Chain of Thought

https://arxiv.org/pdf/2201.11903

1. 서론: 사고의 사슬(Chain-of-Thought) 프롬프트의 혁신적 잠재력

최근 언어 모델(LLMs)의 규모 확장은 성능 향상과 샘플 효율성 증가와 같은 다양한 이점을 가져왔습니다. 그러나 모델 규모만으로는 산술, 상식, 기호 추론과 같은 복잡한 작업에서 높은 성능을 달성하는 데 충분하지 않았습니다. 본 연구는 "사고의 사슬 프롬프트(chain-of-thought prompting)"라는 간단한 방법을 통해 대규모 언어 모델의 추론 능력을 '잠금 해제'하는 방법을 탐구합니다.

"사고의 사슬"은 **"최종 출력으로 이어지는 일련의 중간 자연어 추론 단계"**를 의미합니다. 이 방법은 몇 가지 사고의 사슬 데모를 프롬프트의 예시로 제공함으로써, 모델이 복잡한 추론 작업을 수행하는 능력을 크게 향상시킵니다. 이 접근 방식의 핵심은 인간이 복잡한 문제를 해결할 때 문제를 중간 단계로 분해하고 각 단계를 해결한 다음 최종 답변을 제공하는 사고 과정과 유사하게 작동하도록 언어 모델을 돕는 데 있습니다.

이 연구는 사고의 사슬 프롬프트가 산술, 상식 및 기호 추론 벤치마크에서 표준 프롬프트보다 뛰어난 성능을 보임을 경험적으로 입증하며, 때로는 **"놀라운(striking) 정도"**의 개선을 이룹니다. 특히, PaLM 540B 모델에 단 8개의 사고의 사슬 예시를 제공하는 것만으로도 수학 단어 문제 벤치마크인 GSM8K에서 미세 조정된 GPT-3를 능가하는 최첨단 정확도를 달성했습니다. 이는 대규모 훈련 데이터 세트가 필요하지 않고 단일 모델 체크포인트가 일반성을 잃지 않고 많은 작업을 수행할 수 있다는 점에서 중요한 "프롬프트만 사용하는 접근 방식"의 중요성을 강조합니다.

2. 사고의 사슬 프롬프트의 핵심 특징 및 이점

본 연구에서 제시하는 사고의 사슬 프롬프트는 LLMs의 추론 능력을 향상시키는 데 있어 몇 가지 매력적인 특성을 가집니다.

- **다단계 문제 분해:** "사고의 사슬은 원칙적으로 모델이 다단계 문제를 중간 단계로 분해할 수 있도록 하며, 이는 더 많은 추론 단계가 필요한 문제에 추가 계산을 할당할 수 있음을 의미합니다." 이는 모델이 복잡한 문제를 한 번에 해결하려 하기보다, 작은 논리적 단위로 나누어 처리하게 합니다.
- **모델 행동의 해석 가능성:** "사고의 사슬은 모델의 행동에 대한 해석 가능한 창을 제공하여, 모델이 특정 답변에 어떻게 도달했는지 제안하고 추론 경로가 잘못된 부분을 디버깅할 기회를 제공합니다." 이는 LLMs의 '블랙박스' 문제를 일부 완화하고, 개발자나 사용자에게 모델의 의사결정 과정을 이해할 수 있는 통찰력을 제공합니다.

- **광범위한 적용 가능성:** "사고의 사슬 추론은 수학 단어 문제, 상식 추론, 기호 조작과 같은 작업에 사용될 수 있으며, 잠재적으로 (적어도 원칙적으로) 인간이 언어를 통해 해결할 수 있는 모든 작업에 적용 가능합니다." 이는 이 방법이 특정 도메인에 국한되지 않고 다양한 NLP 작업에 유용할 수 있음을 시사합니다.
- **쉽게 유도 가능:** "마지막으로, 사고의 사슬 추론은 충분히 큰 상용 언어 모델에서 몇 발짜리 프롬프트의 예시에 사고의 사슬 시퀀스 예제를 포함함으로써 쉽게 유도될 수 있습니다." 모델을 미세 조정하거나 새로운 아키텍처를 구축할 필요 없이, 간단한 프롬프트 변경만으로 강력한 추론 능력을 활성화할 수 있다는 점이 큰 장점입니다.

3. 주요 실험 결과

이 연구는 산술 추론, 상식 추론, 기호 추론의 세 가지 주요 영역에서 사고의 사슬 프롬프트의 효과를 평가했습니다.

3.1. 산술 추론 (Arithmetic Reasoning)

수학 단어 문제(Math Word Problems)는 인간에게는 쉽지만 언어 모델에게는 종종 어려운 작업으로 알려져 있습니다. 사고의 사슬 프롬프트는 이 분야에서 놀라운 성능 향상을 보였습니다.

- **PaLM 540B의 최첨단 성능:** "PaLM 540B에 사고의 사슬 프롬프트를 사용하면 GSM8K, SVAMP, MAWPS에서 새로운 최첨단 성능을 달성합니다." 이는 미세 조정된 모델과 비교해도 우월한 결과입니다.
- **모델 규모의 중요성:** "사고의 사슬 프롬프트는 모델 규모의 '창발적 능력(emergent ability)'입니다." 즉, 작은 모델(약 100B 매개변수 미만)에서는 성능에 긍정적인 영향을 미치지 않거나 오히려 저해할 수 있으며, 충분히 큰 모델에서만 성능 향상이 나타납니다.
- **복잡한 문제에서의 더 큰 이득:** "사고의 사슬 프롬프트는 더 복잡한 문제에서 더 큰 성능 이득을 보입니다." GSM8K와 같이 난이도가 높은 데이터셋에서는 가장 큰 GPT 및 PaLM 모델의 성능이 두 배 이상 증가했습니다. 반면, 단일 단계로 해결되는 MAWPS의 SingleOp와 같은 쉬운 데이터셋에서는 성능 향상이 미미했습니다.

3.2. 상식 추론 (Commonsense Reasoning)

언어 기반의 사고의 사슬은 광범위한 상식 추론 문제에도 적용 가능합니다.

- **다양한 상식 데이터셋에서의 개선:** CSQA, StrategyQA, Date Understanding, Sports Understanding, SayCan 등 다양한 상식 추론 데이터셋에서 사고의 사슬 프롬프트는 성능 향상을 가져왔습니다.
- **PaLM 540B의 우수한 성능:** "사고의 사슬 프롬프트와 함께 PaLM 540B는 StrategyQA에서 이전 최첨단 성능(75.6% 대 69.4%)을 능가하고, 스포츠 이해도에

서 보조를 받지 않은 스포츠 애호가(95.4% 대 84%)를 능가하는 강력한 성능을 달성했습니다."

3.3. 기호 추론 (Symbolic Reasoning)

기호 추론은 인간에게는 간단하지만 언어 모델에게는 어려울 수 있는 작업입니다. 사고의 사슬 프롬프트는 이러한 작업의 일반화 능력을 향상시킵니다.

- **길이 일반화 촉진:** "사고의 사슬 프롬프트는 두 가지 기호 추론 작업에서 더 긴 시퀀스로의 일반화를 촉진합니다." 즉, 학습 예시에서 보았던 것보다 더 긴 추론 단계를 요구하는 OOD(Out-of-Domain) 테스트 세트에서도 성능 향상을 보였습니다.
- **충분한 규모의 모델에서만 발현:** "이 세 가지 작업에서 보이지 않는 기호에 대한 추상적인 조작을 수행하는 능력은 100B 모델 매개변수 규모에서만 나타납니다." 작은 모델은 여전히 실패합니다.

4. 심층 분석 및 고려 사항

4.1. 사고의 사슬 프롬프트의 작동 원리 (내부 메커니즘)

사고의 사슬 프롬프트가 왜 작동하는지에 대한 심층적인 이해를 위해, 연구자들은 LaMDA 137B의 GSM8K 오류를 수동으로 분석했습니다.

- **의미 이해 및 논리적 추론 개선:** PaLM 62B에서 PaLM 540B로 스케일링함으로써 **"의미 이해(semantic understanding)"** 및 **"한 단계 누락(one-step missing)"** 오류의 상당 부분이 수정되었습니다. 이는 모델 규모가 증가함에 따라 언어 모델이 다양한 의미 이해 및 논리적 추론 기술을 습득한다는 가설과 일치합니다.
- **변수 계산 단독으로는 불충분:** 사고의 사슬이 더 많은 계산(중간 토큰)을 할당할 수 있다는 직관에도 불구하고, 단순히 출력 토큰 수를 늘리는 "변수 계산만(variable compute only)" 변형은 성능 향상에 거의 기여하지 않았습니다. 이는 중간 단계를 자연어로 표현하는 것 자체가 중요한 유용성을 가짐을 시사합니다.
- **추론 순서의 중요성:** "답변 후 추론(reasoning after answer)" 변형은 표준 프롬프트와 거의 동일한 성능을 보였는데, 이는 사고의 사슬에 담긴 **"순차적 추론(sequential reasoning)"**이 지식을 활성화하는 것을 넘어선 유용성을 가짐을 시사합니다.

4.2. 프롬프트 엔지니어링의 역할 및 견고성

프롬프트 엔지니어링의 민감도는 프롬프트 접근 방식의 주요 고려 사항입니다.

- **다양한 주석자 및 예시:** "사고의 사슬 프롬프트는 다른 주석자가 작성한 프롬프트 예시에서도 다양성을 보이지만, 표준 프롬프팅보다 뛰어난 성능을 보입니다." 이는 성공적인 사고의 사슬 사용이 특정 언어적 스타일에 의존하지 않음을 의미합니다.

- **순서 및 개수 변화에 대한 견고성:** 소수의 예시 순서를 변경하거나 예시 수를 변경해도 사고의 사슬 프롬프트의 성능 이득은 대체로 유지되었습니다.
- **모델 간 전이:** 동일한 프롬프트로 사고의 사슬 프롬프트는 LaMDA, GPT-3, PaLM 세 모델 모두에서 성능을 향상시켰습니다(CSQA 및 StrategyQA의 GPT-3 제외).
- **여전히 중요한 프롬프트 엔지니어링:** 그럼에도 불구하고 **"프롬프트 엔지니어링은 여전히 중요하며, 많은 경우에 성능을 크게 향상시킬 수 있습니다."** 일부 작업에서는 특정 프롬프트 엔지니어링이 필수적입니다.

4.3. 한계점 및 향후 연구 방향

- **진정한 '추론' 여부:** "사고의 사슬이 인간 추론자의 사고 과정을 모방하지만, 신경망이 실제로 '추론'하는지 여부는 여전히 미해결 질문으로 남습니다."
- **고비용의 수동 주석:** 소수의 예시 설정에서는 수동 주석 비용이 미미하지만, 미세 조정을 위한 대규모 고품질 주석 세트를 생성하는 것은 비용이 많이 들 수 있습니다. 합성 데이터 생성이나 제로샷 일반화가 대안이 될 수 있습니다.
- **오류 가능성:** 추론 경로의 정확성을 보장할 수 없으며, 이는 올바른 답변뿐만 아니라 우연히 올바른 답변이나 잘못된 답변으로 이어질 수 있습니다. LLM 생성의 사실성을 개선하는 것은 향후 연구 방향입니다.
- **대규모 모델 필요성:** 사고의 사슬 추론이 대규모 모델에서만 창발적으로 나타난다는 점은 실제 응용 프로그램에서 서비스 비용이 많이 들 수 있음을 의미합니다. 더 작은 모델에서 추론을 유도하는 방법에 대한 연구가 필요합니다.

5. 결론

"사고의 사슬 프롬프트는 언어 모델에서 추론을 향상시키는 간단하고 광범위하게 적용 가능한 방법"임을 입증했습니다. 산술, 기호 및 상식 추론 실험을 통해 사고의 사슬 추론이 **모델 규모의 '창발적 속성'**이며, 충분히 큰 언어 모델이 그렇지 않으면 성능 곡선이 평평한 추론 작업을 수행할 수 있도록 합니다. 이 연구는 표준 프롬프트가 대규모 언어 모델의 기능에 대한 **"하한선(lower bound)만을 제공한다"**는 점을 강조하며, 언어 기반 추론에 대한 추가 연구를 고무할 것입니다.