인지 및 추론 연구를 위한 Mini-ARC 벤치마크 데이터

김수빈 O1, Prin Phunyaphibarn[1], 안동현[1], 김선동[2]
[1]한국과학기술원 전산학과 [2]광주과학기술원 AI대학원
{21supersoo, prin10517, segaukwa}@kaist.ac.kr, sundong@gist.ac.kr

Playgrounds for Abstraction and Reasoning

Subin Kim O1, Prin Phunyaphibarn[1], Donghyun Ahn[1], Sundong Kim[2]
[1]School of Computing, KAIST [2]AI Graduate School, GIST

요 약

프랑소와 숄레(François Chollet)가 제안한 Abstraction and Reasoning Corpus (ARC)는 특정 문제에 매몰되지 않고, 일반화가 가능한 지능의 개발을 위해 디자인된 아이큐 테스트 형태의 벤치마크로, 인간과 컴퓨터 모두의 인지 능력을 측정하기에 적합하다. 대부분의 문제를 풀 수 있는 사람에 반해, 30% 이상의 문제를 풀 수 있는 컴퓨팅 기반 ARC-Solver 는 알려지지 않았다. 이 연구에서는 기존 ARC 의 난이도를 유지하면서도 모델의 복잡도를 간소화하기 위해 탐색 공간을 최소화한 벤치마크 데이터 Mini-ARC 를 소개한다. Mini-ARC 의 수집을 위해, 인간의 풀이 과정을 추적할 수 있는 인터페이스인 O2ARC 를 고안하였으며, 이를 통해 25 명에게 총 3,000 여 개의 풀이를 수집하는 데 성공했다. 이 연구는 간소화된 인간의 인지 과정과 그 풀이 과정을 대량으로 확보하는 시스템을 구축하여, 컴퓨팅 기반 ARC-Solver 개발의 새로운 접근법을 제시한다. Mini-ARC 데이터셋은 https://bit.ly/Mini-ARC 에서 확인할 수 있다.

1. 서 론

구글의 LaMDA, Open AI의 DALL·E 등을 필두로 개발되고 있는 인공지능 모델은 인간이 평생 접하지 못할 크기의 지식을 습득한 후 패턴화하여 분류, 이상 탐지, 이미지 생성 등 특정 임무에서 인간을 능가하는 성능을 보여주고 있다 [1]. 하지만, 아직 접하지 못한 문제에 대한 해결 능력을 갖춘 범용 인공지능 모델(Artificial General Intelligence; AGI)을 위한 연구는 비교적 부진하다. 일례로, '내부 구조를 모르는 집안에 들어가서 커피를 끓이는 문제'로 알려진 스티브 워즈니악의 커피 테스트는 현존하는 AI기술을 활용해서는 해결이 어려운 문제로, 사전에 준비되지 않은 일에 대응할 수 있는 기반 기술의 중요성을 역설한다.

이러한 환경을 단순화한 문제가 바로 Abstraction and Reasoning Corpus (ARC)이다 [2]. ARC는 적은 양의 데이터로부터 문제에 담긴 패턴을 추론하고 새로운 문제에 적용하는 인지 능력을 측정하기 위해 고안되었다. 다른 IQ 테스트들과 마찬가지로, 살면서 축적해 온 다양한 상식을 바탕으로 문제에 담긴 패턴을 발견하고, 이를 적용할 수 있는 능력을 확인한다. 정확한 정답 그리드를 생성해야 하는 ARC 문제의 특성상 출제자의 의도를 정확히 파악한 후 논리에 근거하여 추론하는 방법론의 개발이 필수적이며, Raven's Progressive Matrices와 같은 오지선다형 IQ 테스트의 솔루션들에서 활용된 유사도 기반의 신경망 모델을 활용하기 어렵다 [3]. 아울러 ARC 문제에서 요구하는 다양한 입출력 형태는 일정 수준 이상의 복잡도를 요구하여, 해결책을 고안하는 과정에서 고려해야 할 부분이 배가되었다.

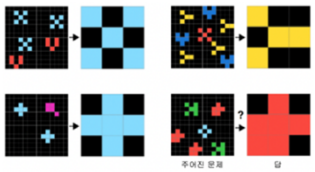


그림 1. ARC의 예시

이 연구에서는 더욱 간소화된 Mini-ARC 벤치마크 테스트를 소개하고, 다양한 기능을 활용해 Mini-ARC를 풀어볼 수 있는 인터페이스인 O2ARC 및 참여자들의 솔루션인 Mini-ARC trace를 소개한다. 입력과 출력의 크기가 다양한 기존의 ARC와 달리, Mini-ARC는 입출력 그리드 공간을 5x5로 제한하여 인공지능 모델의 훈련에 대한 실효성을 높였다. 더불어, 여섯 개의 기능이 추가된 O2ARC 인터페이스는 기존에 비해 높은 수준의 편의성을 제공한다. O2ARC를 통해 수집된 풀이 과정인 Mini-ARC trace는 추후 모방 학습, 프로그램 합성에 기반한 솔루션 개발에 활용할 수 있을 것으로 기대된다.

2. Mini-ARC

이 연구에서 소개하는 Mini-ARC 데이터는 150개의 문제로 구성되었으며, 입출력 크기를 5x5로 제한하여, 인공지능 모델 개발 단계에서의 성능을 효율적으로 확인하고자 한다.

2.1. 제작 원칙

기존의 ARC의 입력과 출력의 크기는 1x1에서 30x30까지 다양하며, 모델 학습에 필요한 탐색 공간은 입출력의 크기에 비례하여 커진다. 따라서 이 연구에서는 학습의 탐색 크기를 제한하여 더욱 가벼운 훈련 환경을 제공함과 동시에, 모든 제작자가 동일한 조건에서 문제를 구상할 수 있도록 Mini-ARC 문제의 입출력 크기를 5x5로 단일화하였다. 크기의 설정

본 논문은 기초과학연구원 (IBS-R029-C2, IBS-R029-Y4) 및 정보통신기획평가원의 지원 (No. 2019-0-01842, 인공지능대학원지원, 광주과학기술원)을 받아 수행된 연구임

배경은 다음과 같다. 1) 입출력 공간이 중심점을 가지도록 크기를 홀수로 설정하였다. 2) 정사각형 모양의 입출력 공간은 문제를 풀 때 전체 그리드의 회전, 뒤집기와 같은 기본적인 변형을 사용하도록 유도한다. 3) 5x5 미만의 경우 회전과 대칭을 구별하기 어렵기 때문에 사용하기 어렵다.

문제 제작을 위해, 25명의 참가자들을 모집하였고, 5x5로 고정된 크기 내에서 ARC 문제의 의도에 맞는 명확하고 유일한 솔루션을 가지는 Mini-ARC를 구상하도록 지시하였다. 패턴 파악을 위해 문제당 최소 4개의 입출력 쌍을 포함하도록 했다. 참가자들은 4시간 동안 새로운 문제 제작에 임하였고, 그 결과 총 150개의 Mini-ARC 문제를 수집할 수 있었다.

2.2. Mini-ARC 유형 분석

Mini-ARC는 각 문제를 구성하는 대표적인 개념에 따라 움직임, 색깔, 객체, 숫자, 기하, 상식의 6가지 범주로 분류할 수 있다.

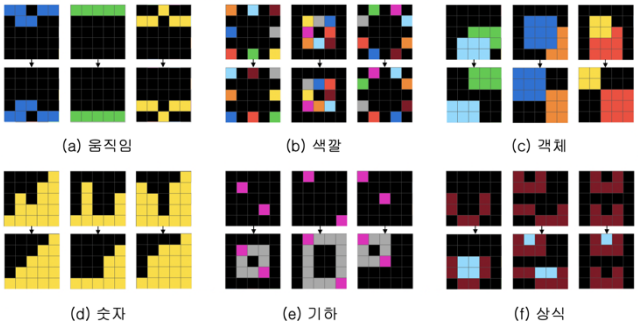
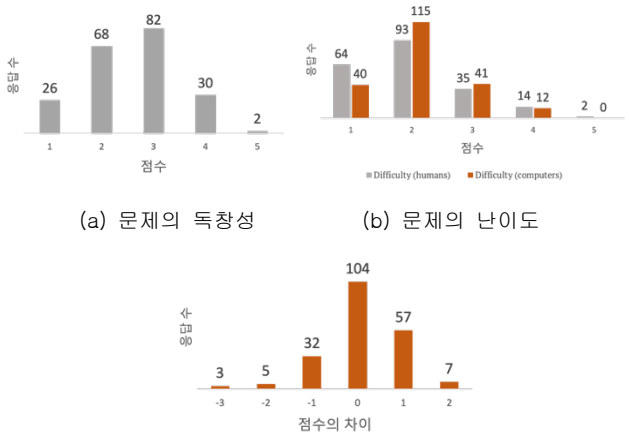


그림 2. Mini-ARC 문제 유형 별 예시

움직임 유형은 뒤집기, 회전 및 옆으로 이동 등과 같은 동적 움직임을 기반으로 한다. 색깔 유형은 각 칸의 색깔을 기반으로 두 칸의 색깔을 교환하는 등의 패턴이 존재한다. 객체 유형은 객체 단위의 상호작용이나 움직임을 파악한다. 여기에서 객체는 배경과 직관적으로 구별할 수 있는 색칠된 영역을 의미한다. 숫자 유형에서는 같은 색깔의 칸의 개수, 또는 한 객체 내에 포함된 칸의 개수 등에 기반한 셈을 요한다. 기하 유형에서는 점, 선, 면 등의 기하학적 개념을 필요로 한다. 상식 유형에서는 높은 수준의 인지 기능을 필요로 하는 문제가 속한다(테트리스, 미로찾기 등). 문제 해결에 쓰이는 기능의 조합 등이 유형별로 다를 수 있다.

2.3. 데이터셋 평가

각 참가자는 다른 참가자들이 생성한 문제를 평가하고 독창성 및 난이도 측면에서 1-5점 사이의 점수를 부여했다. 점수가 높을수록 독창적인 문제이거나, 난이도가 높은 문제라는 것을 의미한다. 문제 난이도는 두 가지로, 평가자가 직접 문제를 풀었을 때 느낀 문제의 난이도와, 해당 문제를 푸는 프로그램을 개발할 때 예상되는 난이도로 나뉜다. 평가자가 직접 풀었을 때 느낀 난이도가 높을수록 고도의 인지 능력을 필요로 한다는 가정을 기반으로, 각 문제가 인간의 인지 능력을 얼마나 활용하는지 측정하기 위한 과정이다. 만약 해당 문제의 프로그램을 개발하는 것이 어렵다면, 문제에 활용된 규칙이 코드를 통해 해결하기엔 직관적이지 않다는 점을 의미한다. 각 문제는 최소 1회 이상의 평가를 받았으며, 총 208회의 평가가 이루어졌다.



(a) 문제의 독창성 (b) 문제의 난이도

(c) 해당 문제를 푸는 프로그램을 개발할 때의 난이도에서 직접 풀었을 때의 문제의 난이도를 뺀 점수의 차이

그림 3. Mini-ARC의 평가 결과 (x축: 점수, y축: 응답의 수)

그림3-(a)에서 볼 수 있듯이, 25명의 참여자들의 기여로 우리는 Mini-ARC를 충분히 독창적인 문제들로 구성할 수 있었다 (평균 2.59). 그럼에도 불구하고, 응답자의 대부분은 주어진 문제를 직관적으로 해결할 수 있다고 평가하였는데 (평균 2.02), 이는 인간은 뛰어난 인지 능력을 토대로 창의적인 문제도 해결할 수 있다는 점을 시사한다. 참여자들은, 각각의 문제를 해결하는 프로그램을 개발하는 것이, 직접 푸는 것보다 더 어렵다고 느꼈다 (평균 2.12). 이를 보여주기 위한 그림이 3-(c)로, 각 응답에서 평가자가 해당 문제에 대한 프로그램을 개발할 때 예상되는 난이도에서 평가자가 느낀 문제의 난이도를 뺀 결과를 보여준다. 두 난이도가 동일하게 느껴졌다는 응답이 전체 응답 수의 절반인 104개이었지만, 나머지 104개의 중 64개의 응답의 경우, 직접 문제를 푸는 것보다 프로그램을 구현하기가 더 어렵다는 의견이 있었다.

2.4. 5x5 공간이 가진 한계 및 의의

5x5로 입출력의 크기를 제한하였기에, 더 큰 공간을 필요로 하는 일부 아이디어는 기각되었다. 그럼에도 불구하고 참가자들은 5x5의 공간 안에서 다양한 종류의 문제를 만들기 위해 노력하였다. 데이터의 크기가 작은 만큼 보다 가벼운 환경에서 Mini-ARC를 위한 인공지능 모델을 개발한 이후, 입출력 크기의 제한이 없는 ARC 데이터셋에 맞는 형태로 모델을 확장해 볼 수 있다.

3. O2ARC

참가자들이 ARC 및 Mini-ARC 문제를 풀어볼 수 있도록 브라우저 기반 인터페이스인 Object-oriented ARC(이하 O2ARC)를 설계하였다. 많은 문제가 객체 지향적이라는 것을

토대로, O2ARC는 ARC 제작자가 제공한 기본적인 툴 위에 문제 해결 과정을 보조하는 6가지 기능을 추가하였다. 파워포인트, 포토샵 등의 편집 도구에서 영감을 받아 추가한 6가지 기능은 상하좌우 움직임, 잘라내기, 되돌리기, 뒤집기, 회전, 여러 칸 동시 선택이다. 더불어, 문제 해결 과정에서 사용한 기능들의 조합과 순서, 그리고 그에 따른 답안을 추적할 수 있다. 알파 테스트를 통해 세부 사항을 개선한 O2ARC는 보다 높은 수준의 편의성을 제공하였다.

4. Mini-ARC Trace

4.1. Mini-ARC trace 수집 과정

O2ARC 인터페이스를 통해 수집된 Mini-ARC 해결 과정을 Mini-ARC trace(이하 trace)로 명명하였다. 이를 위해 20명의 참가자는 2시간 동안 50개의 선별된 Mini-ARC 문제를 해결하였다. 50개의 문제는 2.2장에 언급되었던 6가지 범주 중에서 균일하게 선별되었으나, '상식' 범주에 속한 문제는 최대한 배제하였다. 이 범주에 속한 문제는 높은 수준의 직관을 요구하여 O2ARC에 구현된 기능 만으로는 인지 과정을 제대로 파악하기 힘들기 때문이다. 풀이의 중복을 예방하고 지나치게 단순한 풀이 과정을 지양하기 위해 참가자들은 각 문제 당 최소한 3가지 이상의 다른 풀이 방식을 제출했다. 오답 3개를 연속으로 제출한 경우, 문제의 해결책을 찾지 못하였다고 판단하였다.

4.2. Mini-ARC 예제 및 수집된 trace 분석

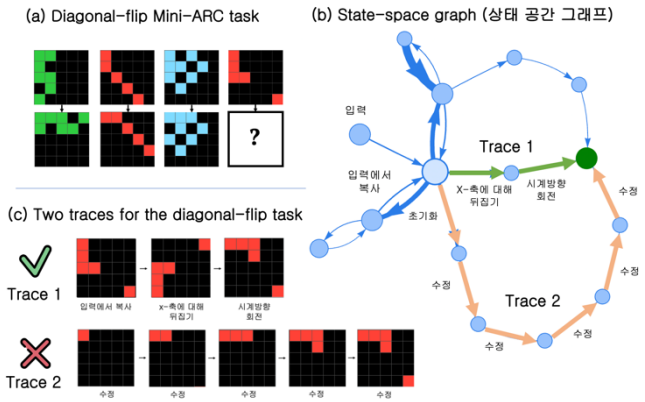


그림4. 대각선 뒤집기 Mini-ARC 문제와 해당 문제를 접근한 두 가지 대표 풀이법을 담은 상태 공간 그래프

그림4-(a)는 그리드 위의 객체를 대각선 축을 중심으로 뒤집는 '대각선 뒤집기' 문제이다. 그림 4-(b)의 상태 공간 그래프는 해당 문제의 서로 다른 풀이법 (trace)들을 나타내며, 상태 공간 그래프 속 초록색 선은 trace 1을, 주황색 선은 trace 2를 나타낸다. 구체적인 풀이 방법은 그림4-(c)를 통해 확인할 수 있으며, 두 참여자가 문제를 풀기 위한 논리를 유추할 수 있다.

Trace 1는 회전, 뒤집기 등 O2ARC에서 새롭게 구현한 기능들을 이용해 대각선 뒤집기 로직을 구현하였다. 반면, Trace 2는 정답에 해당하는 픽셀들을 한 칸씩 빨간 색으로 색칠하는 직관적인 방식으로 문제를 해결하였다. Trace 1에서 구현된 대각선 뒤집기 로직은 모든 예제들에 일반화하여 적용될 수 있지만, trace 2에 사용된 직관은 다른 예제들에 곧바로 적용될 수 없는 불완전한 해결책이다. 결과적으로, 대각선으로 뒤집어야 하는 출제자의 의도를 논리적으로 반영한 trace 1이 완벽한 정답이다.

입출력 크기를 5x5로 제한한 Mini-ARC 문제의 경우 trace 1과 같이 단순한 동작만을 이용하여 푼 솔루션이 상당히 많았다. 이렇게 수집된 trace는 문제 해결 과정을 반영하지 않기 때문에 모델 학습에 사용하기 어렵다. 상태 공간 그래프 내에 존재하는 정보의 옥석을 가리기 위한 고민이 필요하다.

4.3. Mini-ARC trace를 활용한 추후 연구 방안

우리는 인간의 문제 해결 능력을 효율적으로 학습할 수 있는 인공 지능 모델을 만들기 위해 Mini-ARC trace를 수집하였다. Mini-ARC trace를 이용하는 한 가지 방법으로는, 인간을 모방할 수 있는 모델을 훈련하는 imitation learning이 있다. 다양한 문제에서 공통으로 사용된 패턴을 토대로, ARC에서 일반적으로 사용되는 기능의 조합을 도출하고, 이를 활용한 효율적인 프로그램 합성 [6] 및 강화 학습 [7,8] 모델 개발을 계획하고 있다.

5. 결론

이 논문은 간결한 범용 인공 지능 벤치마크인 Mini-ARC 데이터셋과 편리한 문제 해결을 위한 인터페이스인 O2ARC, 그리고 인간의 해결 과정을 담은 Mini-ARC trace를 소개한다. 이를 활용하여 범용 인공 지능에 한 단계 다가갈 수 있는 프로그램 합성 및 강화학습 모델을 연구할 수 있으리라 기대한다.

참 고 문 헌

[1] Aditya Ramesh et al. Zero-shot text-to-image generation. In ICML, 2021.

[2] François Chollet. On the measure of intelligence. arXiv:1911.01547, 2019.

[3] Mikołaj Małkinski and Jacek Mandziuk. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. arXiv:2201.12382, 2022.

[4] Samuel Acquaviva et al. Communicating natural programs to humans and machines. arXiv:2106.07824, 2021.

[5] Aysja Johnson et al. Fast and flexible: Human program induction in abstract reasoning tasks. In CogSci, 2021.

[6] Kevin Ellis et al. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In PLDI, pages 835–850, 2021.

[7] Simon Alford et al. Neural-guided, bidirectional program search for abstraction and reasoning. In Complex Networks, 2021.

[8] Lili Chen et al. Decision transformer: Reinforcement learning via sequence modeling, In NeurIPS 2021.