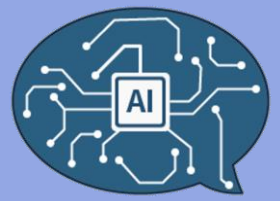
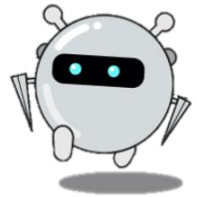


GPT



진행자 : 멘토 현시은

GPT란?

OUTTA

Motivation : 이전의 딥러닝 방식의 비효율성을 인식함

- 대부분의 딥러닝은 지도학습의 방식으로 이루어짐. 즉, 많은 수의 label data가 필요하다.
- 그러나 실제로 우리가 얻을 수 있는 데이터는 Unlabeled data가 압도적으로 많고, Data labeling 작업은 많은 시간과 비용이 필요하다.(인건비 등…)
- 적은 양의 label data로 학습을 진행할 수는 없을까?

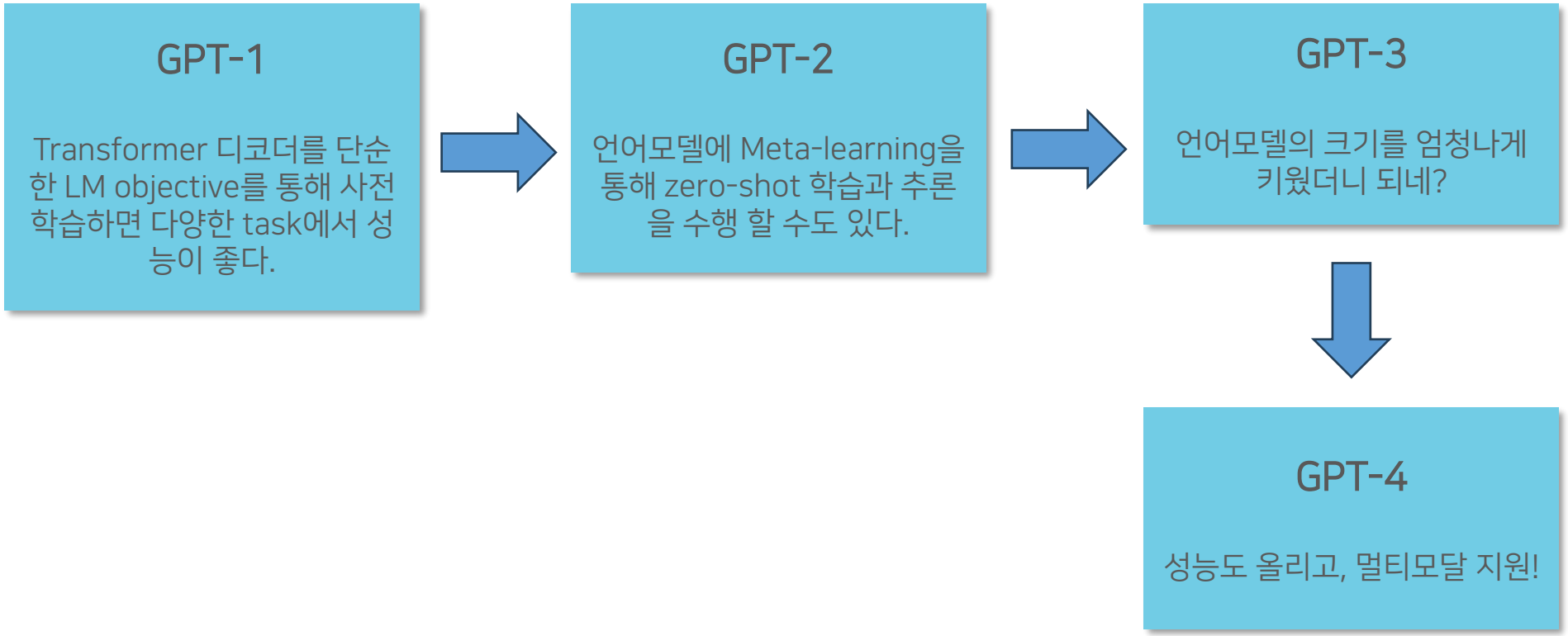
- **GPT란?**

GPT-1

GPT-2

GPT-3

GPT-4

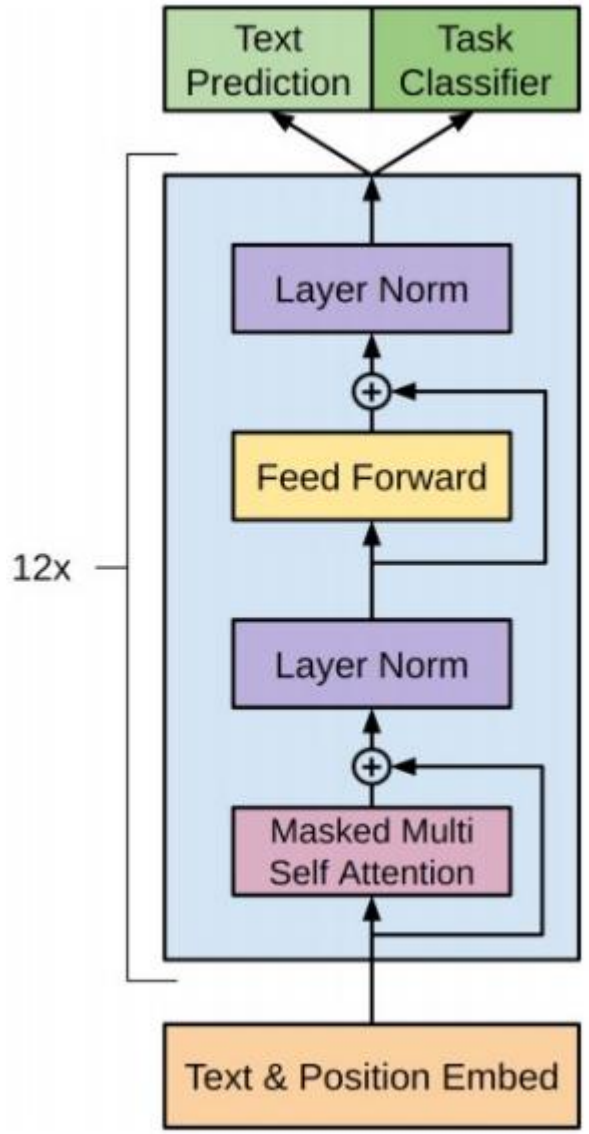


Motivation : 이전의 딥러닝 방식의 비효율성을 인식함

- Unlabeled data를 통해 Pre-train을 수행하고, 이를 바탕으로 각 Task에 맞게 Fine-tuning하여, 전이학습을 활용한 지도학습을 한다.
- GPT-1은, 어떠한 목적을 위한 Pre-train을 수행하는 것이 가장 효율적인 전이학습이 발생하는지 알기 위한 것

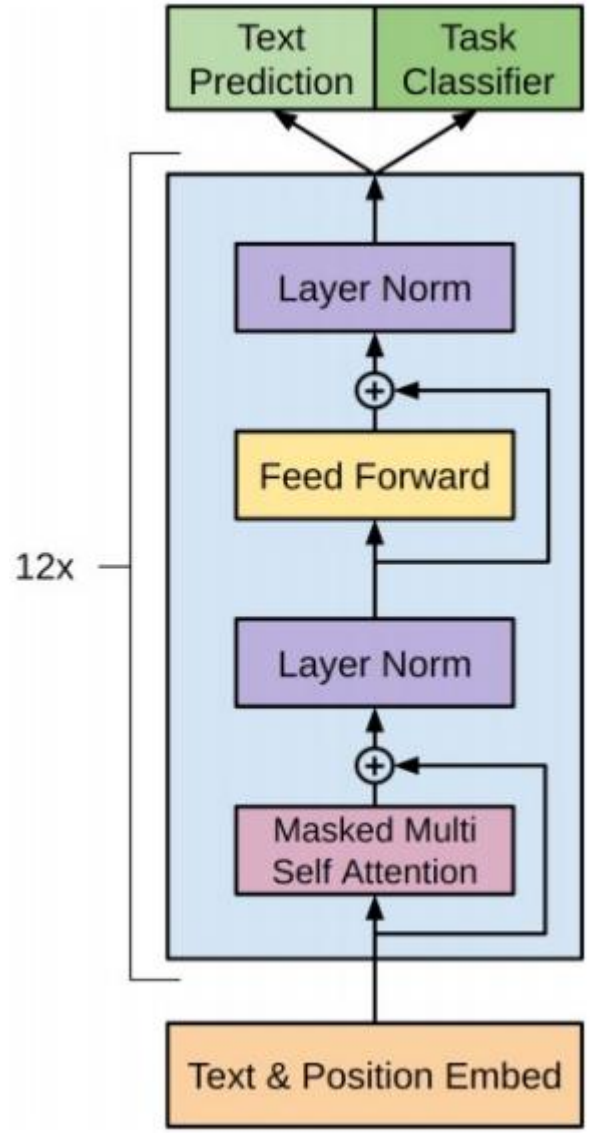
GPT(Generative Pre-trained Transformer) - 1

- Transformer의 디코더를 활용하여 Language Model을 Pre-train하고, 여러 목적에 따른 Task에 전이학습을 수행한다.
- GPT는 auto-regressive 언어 모델이다. 즉, 예측된 값이 다음 예측값을 예측하는 입력으로 사용된다.

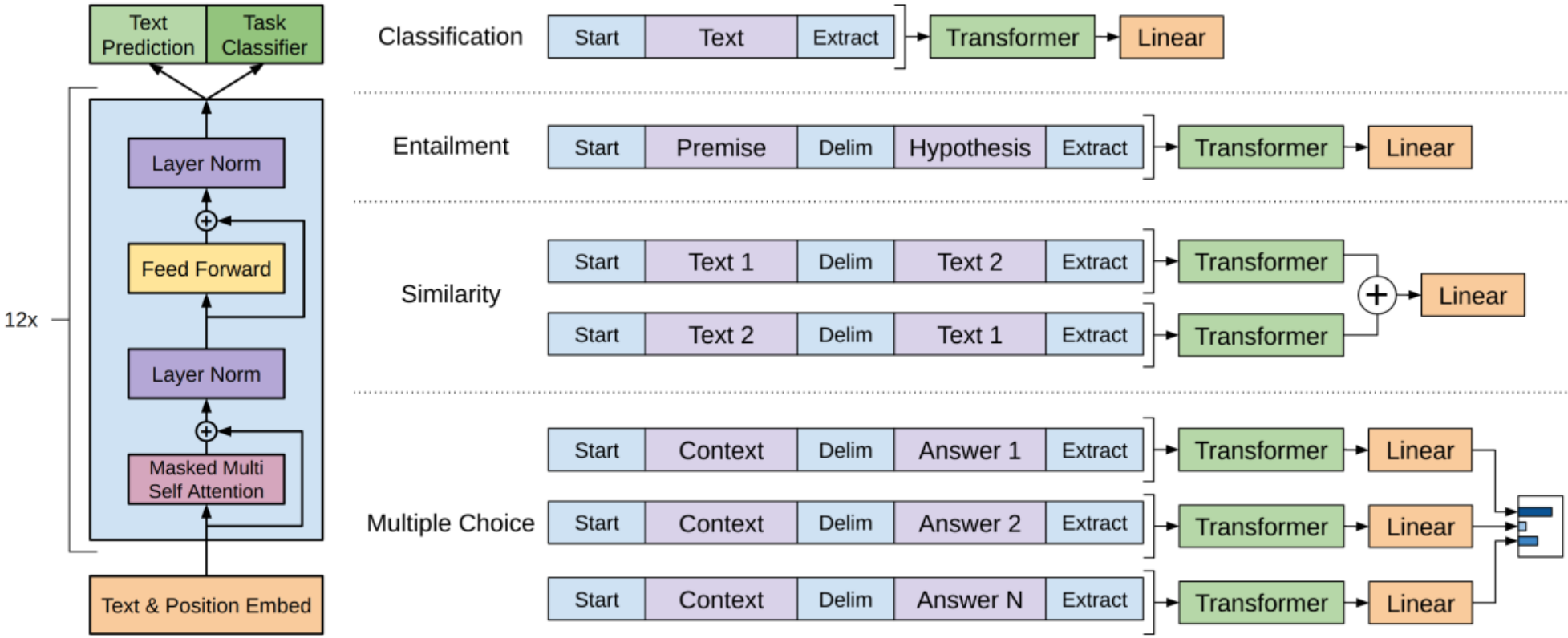


GPT(Generative Pre-trained Transformer) - 1

- Transformer의 디코더를 12개 쌓아 올려 예측한다. 디코더 block은 masked self-attention과 feed forward로 구성된다.
- 12개 각각의 디코더 block은 구조만 같고 그 안의 weight는 모두 다르다.
- Bidirection model이 아니다. 즉, 주어진 것 이전 정보까지만을 사용한다.



GPT(Generative Pre-trained Transformer) - 1



왼쪽 : pre-train된 GPT 아키텍처

오른쪽 : 각각의 다른 task에서 fine-tuning을 위한 입력 시퀀스의 구조

GPT- 1의 결론 및 의의

1. 초기 Pre-trained Language Model(PLM)으로써 이후 GPT 시리즈의 토대가 되었다.
2. RNN 구조를 탈피하고 transformer 구조를 사용하였다.
3. 단순한 objective로 큰 모델을 학습하고 transfer learning을 하였을 때 다양한 task에서 좋은 성능을 보였음

GPT- 1의 한계

특정 task를 위한 transfer-learning을 할 때 성능 향상을 위해 fine-tuning과정과 input transformation이 필요했다. 결국 fine-tuning을 위한 지도 학습이 필요하다.

GPT- 2의 등장 배경

1. GPT-1은 fine-tuning 과정에서 supervised-learning 이 필요로 하는 한계가 있다. 대부분의 다른 PLM들도 사전학습 이후에 fine-tuning 을 통해 task를 학습한다.

2. task를 위해서는 굉장히 많은 데이터가 필요로 한다.

3. 뿐만 아니라 fine-tuning 은 오버피팅과 일반화 성능의 하락, 실제 성능에 비한 과대 평가를 유발할 수 있다.

4. 또한, GPT-1 만으로는 각 downstream task마다 모델이 따로따로 만들어지므로 범용성이 부족하다.

5. 비지도 학습만으로 모델이 만들어지면 더욱 다양하고 범용적인 사용이 가능할 것이다.

6. **따라서, fine-tuning을 하지 않고 사전학습만으로 동작하는 모델 등장!**

GPT- 1과 GPT-2의 차이점

1. 모델의 구조

2. Training 데이터 수집 및 전처리 방식

GPT- 2의 구조 : GPT-1에서 크게 달라진 것이 없음!

1. Layer normalization의 위치가 변경됨

2. Residual layer가 많이 누적됨에 따라 초기화 방식이 변경되었다.

3. 더욱 많은 데이터를 사용하고 배치 사이즈도 늘렸다.

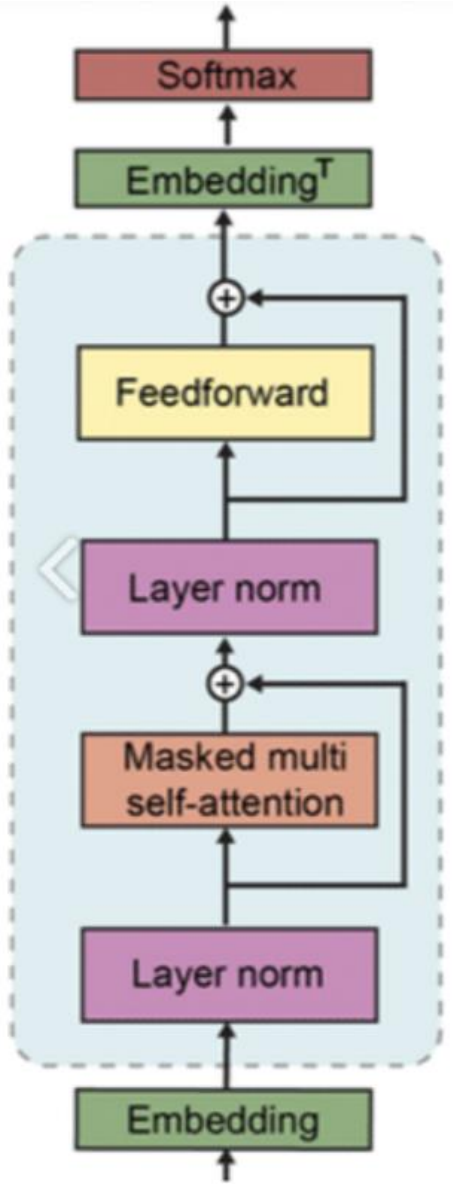


그림 출처 : https://velog.io/@tobigs-nlp/GPT2-Language-Models-are-Unsupervised-Multitask-Learners

GPT- 2의 목적

1. GPT-2의 가장 큰 목적은, Fine-tuning 없이 비지도 사전 학습만으로 zero-shot setting 에서 downstream task를 수행하는 것이다.

2. Zero-shot task 란? : 모델이 학습 과정에서 배우지 않은 task
ex) 셰익스피어처럼 글을 쓰도록 학습한 자연어 생성 모델이 마크 트웨인의 스타일로 글을 쓰는 것

3. Zero-shot setting 이란? : 작업에 대한 명시적인 지시 없이도 작업을 수행하는 것
ex) 번역 작업을 수행할 때, 특정한 예시나 데이터를 주지 않고 '다음의 영어를 한국어로 번역해줘.' 라는 프롬프트를 제공하여 작업을 수행하는 것

$$p(output|input) \Rightarrow p(output|input, task)$$

4. 따라서 GPT-2는 **Fine-tuning 을 하지 않는다!**

GPT- 2의 training dataset

1. 기존과 다르게 Web scraping 을 활용하여 다양한 도메인을 이용한 데이터를 사용했다.

2. 데이터의 품질이 떨어지는 것을 방지하기 위해 직접 필터링을 진행하여 만든 Webtext dataset을 활용하였다.(중복 제거, 대중성 및 다양성 고려)

3. 또한 데이터마다 가중치를 두면서 데이터들을 섞었다. 깔끔한 데이터셋이 실제 양에 비해 상대적으로 높은 빈도로 학습되도록 했다.

4. 뿐만 아니라, 데이터셋 크기 자체도 40GB가 넘는 굉장히 많은 Text를 사용하였다.

5. Byte Pair Encoding(BPE)를 사용하여 토큰화를 수행했다. 이를 통해 모델은 고정된 어휘 크기를 유지하면서 다양한 언어와 도메인에 걸친 데이터를 처리할 수 있게 되었다.

GPT- 2의 한계

Zero-shot(또는 few-shot) learning에 대해서 상용화 할 만큼의 성능이 나오지는 않았다.

GPT- 2의 가치

1. 비지도 학습의 영역의 가능성을 보여준 논문이다.
2. 크고 다양한 도메인과 데이터셋에서 잘 수행될 수 있음을 보여줬다.

GPT- 3의 시작

- 그럼 그냥 더 많은 데이터를 사용하고, 더 거대한 모델을 써보면 되지 않을까? ➡ **되네???**
(GPT-2 : 1.5B / GPT-3 : 175B)

- GPT-2의 Zero shot setting 에서 Downstream task 의 가능성을 보고 Few-shot setting 으로 발전시켜서 fine-tuning 없이도 좋은 성능을 낼 수 있도록 하였다.

- 즉, GPT-3는 fine-tuning 을 전혀 하지 않는다!!!! Pre-train된 모델을 그냥 바로 사용해도 성능이 잘 나온다.

- GPT-3의 모델 구조는 GPT-2의 모델 구조와 거의 유사하다. Attention 패턴과 몇 가지 하이퍼 파라미터나 데이터 전처리 과정만 살짝 다르다.

GPT- 3에서 알고 가면 좋은 용어들

1. in-context learning : pre-trained 된 모델에 풀고자 하는 task를 text input으로 함께 넣어 주는 방식이다. 파라미터 업데이트 없이 feed-forward 를 통해 학습이 이루어지는 것.

• few-shot learning : 10~ 100개 정도 예시를 사용
• one-shot learning : 단 1개의 예시 사용.
• zero-shot learning : 단 하나의 예시도 없이 언어 모델을 바로 NLP task에 테스트

2. meta-learning : 일반화를 향상시키기 위해 관련 task에 걸쳐서 학습 알고리즘을 조정하는 것. 모델이 학습하는 과정에서 스킬과 패턴을 인지하는 능력을 개발한다.

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée        ←
plush girafe => girafe peluche      ←
cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← example
cheese =>                           ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
cheese =>                           ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer          ← example #1
```

gradient update

```
peppermint => menthe poivrée        ← example #2
```

gradient update

...

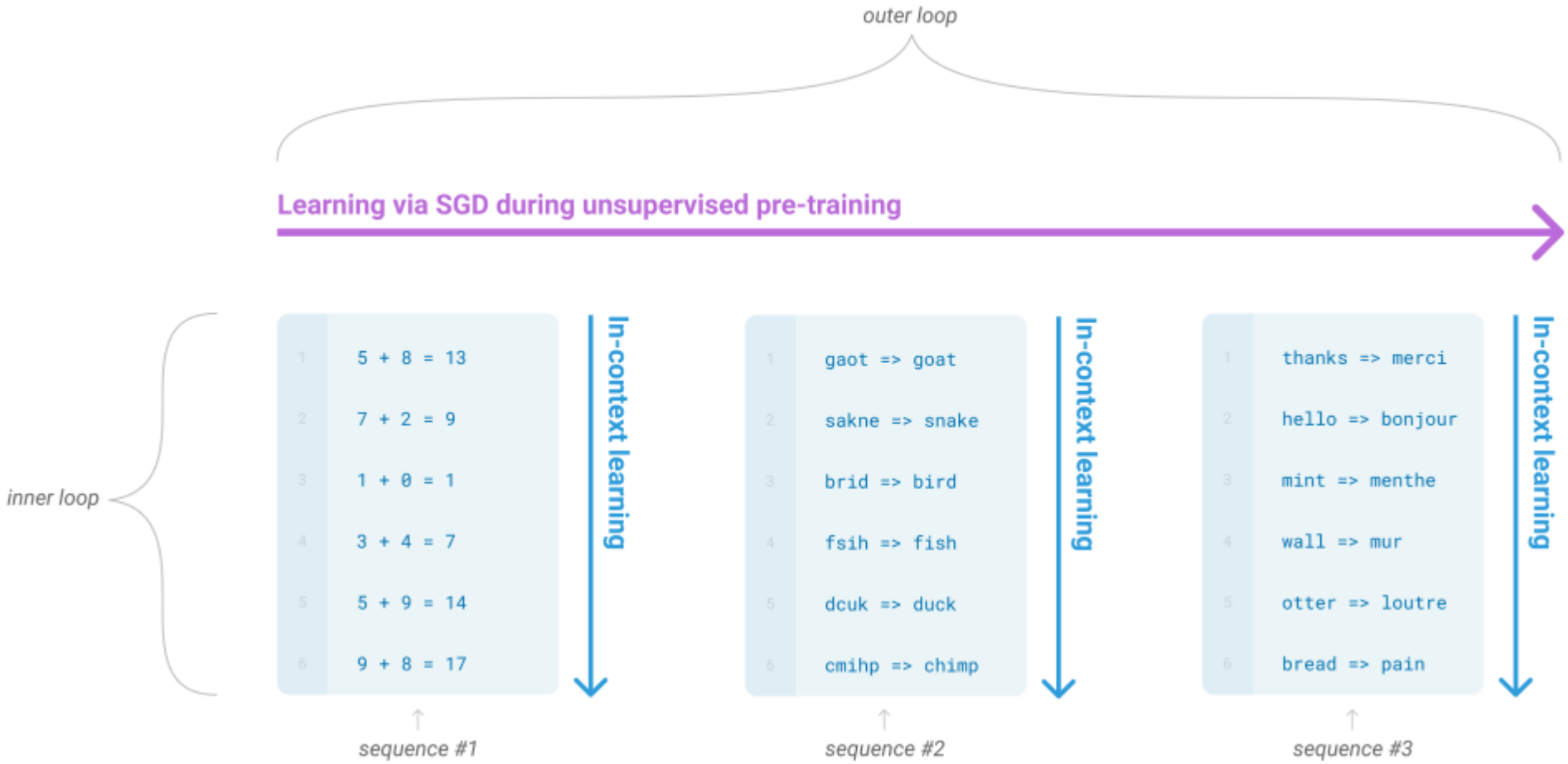
```
plush giraffe => girafe peluche     ← example #N
```

gradient update

```
cheese =>                           ← prompt
```

In-context learning

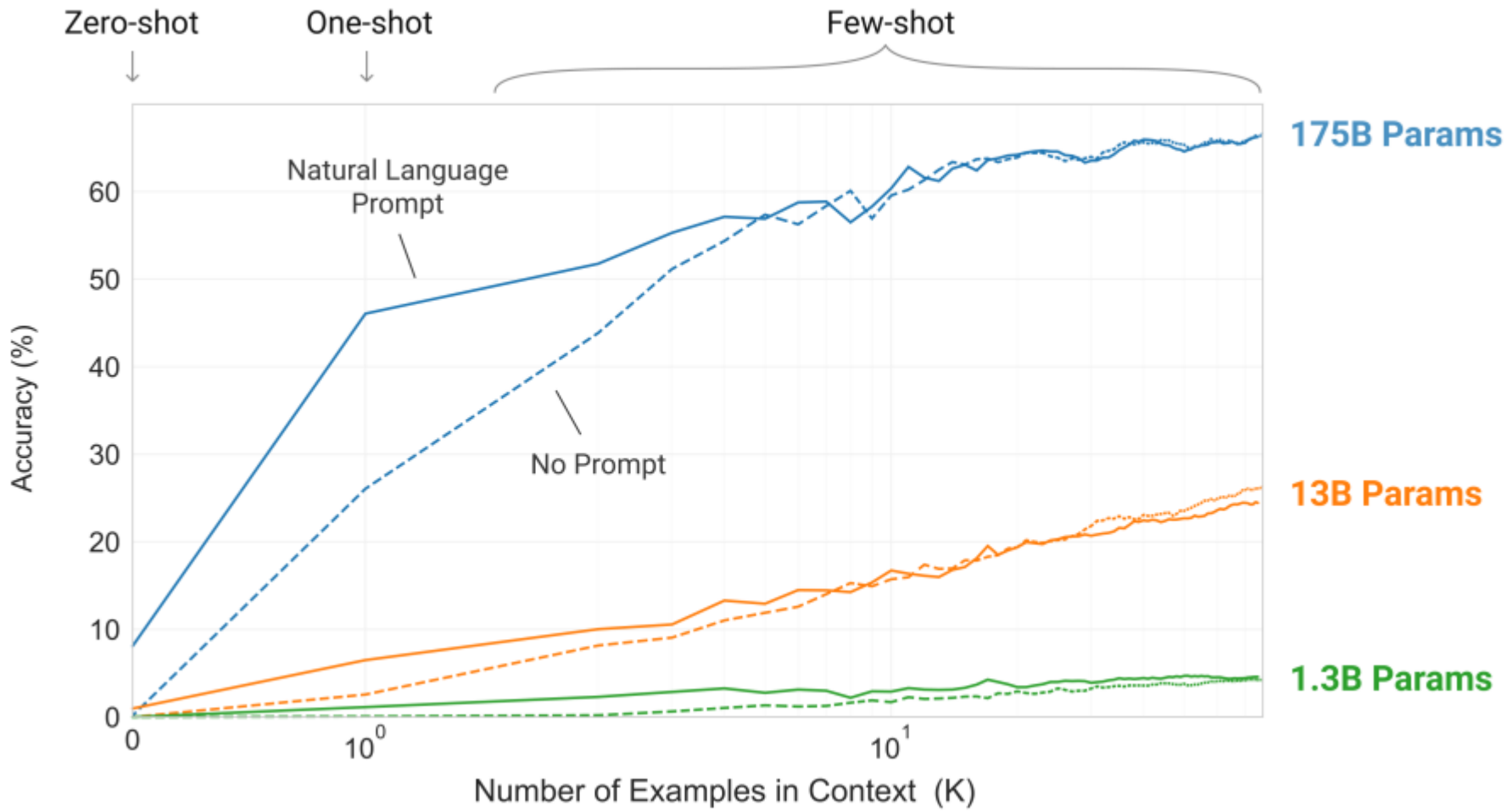
만약 언어모델이 잘 학습되었다면, 주어진 context에 기반하여 나머지 알맞은 문장을 완성하는 방향으로 다음 단어를 예측해 나갈 것이다.



OUTTA

In-context learning의 효율성

모델이 커질수록 In-context learning 의 효율도 높아지며 zero-shot learning 의 효율도 커진다.



GPT-3의 성과

1. 번역, Q&A, 단어 순서 맞추기 등 여러 task에서 뛰어난 성능을 보였다.

2. Text 생성에서 굉장한 성능을 보였다.

이것 외에도 많지만 생략…

GPT-3의 한계

1. 여전히 잘 못푸는 task도 존재한다. 특히 생성에 있어서 동어 반복 현상 등이 발생한다.

2. 양방향적인(bidirectional) 구조는 고려하지 않음. 이로 인해 빈칸 채우기, 긴 문단을 읽고 짧은 답변 생성하는 task에서 성능이 낮게 나왔을 수 있다.

3. 사전 학습에서 너무 많은 데이터가 필요하다.

4. 너무 모델 크기가 크다. 그로 인한 비용이 너무 많이 든다.

5. 해석이 어렵다. 따라서 이 성능이 정말로 few-shot learning에 의해 나오는 것인지 알기 어렵다.

6. 과연 이 모델이 정말 이 지식들을 다 알고 있는지 알 수 없다. 글만 보고 세상을 배운 것이라서 부정확한 정보가 많다.

GPT-3의 의의

1. 파라미터 크기를 175B까지 극대화하여 fine-tuning 없이도 사용이 가능하도록 하였다. 따라서 downstream task에 따라 모델을 각각 만들 필요 없이 범용적으로 사용 가능하다.

2. In-context learning, 그 중에서 Few shot learning 으로 적은 수의 샘플로도 준수한 성능 확보

그럼 chatGPT 는 뭐야?

GPT 모델을 챗봇 형식으로 최적화한 것이며, 이 과정에서 RLHF라는 기법을 추가로 도입했다.
RLHF는 Reinforcement Learning from Human Feedback이라는 기법으로 사람이 직접 피드백을 주는 방식으로 언어모델을 최적화하는 기법이다.
해당 기법이 도입되면서 모델의 성능이 올라갔고 이를 ChatGPT라는 형식으로 OpenAI가 선보였다.

GPT-4

2023년 3월 15일 OpenAI에서 3년 만에 발표한 것이다.

GPT-4에서는 GPT-3와 다르게, 모델의 아키텍처, 사이즈, 데이터셋 종류 및 규모, 학습방법, 학습 시스템 등에 대한 세부정보를 공개하지 않았다!

그러면 이 논문에서 뭘 알 수 있는데?

GPT-4

논문에는 그냥 이전 GPT보다 성능이 잘 나왔고, 어떤 실험을 해서 어떤 실험 결과가 나왔는지에 대한 내용이 대부분이다. 또한 GPT-4의 제한사항(GPT 이전 버전과 동일한 문제)와 그에 대한 위험 완화 방법의 간단한 가이드라인 정도가 제시된다.

GPT-4의 특징

1. Multi-Modal 지원
: 텍스트만으로는 세상을 이해하기 어렵다. 따라서 텍스트와 이미지를 입력할 수 있게 되었다.

2. 입력 context의 길이가 증가했다. 이로 인해 긴 리포트나 장편 소설도 생성 가능할 것으로 예상됨

3. 더 창의적이고 강력하다. 광범위한 지식과 어려운 문제를 더 잘 해결한다.

GPT-4o

1. GPT-4o는 텍스트, 음성, 비전(이미지) 기능을 통합하여 멀티모달 AI로서의 성능을 크게 향상

2. GPT-4o는 GPT-4와 유사한 수준의 텍스트 이해 및 생성 능력을 가지고 있지만 출력 속도가 훨씬 빠르다.

3. GPT-4o는 더 많은 언어를 지원하며, 사용자 설정, 로그인, 파일 업로드 등 다양한 기능을 통해 사용자 경험을 향상시켰다.

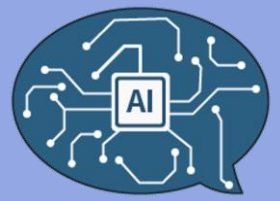
4. GPT-4o는 안전성이 향상되어 위험한 정보나 개인정보를 출력하지 않는 기능이 더 강화되었다.

더 다양한 Large Language Models

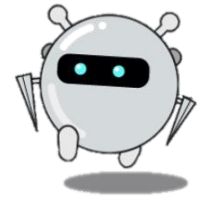
Llama
- 제작 회사 : Meta AI
- 현재 Llama 3까지 개발되었음
- 파라미터 개수 : Llama 3 기준 8B, 70B
- 특이사항 : **오픈소스**로 공개되었음.

Gemini
- 제작 회사 : Google Deepmind
- 현재 Gemini 1.5까지 개발되었음
- 파라미터 개수 : 밝혀지지 않음
- 특이사항 : 갤럭시 S24 시리즈에 일부 기능이 탑재. 나노형 모델이 온디바이스로 내장됨.

Phi
- 제작 회사 : Microsoft
- 현재 Phi 3까지 개발되었음
- 파라미터 개수 : 3.8B, 7B, 14B
- 특이사항 1 : 작은 사이즈임에도 좋은 성능이 나온다고 주장. 온디바이스 LLM을 타겟하는 모델
- 특이사항 2 : **오픈소스**로 공개되었음.



감사합니다



OUTTA

Reference

GPT1 : Improving Language Understanding by Generative Pre-Training
GPT2 : Language Models are Unsupervised Multitask Learners
GPT3 : Language Models are Few-Shot Learners
GPT4 : GPT-4 Technical Report