Language Models are Unsupervised Multitask Learners

Name
조병웅

NLP

2024/05/14

TRAIN AND TEST

Contents

TRAIN AND TEST

Introduction

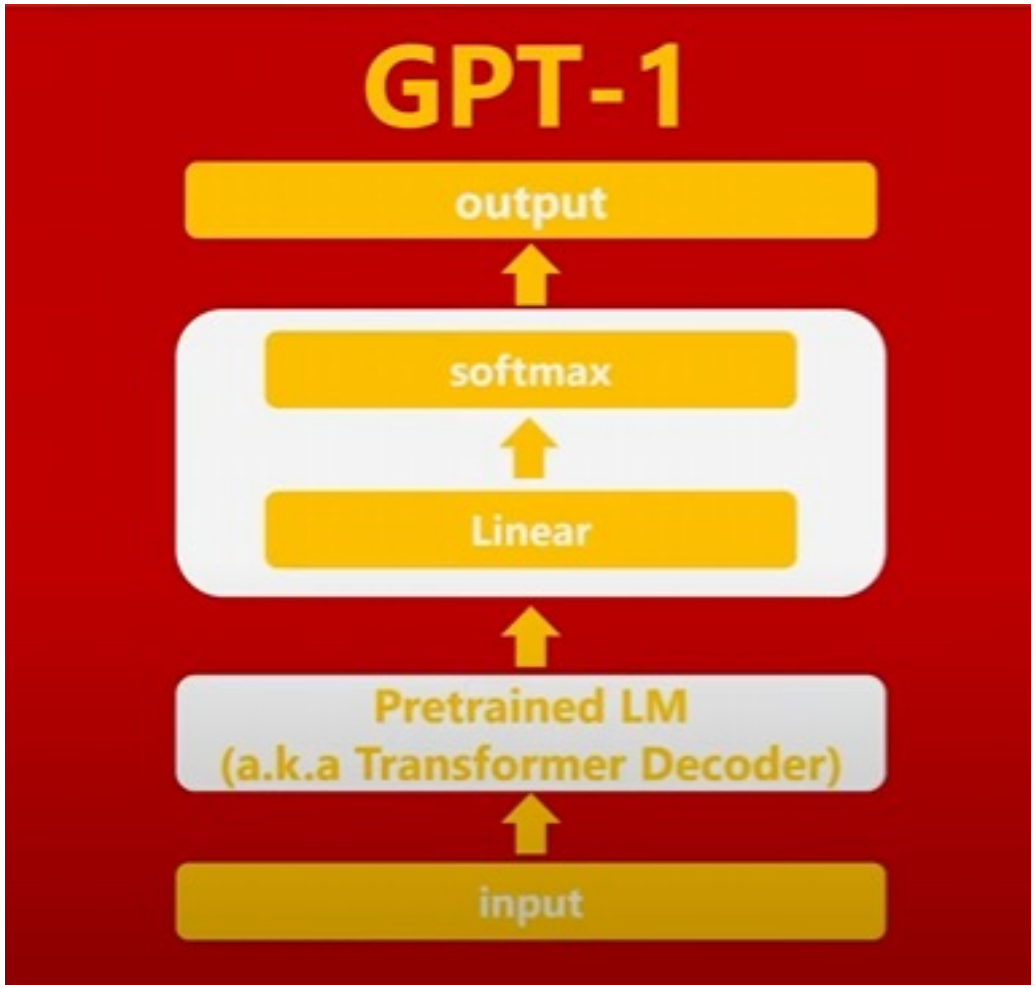
GPT-2의 behind

➔ GPT의 등장

- Transformer 구조를 사용함으로써 약간의 fine-tuning 만으로 여러 Task에서 SOTA를 달성.

- General한 성능향상을 통해 Zero-shot 성능에 대한 가능성을 열어 줌.

여전히 fine-tuning이 존재하며, Task-specific하여 범용적이지 못하다는 한계점 존재.



TRAIN AND TEST

Introduction

GPT-2의 behind

➔ 대규모 데이터 학습의 경쟁

- Transformer 구조의 Encoder를 사용하는 BERT의 등장

- GPT와의 비교를 통한 LM의 크기의 경쟁

#보다 더 큰 LM 확장

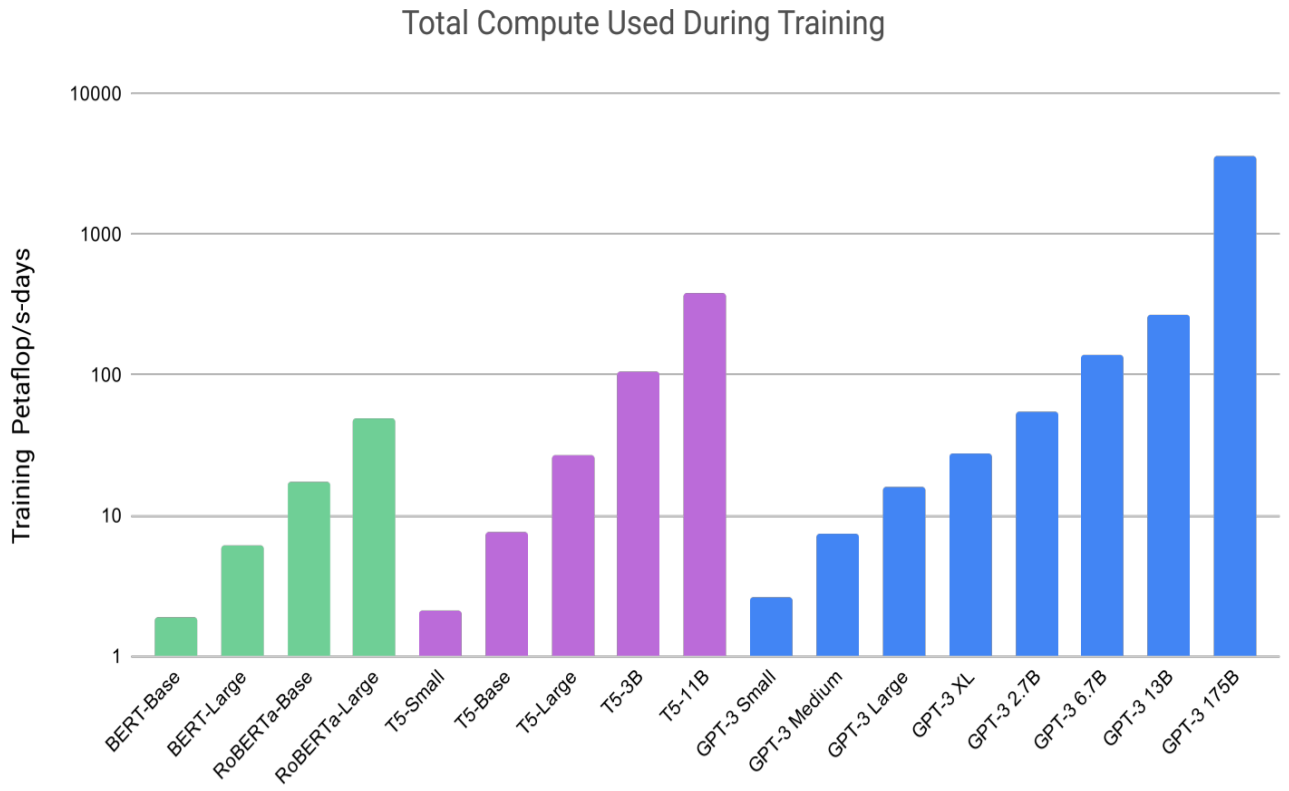


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

TRAIN AND TEST

Introduction

GPT-2의 목적

➔ Fine-tuning없이 zero-shot으로 Task를 수행할 수 있는 General LM을 개발하는 것

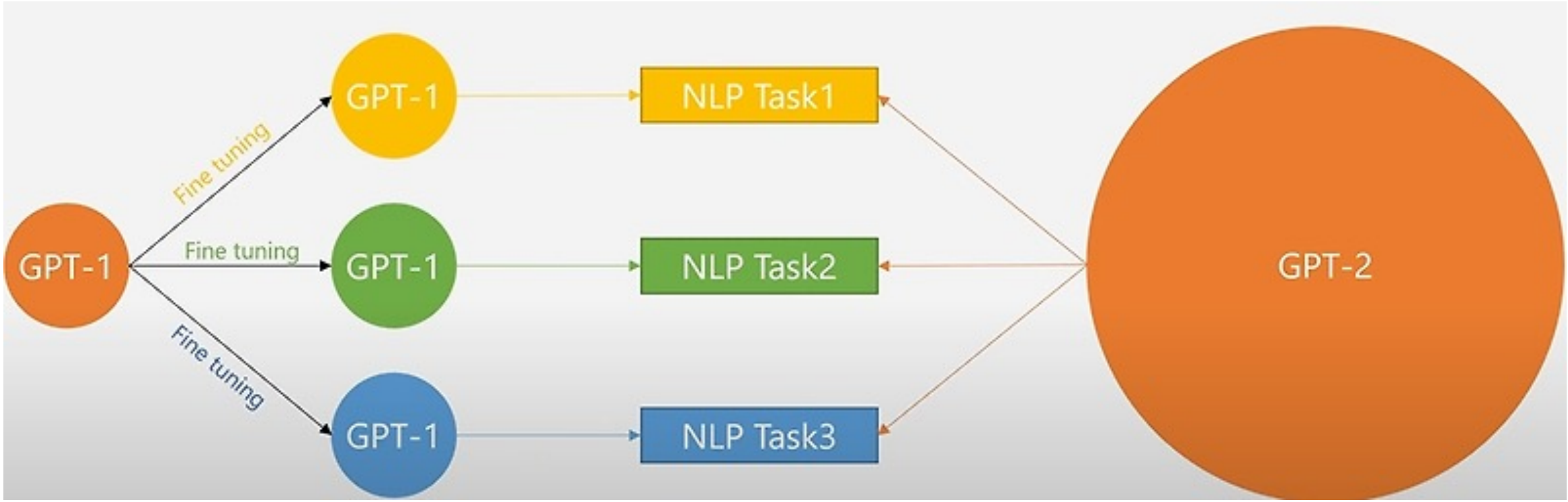
- 더 큰 Dataset, 고용량의 모델의 사용과, 이를 위한 비지도 학습의 도입(기존 GPT)
- Task-specific한 모델이 아닌, General한 모델의 개발
- GPT1과 달리 paramete나 모델 구조의 변화 없이 zero-shot setting 하에서 downstream Task를 수행

TRAIN AND TEST

Introduction

GPT-2의 목적

➔ **Fine-tuning없이 zero-shot으로 Task를 수행할 수 있는 General LM을 개발하는 것**



Approach

Language modeling

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

- 기존 GPT의 조건부 확률 분포 framework는 $p(output|input)$
- GPT-2의 조건부 확률 분포는 범용성을 위해 $p(output|input, task)$

TRAIN AND TEST

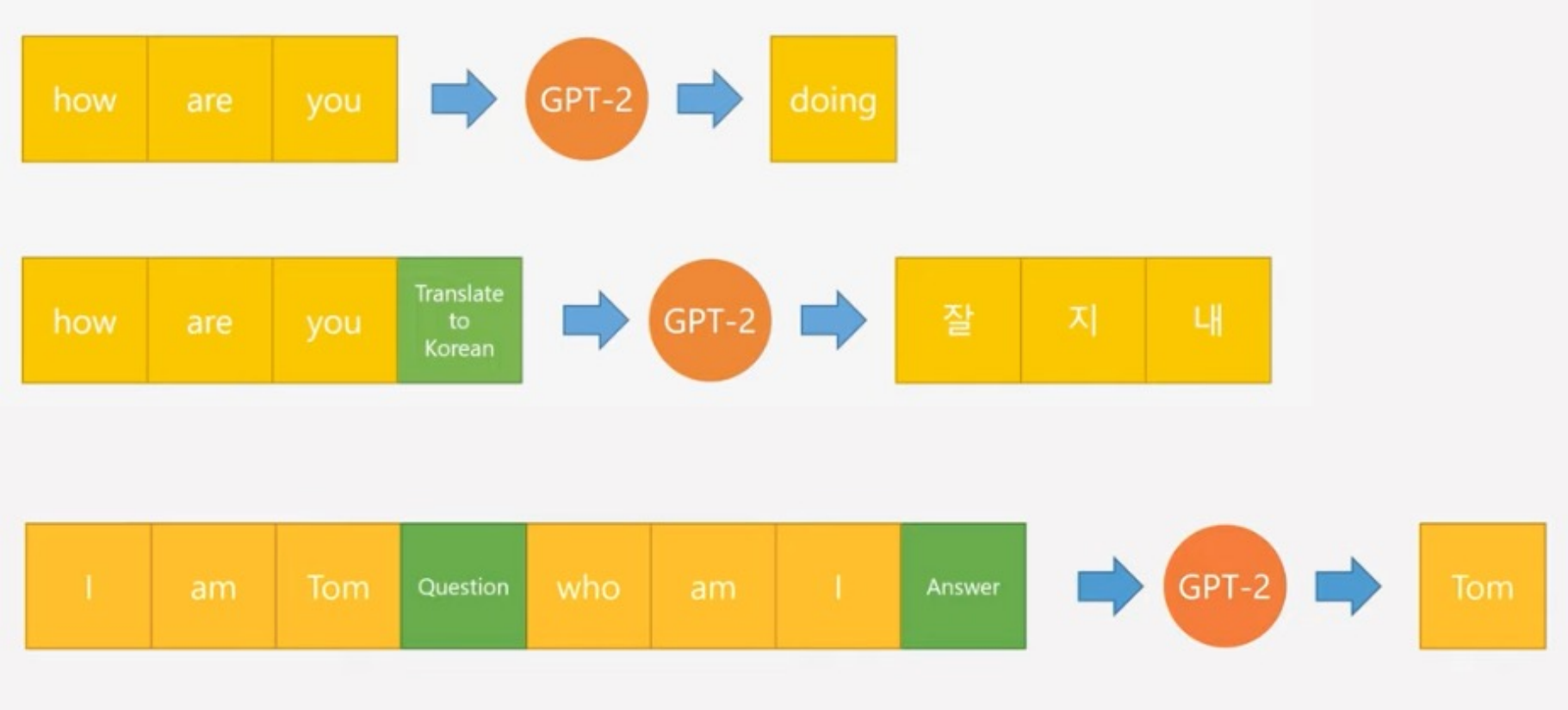
Approach

Language modeling

TRAIN AND TEST

Approach

Language modeling



TRAIN AND TEST

Approach

Training Dataset

기존 Dataset

➔ Wikipedia 등 한 영역에서만 가져온 데이터로 구성(General 하지 않음)

WebText Dataset

➔ 다양한 출처에서 가져오려고 함.

데이터의 품질 문제로 인해 사람에 의해 필터링된 글만 사용

게시글 중 3+

800만개 이상, 40GB

TRAIN AND TEST

Approach

Input Representation

이전에 등장했던 BPE 토크나이저를 사용.

➔ 맥락상의 이해를 높이기 위함.

유의미하지 않은 variation을 포함하여 Vocab의 크기가 증가하는 문제

➔ 문자 수준 이상의 병합을 막음. 최적의 공간활용

TRAIN AND TEST

Approach

Model

- 기존 GPT1 구조와 거의 동일
- Layer normalization이 각 sub-block의 input 부분으로 이동. 또한 추가적인 layer normalization이 self-attention block에 적용됨.
- 모델 깊이에 따른 residual path의 누적에 대한 초기화 방법이 변경

-> layer 수 n에 따라 residual 가중치에 1/루트N이 곱해짐

	GPT-1	GPT-2
parameters	117 millions	1.5 billions
layers	12	48
States dimension	768	1600
Context token size	512	1024
Batch size	64	512
etc		Layer normalization moved to the input of each sub block Layer normalization added after the final self attention block

TRAIN AND TEST

Experiments

Setup

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.

 # 크기가 각각 다른 4개로 실험.

TRAIN AND TEST

Experiments

Zero-shot 환경에서 SOTA 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

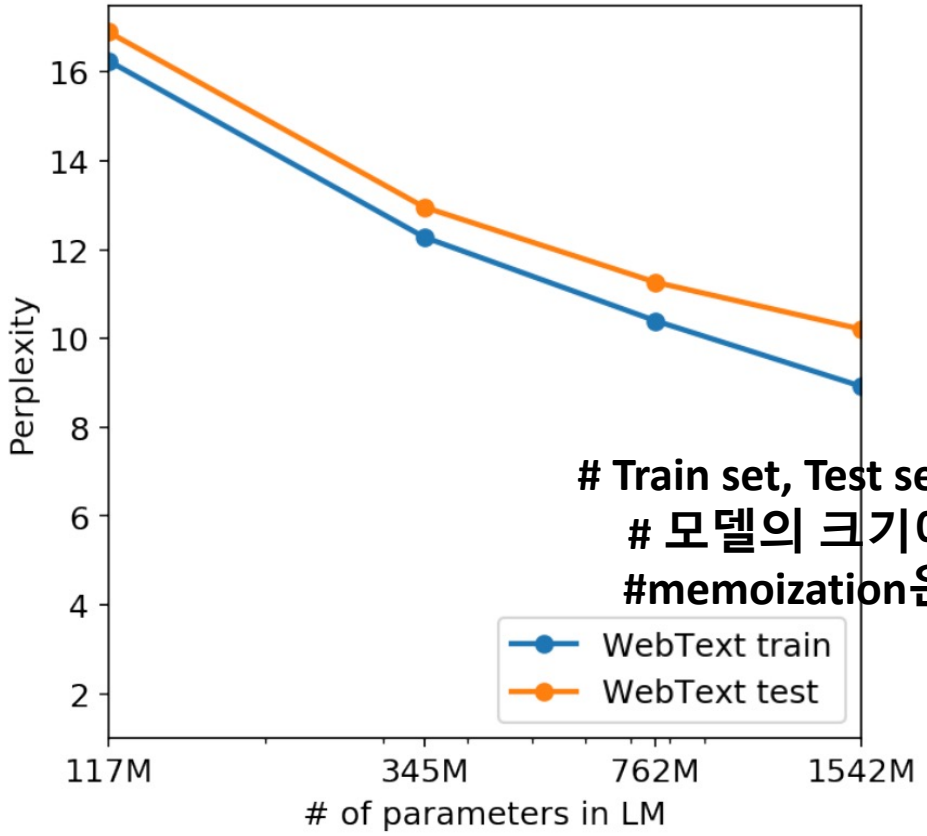
 # 여러 벤치마크에서 실험
 # 그 외 번역이나 QA에서도 높은 성능을 보임.

TRAIN AND TEST

Generalization vs Memorization

문장을 생성하는 것이 아닌, 기억하는 것이 아닌가?

-> Overlap의 가능성 문제



 # Train set, Test set 성능이 거의 비슷.
 # 모델의 크기에 성능이 증가함.
#memoization은 큰 문제가 아님.

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

TRAIN AND TEST

Discussion

남은 연구과제들

• Zero-shot 성능은 독해 등에서 좋은 성능을 보였으나 요약과 같은 문제에서는 기본적인 성능만을 보여줌

• GPT-2의 성능의 베이스라인은 확실하지만, fine-tuning을 통한 그 한계가 얼마인지는 분명하지 않음 -> 연구자들은 추가적인 미세조정을 통해 GLUE같은 벤치마크에 도전할 계획

• 또한 GPT-2의 학습 데이터와 그 크기가 BERT에서 말한 단방향 표현의 비효율성을 극복할 수 있을 만큼 충분한지도 확실하지 않음

TRAIN AND TEST

Conclusion

- LM이 충분히 크고 다양한 dataset에서 훈련되면 많은 도메인과 dataset에서 충분히 좋은 성능을 보인다.

- **GPT-2는 zero-shot임에도 불구하고 8개중 7개의 Task에서 SOTA를 달성했다.**

- **이는 다양한** test corpus**로 훈련된 고용량 모델이** zero-shot setting**에서 명시적임 감독 없이도 놀라운 양의 작업을 수행하는 방법을 배우기 시작함을 보여준다.**

TRAIN AND TEST

TRAIN AND TEST