GPT

GPT (Generative Pretrained Transformer)

- Generative - 생성 모델 : 데이터 전체의 분포를 모델링하는 머신 러닝 기법
- Pretrained - 굉장히 큰 학습 데이터의 양 : GPT3 모델은 5000억개의 단어(token) 데이터셋
- Transformer - GPT는 transformer의 Decoder 사용, BERT는 transformer의 Encoder를 사용

  → Decoder는 다음 단어를 예측, Encoder는 문장 전체를 보고 중간 단어 예측

SKT의 GPT2 데모

GPT 시리즈

- GPT1 - Improving Language Understanding by Generative Pre-training
- GPT2 - Language Models are Unsupervised Multitask Learners
- GPT3 - Language Models are Few-shot Learners

GPT1, 2, 3의 차이 → transformer의 크기와 학습 데이터의 양

GPT2 : 1.5B / GPT3 : 175B

GPT2

pretrain시 풀고자 하는 task를 input으로 넣는 In-context-learning 방법 사용

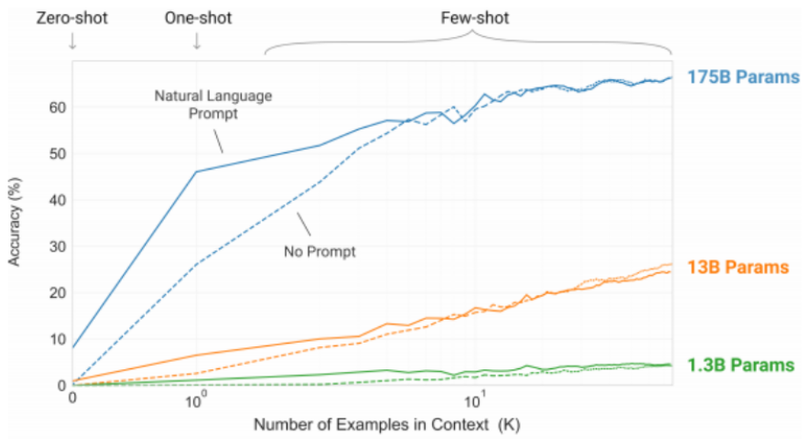
→ fine-tuning 방법의 성능에 미치지 못하는 아쉬움

but, 파라미터수가 늘어날 수록 downstream task에서의 성능 개선이 일어나기 때문에 In-context-learning 방법도 모델의 규모를 키우면 성능이 높아질 것은 기대

GPT3

24개 이상의 NLP task에 대한 GPT3평가 조건

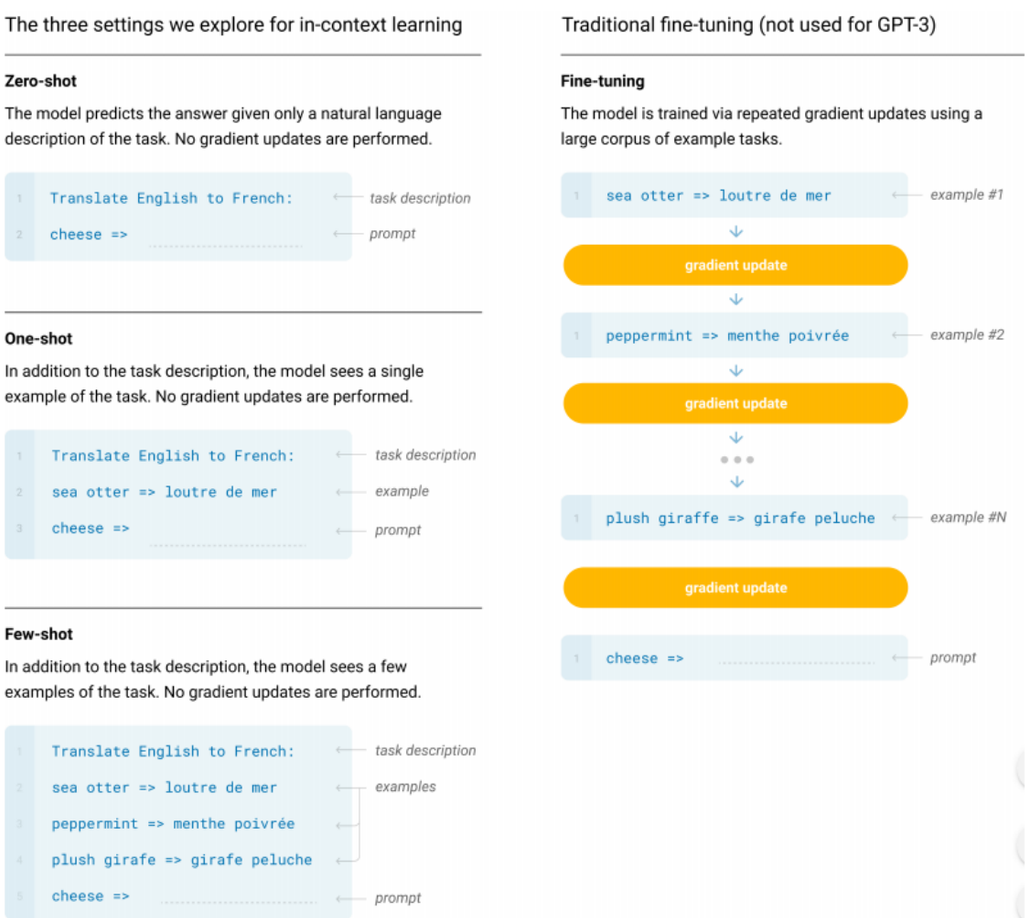
- Few-shot learning : 몇 개의 예제만으로 새로운 문제를 풀 수 있는지?
- Zero-shot learning : 예제를 주지 않고 바로 새로운 문제를 풀 수 있는지?
- One-shot learning : 하나의 예제만 허용



task에 대한 prompt가 주어지면 모델의 성능은 향상

주어지는 예제가 많아지면 모델의 성능 향상

In-context-learning 능력은 모델의 크기가 클수록 향상



1) Fine-tuning : pretrain된 모델의 weight를 downstream task에 대해 미세 조정

→ 매 task마다 라벨링된 데이터가 필요

2) Few-shot : 몇 개의 예제 task를 주어지고, weight업데이트는 일어나지 않음

→ 대부분의 few-shot성능은 fine-tuning을 따라가지 못함

3) One-shot : 하나의 예제 task

4) Zero-shot : task에 대한 예제는 주지 않고, task를 설명하는 자연어 문구만 주어짐

모델 아키텍쳐

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}

GPT2와 같으며 modified initialization, pre-normalization, reversable tokenization 적용

Transformer의 attention 레이어처럼 dense와 sparse attention을 교대로 사용

8개 스케일의 모델이 있으며, 모든 모델은 3000억 token에 대해 학습함