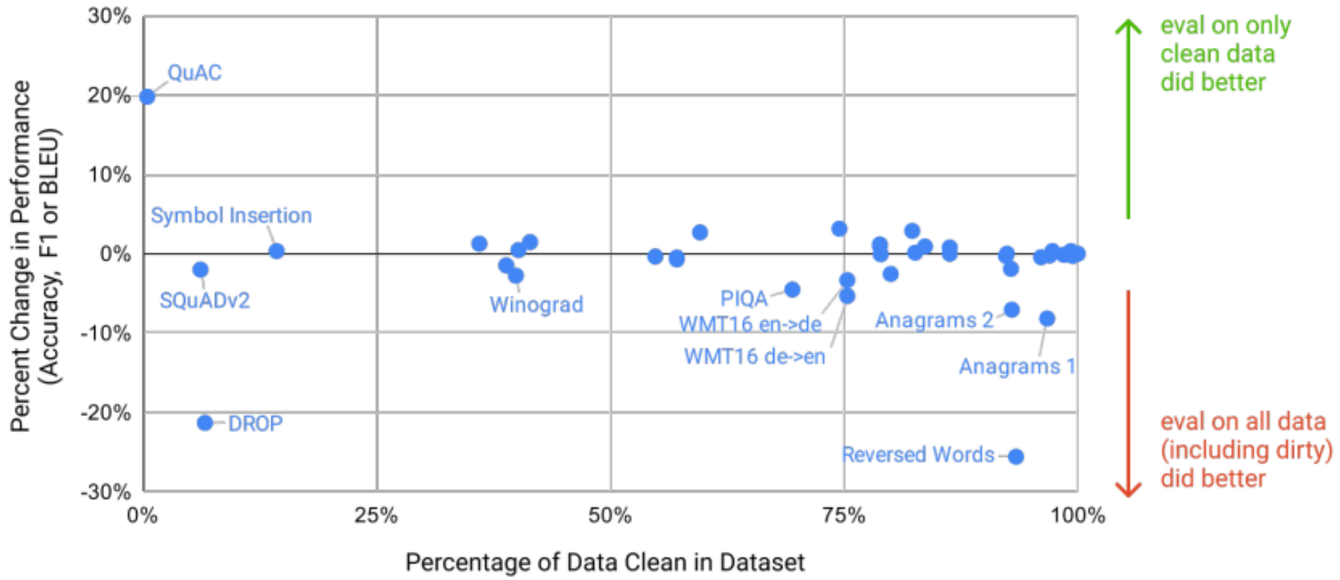
IV. Measuring and Preventing Memorization Of Benchmarks

Data contamination

GPT-3가 학습한 데이터는 대부분 인터넷을 통해 얻은 것 -> 평가할 데이터셋이 섞여 있을 수 있음

그러나, GPT-3 학습 곡선을 볼 때, Validation loss와 Train Loss 감소 추세가 같음

데이터 중복을 제거한 cleaned 버전에서도 기존과 성능이 크게 차이나지 않음



IV. Limitations

1. 성능

- 몇몇 NLP tast에서 성능 저하 보임

2. 구조

- bi-directional이 아니기에 한계

3. 토큰 중요도

- 어떤 토큰이 중요하고, 아닌지 고려하지 않음

4. 비용

- 훈련비용이 매우 큼
- Task-specific한 연구 필요

5. 설명가능성

데이터 자체적으로 존재하는 Bias에 대해 컨트롤 안됨. 사회적 문제 가능성

V. Broader Impacts

1. 모델 악용

- 가짜뉴스, 스팸, 피싱, 가짜 에세이 등 생성 -> 선동을 위해 사용 등 악용

2. 젠더

- 직업과 관련된 경우 남성이라고 예측, 여성의 특성을 묘사하는 경우 외모적인 부분 예측

3. 인종, 종교

- 아시안에 대해 일관되게 긍정적 표현 사용, 흑인에 대해 부정적 표현 사용
- 특정 종교에 부정적인 단어 예측

4. 비용

- 훈련비용이 매우 큼
- Task-specific한 연구 필요

5. 에너지 사용

- 기존 GPT-2의 수백배에 해당하는 전력소비

VI. Conclusion

간단한 모델 아키텍처로 고품질의 단어 벡터 훈련

낮은 계산 복잡성으로 더 큰 데이터 세트에서 정확성 높임

기존의 NN 기반 임베딩 방법론이 NLP 태스크에 적용 되었는데,
이 논문의 방법론으로 더욱 효율적으로 적용 가능 할 것