Transformer 기반 후속 논문 리뷰
: "Training Language Models to Follow Instructions with Human Feedback"

Introduction, GPT-3의 근본적인 문제점

GPT-3은 혁신적인 언어 모델이지만, 여전히 몇 가지 근본적인 문제점을 안고 있다. 이 모델은 때로는 사실과 상관없는 내용을 지어내거나, 편향적이거나 유해한 텍스트를 생성하며, 사용자의 지시를 제대로 따르지 않는 경우가 발생한다. 이러한 문제의 근본적인 원인은 주어진 텍스트 시퀀스를 바탕으로 다음에 올 토큰을 올바르게 예측하는 데 있어 사용자의 의도와 안전성을 고려하지 않아 발생하는 것으로, 이것을 "misaligned"라고 한다.

본 논문의 목표는 이러한 문제를 해결하고, Language Model이 사용자의 의도를 정확하게 이해하고 그에 따라 행동하도록 하는 것이다. 여기서 사용자의 의도는 명시적으로 사용자의 지시를 정확히 따르는 것과 함께, 암묵적으로 편향이나 해로운 내용을 제공하지 않으며, 사용자에게 안전하고 유용한 답변을 제공하는 것을 의미한다. 이러한 의도를 충족시키기 위해 본 논문은 세 가지 중요한 기준을 제시한다. 첫 번째로, Language Model은 사용자의 문제를 해결하는 데 도움이 되어야 한다 (도움 측면). 두 번째로, 정보를 조작하거나 사용자에게 부적절한 정보를 제공해서는 안 되며, 정직해야 한다 (정직 측면). 마지막으로, Language Model은 개인, 사회, 문화적인 측면에서 신체적, 심리적, 또는 사회적 피해를 최소화해야 한다 (무해 측면).

Related work, GPT-3를 사용자의 입맛에 맞게 Fine-tuning 하기

GPT-3 모델을 사용자의 개별 요구에 더 잘 부합하도록 Fine-tuning하는 연구는 인간의 피드백을 기반으로 한 강화학습(Reinforcement Learning with Human Feedback)을 통해 진행된다. 이러한 연구 방법은 GPT-3가 다양한 사용자의 지시사항을 따를 수 있도록 조정하는데 중요한 역할을 한다. 이를 위해, 인간의 평가(선호도)를 reward로 활용하여 모델을 향상시킨다.

Fine-tuning을 위한 dataset을 구축하기 위해 40명의 labeler들이 고용되었고, 이 labeler들은 screening test(다양한 인구통계학적 그룹의 선호도에 얼마나 민감한지, 잠재적으로 유해할 수 있는 결과물을 식별하는 능력 평가)에서 상위에 랭크하신 분들로 선정되었다. 여기서 사용자들이 연구에 참여한 저자들과 labeler들로 한정되어 있어 일반적인 사용자 다양성을 반영하지 못할 수 있다는 한계점이 존재한다. 이러한 한계점을 고려하여 미래 연구에서는 보다 다양한 사용자 그룹을 대상으로 한 데이터 수집 및 Fine-tuning 접근법을 탐구할 필요가 있다.

Fine-tuning의 3가지 절차

1단계, Demonstration 데이터 구축 및 Supervised Fine-Tuning

저자들은 먼저 labeler에게 세 가지 유형(Plain, Few-shot, User-based)의 prompt 작성을 요청한다. Plain은 task가 충분히 다양성을 갖도록 하면서 임의의 task를 제시하도록, Few-shot은 query/response 쌍으로 제시하도록, User-based는 OpenAI API의 여러 use-case에 해당하는 prompt를 제시하도록 요청한다. 아래는 API prompt의 use-case 카테고리의 분포(왼쪽)와 예시 prompt(오른쪽)이다.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

Labeler가 작성한 prompt를 활용하여 SFT dataset을 생성한 후, 사전 학습된 GPT-3 모델을 이 데이터로 fine-tuning한다. 이때, 1.3B, 6B, 175B 버전의 모델을 각각 사용하며, 16 epoch 동안 학습을 진행한다. 학습 과정에서는 Cosine learning rate decay와 residual dropout 0.2가 적용되며, 최종 SFT 모델은 Validation dataset에 대한 Reward Model score를 기준으로 선정된다.

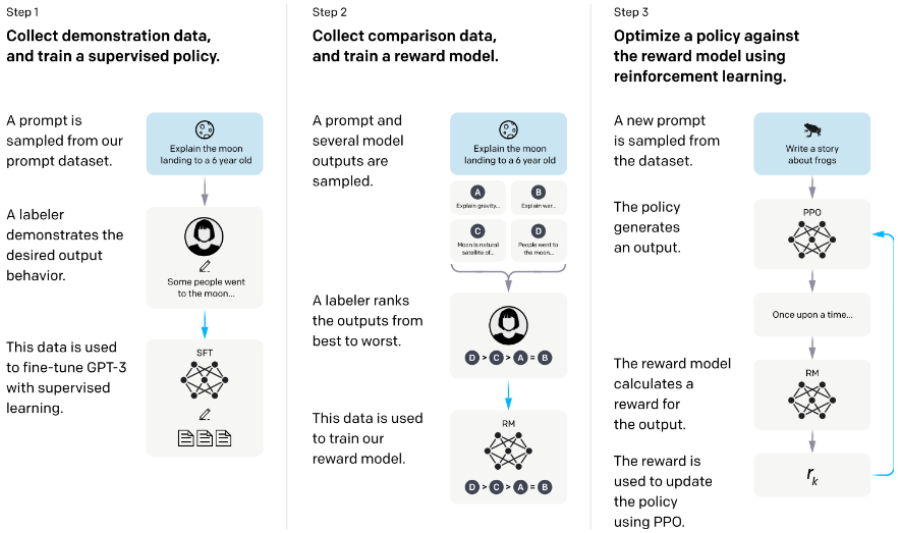
2단계, 인간의 선호도를 반영한 Comparison 데이터 구축 및 Reward Model 학습

두 번째 단계에서는 입력에 대한 인간 labeler의 선호도를 반영하는 Comparison dataset을 생성한다. 이 dataset은 모델의 출력에 대한 labeler의 선호를 포함하며, Reward Model을 학습하기 위한 입력으로 사용된다. Reward Model은 이러한 인간의 선호를 예측하는 데 사용된다.

3단계, Reward Model을 활용하여 PPO 기법으로 GPT-3 fine-tuning

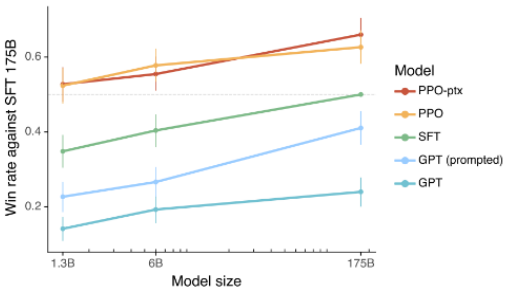
마지막 단계에서는 Reward Model을 활용하여 Proximal Policy Optimization (PPO) 알고리즘을 바탕으로 GPT-3 모델을 fine-tuning한다. 이 단계에서는 PPO 데이터셋을 생성하며, Reward Model의 출력이 스칼라 reward로 사용된다. PPO 알고리즘은 supervised policy를 fine-tuning하여 Reward Model의 출력을 최적화한다. 이러한 프로세스를 반복하여 최종 모델을 구축하며, 이 모델이 InstructGPT라고 불린다.

2단계와 3단계는 필요에 따라 반복될 수 있다.

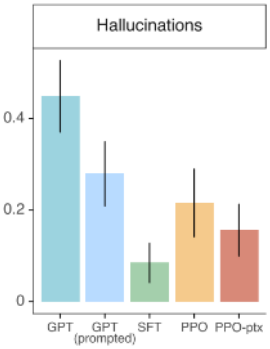


실험 평가 방법 및 결과

이 연구에서는 Language Model이 사용자의 의도를 얼마나 잘 이해하고 실행하는지를 평가하기 위해 다음과 같은 주요 기준으로 평가를 진행하였다. 도움측면에서는 labeler가 전적으로 판단하게 되고, 정직측면에서는 TruthfulQA와 closed-domain 결과로 평가하게 되고, 무해측면에서는 독성 측정용 RealToxicityPrompts dataset을 사용하여 평가하게 된다. 몇 가지 결과들을 아래 나열해보면,



- 오른쪽은 API prompt 분포에서 다양한 모델에 대한 인간 평가 결과로, 175B SFT model와 비교할 때 얼마나 선호되는 지를 측정한 것. Labeler들은 GPT보다 InstructGPT의 결과를 훨씬 선호했음을 직관적으로 살펴볼 수 있음. 또한, 우리의 1.3B PPO-ptx 모델의 출력이 175B GPT-3 모델의 출력보다 약 100배 이상 적은 파라미터를 가지고 있음에도 불구하고 선호된다는 점도 주목할만한 부분.



- GPT-3에 비해 Truthfulness가 개선됨.
- GPT-3에 비해 Toxity가 약간 개선되었지만 bias는 개선되지 않음. 여기서 bias에 대한 한계가 존재함을 파악할 수 있음.
- RLHF fine-tuning 절차를 수정하여 public NLP dataset의 성능 회귀를 최소화.
- InstructGPT는 학습 데이터를 생성하지 않은 hold-out labeler의 선호도에 대해 일반화.

- Public NLP datasets은 언어 모델이 사용되는 방식을 반영하지 않음.
- Fine-tuning 과정에서 거의 학습되지 않은 지시사항에도 잘 답변함.
- 하지만 여전히 InstructGPT는 완벽하지 않고, 실수를 범할 수 있음.

전반적으로, 이 논문의 결과는 인간의 선호도를 활용하여 대규모 언어 모델을 fine-tuning함으로써 모델이 다양한 작업에서 획기적으로 개선될 수 있음을 보여주고 있다. 그러나 이러한 개선에도 불구하고, 모델의 safety과 reliability을 높이기 위해서는 더 많은 연구와 개발이 필요해 보인다.

결론 및 향후 연구 방향

이 연구는 아직 몇 가지 간단한 실수를 포함하고 있지만, 인간의 피드백을 활용한 미세 조정이 언어 모델을 인간의 의도와 조화롭게 만드는 유망한 방향임을 입증하였다. 이러한 연구 결과는 인간 피드백을 기반으로 한 언어 모델의 개선 가능성을 강조하며, 향후 연구와 개발을 지속적으로 진행할 필요가 있음을 시사한다.

최근 ChatGPT와 같은 모델이 인공지능 시장에서 게임체인저로 등장하면서 굉장히 많은 관심이 이어지고 있는데, 이는 제로베이스에서 등장한 게 아닌 InstructGPT를 계승한 모델이라고 할 수 있다. 그렇기에 InstructGPT의 주요 특징들을 더 자세히 조사하고 평가하는 것은 중요한 의미를 가지고 있다. 앞으로도 OpenAI와 같은 연구기관이 개발하는 대규모 언어 모델의 발전 흐름을 계속 관찰하고 모니터링하는 것은 학계 및 업계에서 중요한 이슈가 될 것으로 예상된다. 이에 따른 윤리적, 안전성, 및 신뢰성 문제에 대한 연구가 더욱 필요하다는 점을 염두에 두어야 한다.

[References]

· https://arxiv.org/abs/2203.02155
· https://www.youtube.com/watch?v=vx3hxa9Bi5c
· https://kimjy99.github.io
· https://arxiv.org/abs/1909.08593
· https://arxiv.org/abs/2009.01325