ACES-SeqPlan: 순차적 계획 탐색을 활용한 자율적 고난이도 퍼즐 생성 프레임워크 제안

임도연, 인수진, 조현중*

고려대학교, *고려대학교

imyryryr@korea.ac.kr, isjin8707@korea.ac.kr, *raycho@korea.ac.kr

ACES-SeqPlan: Using sequential planning exploration Propose an autonomous high-level puzzle generation framework

Im Do Yeon, In Su Jin, Cho Hyeon-Joong*

Korea Univ, *Korea Univ.

요 약

본 논문은 신뢰성 높은 코드 생성 벤치마크를 구축하기 위해, 자율 문제 생성 프레임워크(ACES)의 검증 단계에 추론 중심의 '순차적 계획 탐색(Sequential PlanSearch)'을 결합한 ACES-SeqPlan 프레임워크를 제안한다. 최근 자율 생성 모델은 데이터 부족 문제를 완화할 수 있어보이나, 기존의 단순 샘플링 검증 방식은 풀이 과정이 복잡한 고난이도 문제들을 '해결 불가능'으로 오판하여 폐기하는 '거짓 부정(False Negative)'의 구조적 한계를 보이고 있다. 이에 본 연구에서는 단순 반복(Random Sampling) 시도 대신 문제의 논리적으로 추론하여 해결 가능성을 입증하는 강화된 검증 알고리즘을 도입하여, 기존에는 버려졌던 고가치의 문제들을 선별할 수 있도록 설계하였다. 이를 통해 생성된 문제의 난이도 및 다양성(Quality-Diversity) 간의 상관관계를 분석하였다. 실험 결과를 통해 단순 생성 방식이 갖는 난이도 향상의 임계점을 규명하고, 베이스라인 대비 생성된 문제의 평균 난이도를 획기적으로 상승시킨 성과를 입증함으로써 향후 고품질 벤치마크 구축을 위한 추론 기반의 데이터 검증 전략을 제시한다.

Ⅰ. 서론

최근 GPT-4[1] 및 Llama-3[2]와 같은 대규모 언어 모델(LLM)은 HumanEval[3], MBPP[4] 등 표준 코드 생성 벤치마크에서 비약적인 성능 향상을 기록하고 있다. 그러나 이러한 벤치마크들은 이미 성능 포화(Saturation) 상태에 이르러 모델 간 우열을 가리는 변별력을 잃어가고 있다. 훈련 데이터 오염(Data Contamination) 가능성으로 인해, 모델이 실질적인 추론(Reasoning)을 수행하여 해법을 도출한 것인지 단순히 학습된 패턴을 암기(Memorization)한 것인지 판별하기 어려워졌다.

이러한 정적인 벤치마크의 한계를 극복하기 위한 대안으로, 모델 스스로 다양한 난이도의 새로운 문제를 생성하여 학습 데이터를 확보하는 자율적 탐색(Autotelic Exploration) 방법론이 주목받고 있다. 그중 대표적인 프레임워크인 ACES(Autotelic CodE Search)[5]는 LLM을 생성 주체(Proposer)로 활용하여, 창의적이고 다양한 프로그래밍 퍼즐을 생성함으로써 데이터의 희소성 문제를 해결하고자 제안되었다. 또한 모델이 자신의 능력을 스스로 탐색하고 확장하는 자동화된 능력 발견(Automated Capability Discovery)연구로까지 이어지고 있다.

하지만 기존 ACES 방식은 생성된 문제의 유효성을 검증하는 과정(Solver)에서 단순한 샘플링 기법에 의존한다는 치명적인 한계를 지닌다. 이는 논리적으로는 합당하지만 복잡한 추론 과정을 요구하는 고난이도 문제들을 검증기가 풀지 못해 '해결 불가능'한 문제로 오판하여 폐기하는 구조적 모순을 야기한다. 결과적으로 이러한 '거짓 부정(False Negative)' 현상은 생성될 수 있는 문제의 난이도 상한선을 검증기의 낮은 성능 수준으로 제약하며, 벤치마크의 고도화를 저하시키는 주요 원인으로 작용한다.

본 연구에서는 앞서 언급한 구조적 한계인 '거짓 부정(False Negative)' 문제를 해결하고 벤치마크의 난이도 상한선을 실질적으로 확장하기 위해, ACES 프레임워크의 검증 단계에 PlanSearch[6]를 결합하는 새로운 접근법을 제안한다. 제안하는 방법론은 단순 샘플링에 의존하던 기존 Solver를 계획(Plan) 기반의 추론 과정으로 대체함으로써, 논리적으로 타당하지만 복잡하여 폐기되던 고난이도 문제들의 해결 가능성을 입증한다. 이 제안을 통해 벤치마크의 유효 탐색 공간을 넓히는 동시에, LLM의 심층적인 추론 능력을 보다 정밀하게 평가 및 향상시키고자 한다.

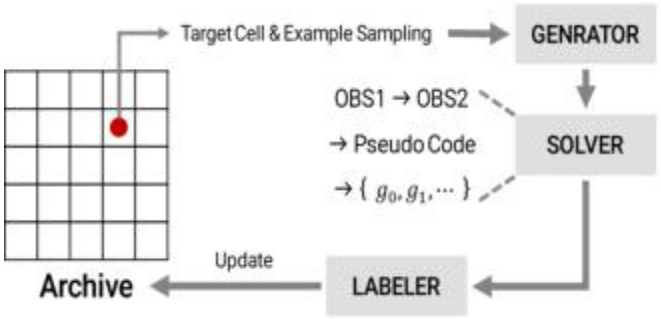
Ⅱ. 본론

1. 데이터셋 정의

본 연구에서 다루는 문제 생성 및 해결의 대상 도메인은 Python Programming Puzzles (P3)의 형식을 따른다. P3는 자연어 문제 설명이 갖는 내재적 모호성을 배제하고, 검증 함수 f의 실행 결과만으로 해답(Solution)의 유효성을 즉각적이고 객관적으로 판별할 수 있는 구조를 갖는다. 이는 평가 과정에서의 인적 개입을 최소화하여 대규모 자동화 검증을 가능케 한다. 따라서 본 연구에서는 이러한 객관적 검증의 이점과 선행 연구인 ACES 프레임워크와의 방법론적 일관성을 유지하기 위해, P3 형식을 실험의 핵심 도메인으로 채택하였다.

2. ACES-SeqPlan: 프레임워크 개요

본 논문에서는 프로그래밍 퍼즐의 자율적 탐색을 위해, 목표 지향적 생성과 계획(Plan) 기반 검증을 통합한 ACES-SeqPlan 프레임워크를 제안한다. 본 프레임워크는 Map-Elites 알고리즘에 기반한 아카이브(Archive)를 중심으로 작동하며, 생성기(Generator)와 검증기(Solver), 그리고 라벨러(Labeler)가 유기적으로 상호작용하는 순환 루프로 구성된다.



(그림 1) Generation 당 ACES-SeqPlan 루프 구조도

Solver는 Generator가 만든 퍼즐의 소스 코드(f)를 입력받아 그 안에 숨겨진 제약 조건(예: 소수 판별, 회문 구조 등)을 분석한다. 이 과정을 통해 도출된 자연어 관찰(Observation)은 복잡한 수식이나 코드를 '사람이 이해할 수 있는 명확한 지시문'으로 변환하는 역할을 하며, 이는 이후 단계에서 정답 함수(g)를 설계할 때 논리적 오류를 방지하게 된다.

3. ACES-SeqPlan: 설계 의도

ACES-SeqPlan의 핵심은 기존의 단순 반복(Random Sampling) 검증 방식을 PlanSearch 기법에 기반한 '추론 중심 검증(Reasoning-based Verification)'으로 대체하였다는 점에 있다. 본 연구에서 연산 비용의 증가를 감수하면서까지 이러한 구조를 채택한 당위성은 다음과 같다.

가. '거짓 부정(False Negative)'의 최소화

기존의 단순 검증기는 논리적으로 완벽하지만 난이도 높은 문제를 '해결 불가능'으로 오판하여 폐기하는 한계가 있다. 본 연구는 추론 능력이 강화된 검증기를 통해, 이러한 높은 가치를 가질 수 있는 문제들이 아카이브에 생존하게 함으로써 벤치마크의 난이도 상한선에 대하여 비약적인 향상을 시도하였다.

나. 답변 함수(g)의 질적 향상

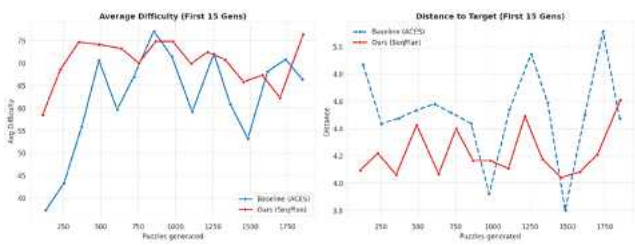
더 까다로운 기준의 검증기를 통과해야만 아카이브에 남을 수 있으므로, 결과적으로 아카이브에는 단순한 속임수가 아니라 깊이 있는 논리를 가진 문제들만 축적된다. 중요한 점은 이들이 다시 다음 세대의 퓨샷 예시(Example Puzzles)로 사용됨으로써, 전체적인 생성 품질이 상향 평준화되는 선순환 구조를 기대할 수 있을 것이다.

Ⅲ. 실험

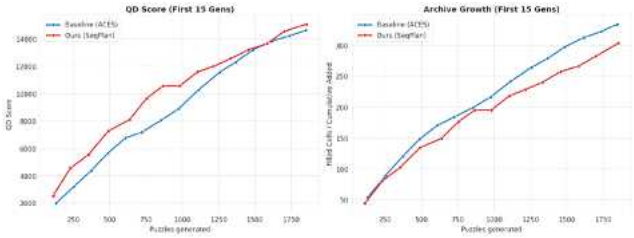
1. 실험 환경 설정

본 연구는 경량화된 환경에서의 프레임워크 검증을 위해 Llama-3-8B-Instruct를 백본 모델로 채택하였으며, 총 15세대(Generations)에 걸쳐 실험을 수행하였다.

2. 실험 결과 및 분석



(그림2) 평균 난이도 및 목표 거리 지표 비교 그래프



(그림 3) QD Score 및 Archive Coverage 비교 그래프

(그림 2, 3)으로 보아 ACES-SeqPlan 프레임워크는 초기에 평균 난이도 58.4를 기록하며, 세대가 지날수록 10점대로 하락하는 베이스라인 대비 월등히 높은 난이도 상한선을 유지하였다. 또한 Target distance 지표에서도 4.09를 기록하여 베이스라인(5~6)보다 정교한 탐색 능력을 입증하였다. 비록 연산 비용으로 인해 생성량은 약 15% 감소하였으나, 난이도와 정밀도가 획기적으로 향상했음을 알 수 있다.

Ⅳ. 결론

본 연구를 통해 우리는 경량화된 연산 자원(Llama-3-8B-Instruct) 환경에서도 제안된 ACES-SeqPlan 프레임워크가 유효하게 작동함을 확인하였다. 특히, 추론 능력이 강화된 검증 과정을 통해 기존 단순 샘플링 방식이 '거짓 부정(False Negative)'으로 오판하여 폐기했던 고난이도 문제들을 성공적으로 발굴할 수 있었다. 이는 연산 비용의 증가를 감수하더라도 개별 문제의 품질과 탐색의 정밀도를 높이는 것이 벤치마크의 실질적 가치를 향상시키는 전략으로서 유효함을 입증한 결과이다.

다만, 본 연구는 컴퓨팅 자원의 제약으로 인해 기존 PlanSearch의 병렬 탐색 구조를 단일 트리 탐색(Single Tree Search)으로 축소 적용하였으며, 생성 횟수(Generations)를 기존 40회에서 15회로 제한함에 따라 다양성(Diversity) 확보에 있어 일부 구조적 한계를 보였다.

향후 연구에서는 병렬 트리 탐색(Parallel Tree Search)의 도입과 생성 규모의 확대를 통해 이러한 한계를 극복하고 모델의 일반화 성능을 강화할 계획이다. 나아가 다양한 관찰(Observation) 경로의 조합을 유도함으로써, 문제 해결의 다양성을 획기적으로 높이고 단순 난이도 상승을 넘어 해결 경로의 창의성(Creativity)까지 담보하는 차세대 벤치마크를 구축하고자 한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부의 재원으로 한국연구재단 중견연구(창의연구형) 사업의 지원을 받아 수행되었음(과제번호: RS-2025-16066493).

참 고 문 헌

[1] OpenAI. "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[2] Dubey, Abhimanyu, et al. "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[3] Chen, Mark, et al. "Evaluating Large Language Models Trained on Code," arXiv preprint arXiv:2107.03374, 2021. (HumanEval)

[4] Austin, Jacob, et al. "Program Synthesis with Large Language Models," arXiv preprint arXiv:2108.07732, 2021.

[5] Pourcel, J., et al. "ACES: Generating a diversity of challenging programming puzzles with autotelic generative models," Advances in Neural Information Processing Systems 37, pp. 67627-67662, 2024.

[6] Wang, E., et al. "Planning in Natural Language Improves LLM Search for Code Generation," arXiv preprint arXiv:2409.03733, 2024.