거대언어모델의 추론능력 평가를 위한 MC-LARC 데이터셋*

신동현[O2], 황산하[1], 이석기[1], 김윤호[2], 이승필[2], 김선동[1Ψ]
광주과학기술원 AI대학원[1] 광주과학기술원 전자전기컴퓨터공학부[2]
{dong97411, hsh6449j, sklee1103, dbsgh797210, iamseungpil}@gm.gist.ac.kr, sundong@gist.ac.kr

MC-LARC Benchmark to Measure LLM Reasoning Capability

Donghyeon Shin[O2] Sanha Hwang[1] Seokki Lee[1] Yunho Kim[2] Seungpil Lee[2] Sundong Kim[1Ψ]
GIST AI [1] GIST EECS [2]

요 약

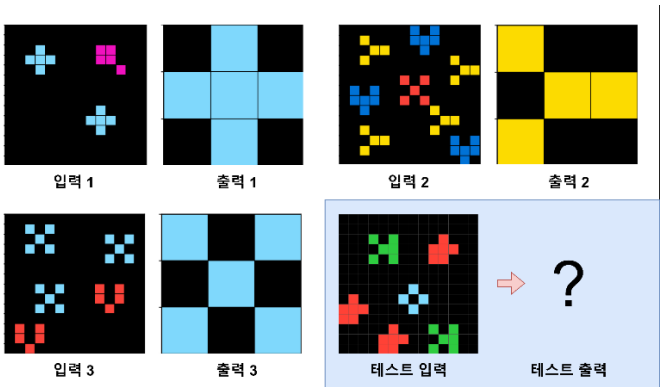
Abstract Reasoning Corpus (ARC)는 적은 이미지 데이터만 주어져 있기 때문에 많은 데이터를 요구하는 기존의 인공지능 모델로는 해결하기 어렵다. 한편, 거대언어모델은 다양한 분야에서 높은 성능을 보이고 있다. 따라서 본 연구에서는 거대언어모델의 추론 능력을 이용하여 ARC를 풀기 위한 새로운 데이터셋 MC-LARC를 제안한다. 본 연구에서 제시하는 MC-LARC 데이터셋은 *'ARC의 예시 입력 이미지를 설명하는 문장 1 개'*와 *'문제 풀이 규칙을 설명하는 문장 5 개'*로 이뤄진 데이터셋이다. 이미지 설명 문장은 사람이 직접 작성하였고, 문제 풀이 규칙 문장은 GPT4-32k을 활용하여 만들었다. 이후, MC-LARC를 이용하여 ChatGPT4 와 사람을 대상으로 주어진 이미지에 대해서 적절한 설명을 선택할 수 있는지에 대한 추론 능력 테스트 실험을 수행하였다.

1. 서 론

현시점에서 거대언어모델은 자연어처리의 여러 분야에서 높은 성능을 보여주고 있다. 하지만, 거대언어모델에게 추론 능력이 있는지에 대해서는 여러 비판이 제기되고 있다 [1]. 몇몇 선행 연구들은 추론 능력을 평가하기 위해 제안된 데이터셋인 Abstraction and Reasoning Corpus (ARC) 데이터셋[2]에 대해 거대언어모델이 좋은 성능을 보이지 못한 점에서 추론 능력을 가지지 못했다고 평가한 바 있다 [3][4][5]. 하지만, 거대언어모델이 추론능력을 평가하는 과정에서 1) 주어진 데이터 사이의 관계 추론과, 2) 이를 문장으로 생성하는 두 단계로 이루어진다는 점에서 ARC를 포함한 기존의 데이터셋은 추론 능력을 평가하는 적절한 벤치마크로 작동하지 못하고 있을 가능성이 크다. 따라서 본 논문에서는 거대언어모델의 추론 작업에서, 문장으로 생성하는 과정을 제외하고 객관식 문제로 추론 능력만을 적절히 평가할 수 있는 새로운 데이터셋 MC-LARC (Multiple-Choice LARC) 를 제안하고자 한다.

우리의 MC-LARC는 ARC를 기반으로 하고 있다. 기존의 ARC는 이차원 행렬로 이뤄진 작은 이미지로, 입력 이미지와 출력 이미지 사이의 규칙을 추론하는 문제이다. LARC (Language-complete ARC) [6]는 ARC의 입력 이미지와 규칙을 문장으로 기술함으로써 ARC와 자연어 처리 분야 연계 가능성을 제시했다. 하지만 기존의 LARC는 Amazon Mechanical Turk 크라우드 소싱 (Crowdsourcing)의 통제되지 않은 환경에서 참여자들이 자율적으로 작성해서 수집했기 때문에 문제에 대한 정보를 충분히 포함하지 않는 잘 정제되지 않은 데이터라는 문제가 있다. 따라서 기존 LARC 데이터셋을 수작업으로 정제하고, 이후에 정제된 데이터셋과 ChatGPT4-32k 모델을 활용하여 MC-LARC 데이터셋을 생성하였다.

MC-LARC는 *1) ARC의 예시 입력 이미지를 설명하는 문장 1 개*와 *2) 문제 풀이 규칙을 설명하는 문장 5 개*로 이뤄져 있다. 규칙을 설명하는 문장은 문제의 규칙을 올바르게 설명한 정답 문장 1 개와 오답 문장 4 개로 구성된 객관식의 형태이다. 이처럼 오지선다 객관식 문제로 치환함으로써 MC-LARC는 이미지 기반의 ARC 문제를 텍스트 기반의 문제로 확장할 수 있었다. 더불어, 출력 이미지를 생성하는 프로그램을 강제하는 기존의 ARC 문제와 달리 유사도 기반의 모델의 사용을 가능케 하여 거대언어모델의 추론 능력을 더욱 손쉽게 평가할 수 있다.



[그림 1] (하얀 배경) 입력 이미지와 출력 이미지를 보고 규칙을 유추하여 (파란 배경) 알맞은 출력 이미지를 추론한다.

2. 데이터셋

2.1 원본 ARC 데이터셋

Abstraction and Reasoning Corpus (ARC) 데이터셋 [2]은 컴퓨터 시스템의 지능 측정을 위한 목적으로 만들어졌다. 이 데이터셋은 산수 능력, 기하학적, 위상적 이해 등의 복합적인 사전 지식을 바탕으로 한 깊은 사고와 추론을 요구한다. [그림 1]은 ARC 데이터의 예시로, 세 개의 예제로부터 공통된 규칙을 도출하고, 테스트 입력 이미지에 적용하여 알맞은 출력 이미지를 추론해 내는 것을 목표로 한다. 각 문제에는 입력 이미지와 출력 이미지의 2~5 쌍이 예시로 주어진다. 원본 ARC 데이터셋은 400 개의 학습 데이터와 400 개의 평가 데이터 그리고 200 개의 테스트 데이터로 구성되어 있다. ARC 데이터셋은 2 차원 행렬로 표현되는데 이를 시각화 하면 [그림 1]과 같은 이미지로 표현된다.

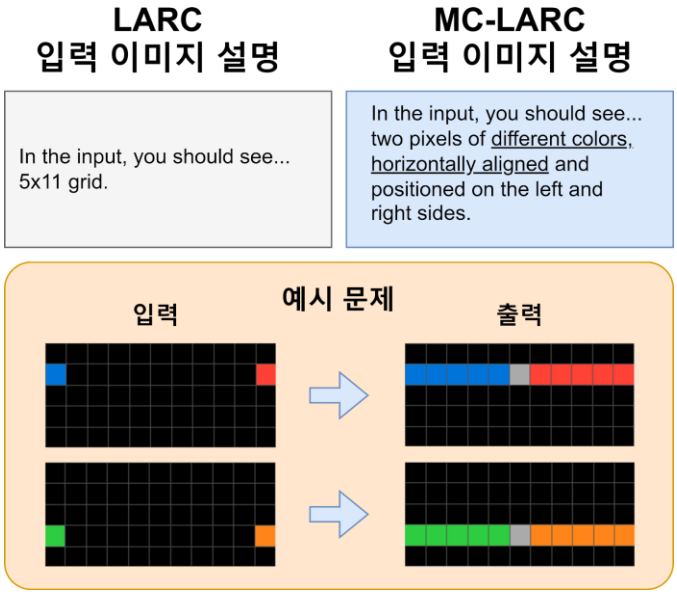
본 연구에서 제안하는 MC-LARC의 입력 이미지에 대한 설명문을 작성할 때, [그림 1]에서 확인할 수 있는 시각화 된 원본 ARC의 입력 이미지를 참고하였다.

* 이 논문은 과학기술정보통신부의 재원으로 한국연구재단과 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00240062, RS-2023-00216011, 2019-0-01842)

2.2 텍스트 데이터 – LARC 데이터셋

LARC 데이터셋은 원본 ARC 학습 데이터셋 400 개에 대해 각각의 1) 입력 이미지에 대한 설명, 그리고 입력 이미지와 출력 이미지 사이의 2) 규칙이 기술되어 있다. 여기에 참여자들이 본인이 기술한 문장에 대해 어느 정도 확신이 있는지를 나타내는 지표인 신뢰도(confidence) 항목을 추가하였다.

하지만 기존의 LARC의 경우 문제 해결에 있어서 충분한 정보를 제공하지 못한다는 한계를 지니고 있다. 먼저, 기존의 LARC 데이터셋의 경우 한 명의 전문가가 아닌 다수의 비전문가가 작성하였기 때문에 일관성이 떨어진다. 예를 들어, 동일한 색의 픽셀에 대하여 서로 다른 표현을 사용하는 경우가 존재한다. 또한 LARC에는 신뢰도가 높음에도 불구하고 [그림 2]의 왼쪽 설명과 같이 문제와 관련 없는 텍스트가 기술되어 있는 경우도 있어 신뢰도 항목을 같이 조회하더라도 데이터를 온전히 신뢰할 수 없는 문제가 있었다. 따라서 LARC의 문제 설명 부분을 정제하여 의미 있는 정보를 담았다.



[그림 2] (아래) 예시 문제의 입력 이미지에 기반하여 (왼쪽 위) 기존 LARC 입력 이미지 설명에서 (오른쪽 위) MC-LARC 입력 이미지 설명으로 수정하였다.

3. Multiple Choice LARC (MC-LARC) 데이터셋

제안하는 MC-LARC는 ARC 문제를 설명하는 LARC와 유사한 오답을 추가하여 오지선다 형태로 만든 벤치마크이다. 기존의 LARC와 동일하게 입력 이미지에 대한 설명과 문제 풀이 규칙에 대한 설명 두 가지를 포함하고 있다. 앞서 언급한 기존 LARC의 문제점을 해결하기 위해서 먼저 기존 LARC의 입력 이미지에 대한 설명과 규칙 문장을 LARC의 신뢰도 데이터를 이용하여 1 차 필터링을 한 뒤에, 사람 수준에서 한 번 더 확인하여 정제하였다. 이 오지선다 문제를 푸는 과정을 통해 거대언어모델이 텍스트를 생성하는 과정을 생략하고 추론하는 능력만을 평가하고자 하였다.

3.1 입력 이미지에 대한 문장 정제

MC-LARC의 입력 이미지에 대한 설명을 만들어 내기 위해 기존 LARC 400 개의 입력 이미지에 대한 설명 문장 정제를 진행하였다. 모델의 추론능력을 더 엄밀하게 평가하기 위해 입력 이미지에 대한 설명이 규칙을 암시하지 않도록 했고, 각 문장은 색, 객체 정보, 수리, 기하 및 위상, 상식의 5 가지를 기준으로 이에 대한 설명이 잘 드러나도록 정제했다. 이에 대한 자세한 설명은 아래와 같다.

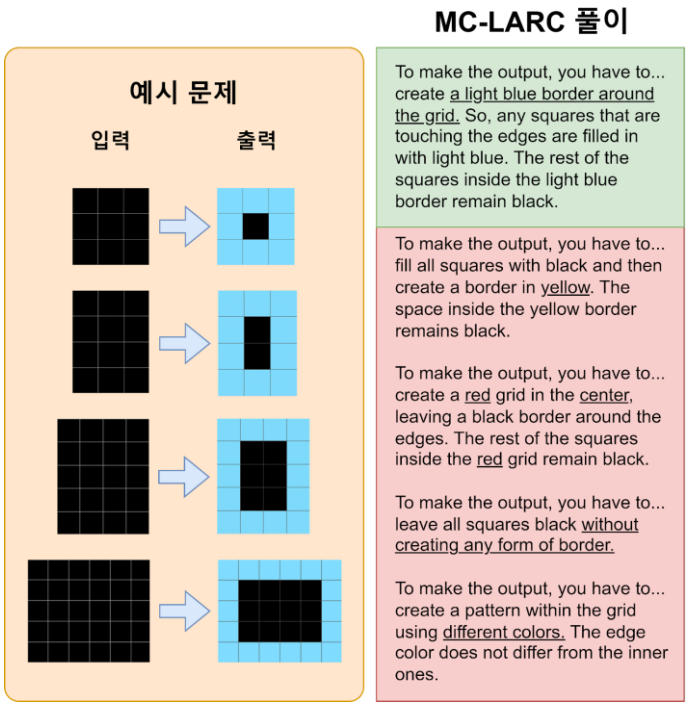
•색깔: ARC 데이터셋은 규칙과 색이 연관된 경우가 많기 때문에 픽셀의 색깔에 대한 정보를 인지하는 것은 중요하다. 이를 고려하여 색깔 정보를 포함하였다.

•객체 정보: ARC 문제는 문제의 50%가 객체와 연관되어 있을 정도로 객체에 대한 정보가 문제 해결에 중요하다[4]. 인접한 픽셀, 또는 동일한 색의 픽셀, 그리고 비슷한 패턴의 집합을 객체라고 볼 수 있다.

•수리: 일부 문제에서 픽셀의 개수 또는 객체의 개수를 인지하여야만 풀 수 있는 문제들이 있다. 픽셀 또는 객체의 개수 정보를 통해 해당 정보를 인지하도록 하였다.

•기하 및 위상: 픽셀들이 모여 삼각형과 사각형 등의 기하 형태와 픽셀 간의 위치 관계가 드러난 문제에서 이에 대한 정보를 기술하였다.

•상식 : 일상에서 쉽게 접할 수 있는 물리 개념, 또는 방사형 패턴, 체크 패턴 등에 대한 정보들을 인간의 기본 상식으로 간주하여 작성하였다.



[그림 3] (왼쪽) 예시 문제에 대한 (오른쪽) MC-LARC (초록 바탕) 정답 선택지 1 개와 (빨강 바탕) 오답 선택지 4 개를 확인할 수 있다.

3.2 규칙에 대한 문장(풀이) 증강

거대언어모델인 ChatGPT4-32k를 활용하여, 정제된 LARC의 규칙에 대한 문장(정답)과 비슷하면서도 의미는 완전히 다른 오답 선택지 4 개를 [그림 3]과 같이 추가적으로 만들었다.

적절한 오답 문장을 생성하기 위해서 본 연구에서는 프롬프트 수준에서 크게 두 가지의 제약을 걸었다. 첫 번째로, 오답 생성 시 동의어로 바꾸지 못하게 제한하여 단순히 동의어로 바꾼 문장을 오답으로 생성하는 현상을 방지하였다. 두 번째로, ARC 데이터의 환경에 대한 배경지식을 제공하여 ARC 데이터셋 맥락에 맞지 않는 문장 생성을 최소화하도록 하였다. 위에서 설명한 프롬프트를 통해 정답 문장과 비슷하지만, 내포된 의미는 완전히 다른 오답 문장을 증강하였다.

또한, 거대언어모델이 정답에만 존재할 수 있는 특정 부분을 보고 정답을 찾는 지름길 학습 현상을 방지하기 위해 문장 형식을 통일하였다. 같은 논리로 정답의 순서 역시 무작위로 변화를 주어 정답의 순서를 통해 정답을 맞히는 것을 방지하였다.

4. 실험 및 결과

본 연구에서는 사람과 거대언어모델이 MC-LARC를 얼마나 잘 푸는지 측정하였다. 이후 입력 이미지에 대한 설명을 힌트로 제공했을 때와 제공하지 않았을 때의 차이를 비교하였다. 이를 통해 LARC와 MC-LARC를 비교하여 우리의 MC-LARC의 품질을 평가하였다.

4.1 MC-LARC 평가 – ChatGPT

언어모델 실험은 ChatGPT Plus에서 제공하는 GPT-4 를 활용하여 실험을 진행하였다. 현재 ChatGPT Plus는 최대 4 개의 이미지를 첨부하여 질문을 할 수 있기 때문에 입력 이미지-출력 이미지 쌍은 최대 4 개까지 제공하였다.

실험은 총 두 방법으로 진행되었으며, 수정된 입력 이미지 설명 데이터와 ARC 데이터셋에 대한 배경지식을 설명하는 경우와 그렇지 않은 경우로 나누어 실험을 진행하였다.

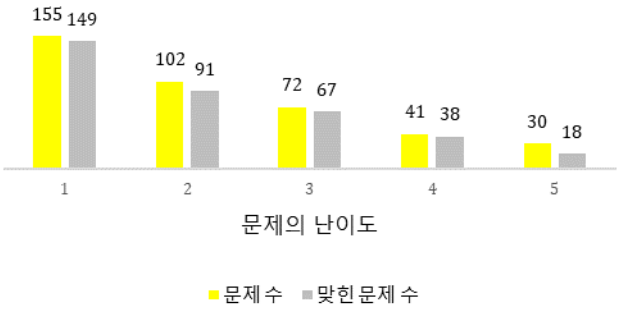
[표 1] 프롬프트/input LARC 제공 여부에 따른 정답 비율과 선행 연구[5]의 결과를 보여준다.

MC-LARC	정답 비율 (맞춘 문제 수/총 문제 수)		
구분	입력 이미지 설명 O	입력 이미지 설명 X	GPT4-0613[5]
ChatGPT4	337/400 **(84.25%)**	316/400 **(79.00%)**	77/800 **(9.625%)**

결과를 살펴보면, ChatGPT4 는 두 경우 모두에 대해 약 80%의 정확도를 달성했다. 이는 거대언어모델을 이용하여 원본 ARC 를 푼 선행 연구[5]와 비교하여 매우 높은 정답률을 보여준다.

4.2 MC-LARC 평가 – 사람

사람을 대상으로 한 MC-LARC는 **90.75%**의 정답률을 보였으며, ARC 데이터셋에 대한 선행 연구[7]에서 83.8%의 정답률을 보인 것에 비해 더 높은 정답률을 보였다. [그림 4]는 총 400 개의 문제에 대해 사람이 평가한 난이도의 분포와 맞힌 문제 수를 보여준다.



[그림 4] 문제의 난이도 (1~5)에 따른 문제 수 (노란색)와 실제로 맞힌 문제 수 (회색)를 보여준다.

문제 난이도의 평균 값은 2.22 로 전체적인 문제 난이도는 어렵지 않았음을 알 수 있다. 문제 난이도 1~4 에 대해서는 약 90%의 정답률을 보였으나 난이도 5 에 대해서는 정답률이 60%로 떨어졌다.

5. 향후 연구 제안

첫 번째로, 입력 이미지에 대한 설명을 사람이 만든 수준으로 생성할 수 있는 모델 연구를 제안한다. MC-LARC 데이터셋을 정제할 때, 선별된 문장을 다시 일일이 보며 수작업으로 진행했기 때문에 만드는 비용이 많이 필요하였다. 또한 ARCathon [8]에서 성능을 평가하는 테스트 데이터셋에는 입력 이미지에 대한 텍스트 설명이 주어지지 않기 때문에 텍스트 데이터를 활용하기 위해서는 직접 텍스트 데이터를 생성하는 모델이 필요하다. 따라서 입력 이미지 설명 생성 모델을 구현하여 앞서 언급한 한계들을 극복하고, 이미지 추론 문제를 텍스트 추론 문제로 변환하여 해결할 수 있는 기반을 마련하고자 한다.

두 번째로, 이미지와 텍스트 데이터에 대한 정보를 동시에 사용하는 멀티모달 모델 구조를 ARC 문제 해결에 사용해 보고자 한다. 이미지 정보를 통해 거대언어모델의 추론 능력을 한층 더 활용할 수 있고, 더 나아가 해당 모델의 파라미터를 이용해서 이미지를 생성할 수 있는 모델을 구축함으로써 궁극적으로 ARC 문제 해결을 목표로 하고 있다.

6. 결론

본 연구에서는 ARC 문제를 이미지 추론 문제에서 텍스트 추론 문제로 바꿨다. 또한 기존의 거대언어모델이 추론 능력에 취약했던 한계를 감안하여 문장 생성 작업을 제외하고, 추론 능력만을 평가하기 위해 다지선다 문제 데이터셋 MC-LARC를 제안하였다. 이 데이터셋 생성을 위해 사용했던 LARC 데이터셋의 신뢰성이 떨어진다는 점에서 사람 수준에서 데이터셋을 평가하고 수정하는 작업이 추가로 필요하다는 점을 인지하였다.

그럼에도, MC-LARC와 같은 다양한 시도를 통해 인공 일반 지능(AGI)으로서 거대언어모델의 강점과 취약점, 한계점을 발견하는 데 기여할 수 있고, 이를 통해 인공 일반 지능의 발전에 기여할 것이라 기대한다.

참고문헌

[1]Nouha Dziri et al., “Faith and Fate: Limits of Transformers on Compositionality”, arXiv:2305.18654, 2023.
[2]Francois Chollet, “On the Measure of Intelligence”, arXiv:1911:01547, 2019.
[3]Arseny Moskvichev et al., “The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain”, arXiv:2305.07141, 2023.
[4]Yudong Xu et al., “LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations”, arXiv:2305.18354, 2023.
[5]Suvir Mirchandani et al., “Large Language Models as General Pattern Machines”, arXiv:2307.04721, 2023.
[6]Sam Acquaviva et al., “Communicating Natural Programs to Humans and Machines”, NeurIPS, 2022.
[7]Aysja Johnson, Wai Keen Vong, Brenden M. Lake, Todd M. Gureckis, "Fast and flexible: Human program induction in abstract reasoning tasks", arXiv:2103.05823, 2021.
[8]“ARCathon”, https://lab42.global/arcathon/ (accessed October 31th, 2023).