

# Language Models are Few-Shot Learners (GPT-3)

NIPS 2020

Tag: Few-shot learning

## 0. Summary

- Few-shot learning을 사용하여 fine-tuning 없이도 좋은 결과를 내게끔 학습한 모델

## 1. Background

### > Limitation of PLM (Pre-trained Language Model)

- Pre-training step에서 많은 지식을 흡수하지만, 그중 1개의 downstream task에만 fine-tune됨
- Downstream task를 풀 때마다 fine-tuning 필요
- 실제 인간은 학습을 위해 많은 양의 데이터를 필요로 하지 않음

### > GPT-2

- Encoder 없이 Transformer decoder를 사용한 모델
- Fine-tuning 과정을 없애고 pre-training만으로도 좋은 결과를 낼 수 있도록 학습

### > Few, One, Zero-shot, Learning

- **Few-shot Learning:** 적용하고자 하는 task에 대한 data k개를 모델에게 알려준 후 원하는 결과를 얻어내는 것. 가중치 업데이트는 되지 않음 (본 논문에서 k=10~100)

1	Translate English to French:	← task description
2	sea otter => loutre de mer	← examples
3	peppermint => menthe poivrée	←
4	plush girafe => girafe peluche	←
5	cheese => .....	← prompt

- **One-shot Learning:** 1가지 예시와 task에 대한 지시문(ex: Translate English to French)을 제공받음

1	Translate English to French:	← task description
2	sea otter => loutre de mer	← example
3	cheese => .....	← prompt

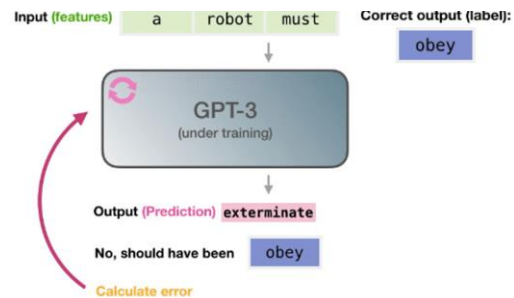
- **Zero-shot Learning:** task에 대한 지시문만 받음

1	Translate English to French:	← task description
2	cheese => .....	← prompt

## 2. GPT-3

- **Model:** GPT-2와 거의 같은 구조의 모델의 크기를 변화시키며 실험함 - 8가지(S, M, L, XL, ...)
- **Dataset:** 570GB의 corpus 사용  
→ Filtered Common Crawl, WebText2, Books1,2, Wikipedia

## Unsupervised Pre-training



## 3. Task

- Task마다 zero-shot, one-shot, few-shot learning에 대하여 검증함

### > 일부 Data에서만 SOTA 달성

- **Cloze and Completion task:** 문장 완성 or 알맞은 문장 고르는 task, SOTA 달성하지 못한 task에서도 기존 zero-shot의 SOTA를 넘어섬
- **Closed Book QA**
- **Translation:** Pre-training data 대부분이 영어, 7%가 타 언어 포함, Fr, De, Ro→EN에서 SOTA 달성
- **Common Sense Reasoning:** 기본 상식 질문, SOTA 달성한 data의 경우 pre-training data에 test data 포함 가능성이 있음.

### > SOTA에 준하는 성능 달성

- **Winograd-Style Tasks:** 대명사 찾는 task

### > SOTA보다 낮은 성능

- **Reading comprehension**
- **SuperGLUE:** 모든 항목 SOTA보다 낮음
- **NLI:** Natural Language Inference, 다음 문장 찾는 task

### > Synthetic & Qualitative Tasks

- 특이한 task에도 적용할 수 있다는 것을 보임
- Arithmetic, Word Scrambling & Manipulation, SAT analogy, News Article Generation, Learning and Using Novel Words, Correcting English Grammar
- 나름 우수한 성능을 보임

## 4. Limitation

- 여전히 일부 task에서 한계 보임
- 단방향 masked attention으로 인하여 성능 저하 가능성 존재 (다음 단어 예측 방식으로만 학습)
- 비용 많이 듦, 해석 가능성 떨어짐

📢 끝!!

질문 있으시면 정민지에게 DM 주세요!!