경량 Vision-Language Model 의 강화학습 최적화: GymCards 환경에서 CoT 및 LoRA 적용 사례 연구

김태수 [1], 유헌창 [2]
[1] 고려대학교 SW·AI 융합대학원 인공지능융합학과 석사과정
[2] 고려대학교 정보대학 컴퓨터학과 교수

kimts92@korea.ac.kr, yuhc@korea.ac.kr

Reinforcement Learning Optimization for Lightweight VLMs via CoT and LoRA : A Case Study in GymCards

Taesoo Kim[1], Heonchang Yu[2]
[1]Dept. of Applied Artificial Intelligence, Graduate School of SW·AI Convergence, Korea University
[2]Dept. of Computer Science & Engineering, Korea University

요 약

Vision-Language 모델(VLM)은 시각적 환경과 자연어 지시의 결합을 통해 강화학습(RL)의 성능을 높일 수 있으나, 대규모 모델 적용 시 막대한 계산 자원과 메모리 제약이라는 어려움이 있다. 본 연구는 경량 VLM 인 SmolVLM 에 파라미터 효율적 파인튜닝(LoRA)과 명시적 추론(CoT)을 결합하여 효율적인 RL 에이전트를 제안한다. GymCards 환경에서 Proximal Policy Optimization(PPO)으로 에이전트를 학습하고 정책 생성 과정에 CoT 를 통합해 성능과 해석 가능성을 높였다. 실험 결과, 제안된 SmolVLM-CoT-LoRA 모델은 CoT 가 없는 SmolVLM-LoRA 와 제로샷 평가 대비 우수한 성능과 효율성을 보였으며, CoT 의 효과성과 RL 파인튜닝의 필요성을 확인했다. LoRA 는 제한된 GPU 자원에서도 효과적인 학습을 가능하게 하였으며, 본 연구는 자원 제약 환경에서 실용적이고 효과적인 경량 VLM 기반 RL 접근법을 제시한다.

1. 서론

강화학습(RL)은 에이전트가 환경과 상호작용하며 최적의 행동 정책을 학습하는 방법으로, 복잡한 시각 정보와 언어 기반 지시를 처리하는 현실적 작업 환경에서 큰 가능성을 제시하고 있다. 최근 RL 연구는 시각적 입력과 언어 명령을 동시에 이해하여 보다 일반화된 의사결정을 수행하는 방향으로 확장되고 있으나, GPT-4V 나 LLaVA 와 같은 대규모 Vision-Language 모델(VLM)을 RL 환경에 적용하기 위해서는 막대한 계산 자원과 GPU 메모리 요구라는 현실적 어려움이 존재한다[1]. 특히 RL 학습의 반복적이고 탐색적인 특성은 이러한 자원 부담을 더욱 심화시킨다.

본 연구는 이러한 한계를 극복하기 위해 경량 VLM 인 SmolVLM [3]을 기반으로 RL 에이전트를 개발하고, GymCards 라는 시각 기반 숫자 추론 환경에서 그 의사결정 능력을 평가한다. 특히, 에이전트의 추론 능력과 해석 가능성을 높이기 위해 Chain-of-Thought (CoT) 프롬프팅[4]을 행동 생성 과정에 통합하며, 경량 모델의 파라미터 효율적 파인튜닝을 위해 Low-Rank Adaptation (LoRA) [7] 기법을 적용한다. 이 접근법은 제한된 자원 내에서 VLM 기반 RL 에이전트의 실용성을 높이는 것을 목표로 한다. Proximal Policy Optimization (PPO) 알고리즘[8]을 사용하여 SmolVLM-CoT-LoRA 에이전트를 학습시키고, GymCards 환경에서의 성능과 CoT 및 LoRA 의 효과를 분석한다.

2. 관련 연구

본 연구는 VLM 기반 RL 에이전트, CoT 추론, 그리고 LoRA 기법과 관련된다.

VLM 기반 RL 에이전트는 시각 및 언어 이해 능력을 활용하여 환경과 상호작용하는 에이전트를 개발하는 연구 분야이다. RT-1 [5]이나 Gato [6] 같은 대규모 모델 기반 연구는 가능성을 보여주었으나, 계산 비용과 학습 안정성 문제가 있다. 본 연구는 경량 VLM 을 사용하여 이러한 문제를 완화하고자 한다.

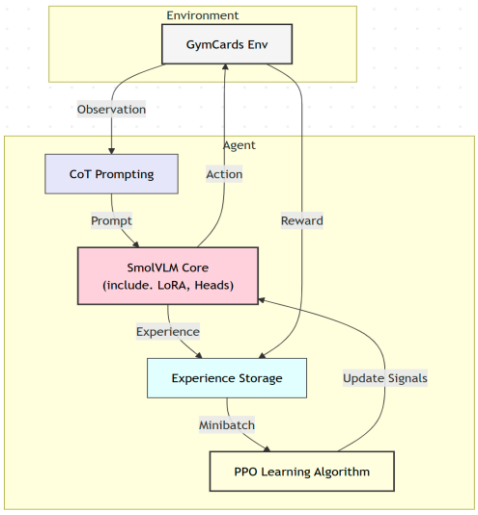
CoT 는 모델이 명시적인 중간 추론 단계를 생성하

도록 유도함으로써 복잡한 문제를 효과적으로 해결할 수 있도록 돕는 기법이다. 이를 강화학습(RL)에 적용하면 에이전트의 행동 결정 과정에 대한 해석 가능성을 높일 뿐 아니라, 성능 향상에도 기여할 것으로 기대된다. 본 연구에서는 이러한 CoT 기법을 VLM 기반의 RL 에이전트 행동 생성 과정에 직접 통합하여, 에이전트의 추론 능력과 의사결정 성능을 동시에 강화하고자 한다.

LoRA 는 가중치 업데이트를 저차원 행렬로 근사하여 학습 파라미터 수를 크게 줄인다. 본 연구는 LoRA 를 SmolVLM 에 적용하여, 제한된 자원으로 경량 VLM 을 RL 작업에 효과적으로 파인튜닝하는 것을 목표로 한다.

3. 시스템 설계 및 학습 프레임워크

본 연구는 경량 VLM, CoT 프롬프팅, LoRA 파인튜닝, 그리고 PPO 강화학습 알고리즘을 결합하여 GymCards 환경에서 효율적인 의사결정 에이전트를 구축하는 것을 목표로 한다. (그림 1)은 제안하는 강화학습 에이전트의 주요 구성 요소와 로직 흐름을 도식화한 것이다. 시스템은 크게 환경(Environment)과 에이전트(Agent)로 구성되어 있다.



(그림 1) 전체 시스템 아키텍처

상호작용 루프 (Interaction Loop): GymCards 환경은 현재 상태에 대한 관측(Observation)을 제공한다. 이 관측은 CoT Prompting 모듈로 전달되어, VLM 이 추론 과정을 포함한 응답을 생성하도록 유도하는 프롬프트로 변환된다. LoRA 어댑터와 PPO 헤드(Value/Action)가 통합된 SmolVLM Core 는 이 프롬프트를 입력 받아 환경에 적용될 행동(Action)을 생성한다. 생성된 행동은 다시 환경에 영향을 미치게 된다.

학습 루프 (Learning Loop): 환경으로부터 받은 보상(Reward)과 SmolVLM Core 에서 생성된 경험 관련 데이터(Experience, 예: 상태 표현, 행동 등)는 Experience Storage 에 수집된다. PPO Learning Algorithm 은 저장된 경험 데이터로부터 미니배치(Minibatch)를 샘플링하여 정책 및 가치 손실을 계산한다. 계산된 손실을 기반으로 생성된 업데이트 신호(Update Signals, 그래디언트)는 SmolVLM Core 내의 LoRA 어댑터 파라미터와 PPO 헤드 파라미터를 업데이트하는데 사용되어 정책을 개선한다.

3.1. LoRA 기반 Actor-Critic 모델

에이전트 정책 및 가치 함수의 핵심은 경량 VLM 인 SmolVLM[2]을 기반으로 구축된 통합 모델(SmolVLM Core)이다. 전체 모델 파인튜닝에 따르는 높은 계산 비용과 메모리 요구량을 완화하기 위해, 파라미터 효율적 파인튜닝 기법인 LoRA 를 적용한다. LoRA 는 사전 학습된 SmolVLM 의 대부분 가중치를 동결시킨 상태에서, 특정 트랜스포머 레이어(어텐션 및 피드포워드 관련 프로젝션) 대상으로 소수의 학습 가능한 저차원 행렬(어댑터)을 추가한다. 이 접근 방식은 최적화해야 할 파라미터 수를 대폭 감소시켜 GPU 메모리 사용량을 줄이고 학습 효율성을 높인다. 중요하게도, PPO 학습에 필요한 가치 예측 헤드(Value Head)와 행동 정책 헤드(Action Head)가 이 LoRA 적용된 SmolVLM 모델 내에 직접 통합되어, 단일 모델이 Actor 와 Critic 역할을 동시에 수행하는 통합 Actor-Critic 구조를 형성한다.

3.2. CoT 기반 행동 생성

에이전트의 단계별 추론 능력을 활용하고 행동 결정 과정의 해석 가능성을 높이기 위해 행동 생성 과정에 CoT 프롬프팅 기법을 적용한다. 에이전트는 환경으로부터 현재 관측(Observation, 예: 카드 이미지)을 받으면 다음 단계를 거쳐 행동을 결정한다: (1) 현재 관측과 게임 상태 정보, 그리고 CoT 를 유도하는 템플릿을 결합하여 상세한 프롬프트를 구성한다. 이 프롬프트는 모델이 최종 행동을 선택하기 전에 생각의 연쇄과정(thoughts)을 먼저 생성하도록 지시한다. (2) 구성된 프롬프트를 LoRA 가 적용된 통합 SmolVLM 모델의 generate 메소드에 전달하여, 추론 과정과 최종 선택 행동이 포함된 텍스트를 생성한다. (3) 생성된 텍스트에서 최종 행동에 해당하는 부분("[Action]: ...")을 파싱하여 환경에서 사용할 수 있는 정수 형태의 행동(Action)으로 변환한다. 이 CoT 기반 행동 생성은 에이전트가 환경과 상호작용하며 경험 데이터를 수집하는데 사용된다.

3.3. PPO 기반 강화학습

에이전트는 GymCards 환경과 상호작용하며 수집한 경험 데이터(Experience Storage 저장)를 사용하여 PPO 알고리즘으로 정책을 학습한다. 학습 과정은 다음과 같다: (1) Experience Storage 에서 미니배치 데이터를 샘플링한다. (2) 각 미니배치 샘플에 대해 통합 Actor-Critic 모델로부터 현재 상태에 대한 주요 특징 표현을 추출한다. (3) PPO 업데이트 루프 내부에서 추출된 hidden state 를 기반으로 통합 모델 내의 Value Head 와 Action Head 를 직접 호출하여 현재 정책의 가치 추정치(values)와 행동 로짓(action_logits)을 계산한다. (4) Generalized Advantage Estimation (GAE)[9]를 통해 계산된 Advantage 와 함께, PPO 의 목적 함수(클리핑 된 정책 손실, 가치 함수 손실, 엔트로피 보너스 항의 조합)를 구성한다. (5) 계산된 최종 손실에 대해 역전파를 수행하여 그래디언트를 얻는다. (6) Optimizer(Adam)는 이 그래디언트를 사용하여 LoRA 어댑터 파라미터와 PPO 헤드 파라미터만 업데이트한다.

4. 실험 및 성능 평가

SmolVLM-CoT-LoRA 에이전트의 성능을 GymCards 환경에서 평가하고, CoT 와 LoRA, 그리고 RL 파인튜닝 자체의 효과를 분석한다.

4.1. 실험 환경

본 연구의 실험은 시각적 추론 능력이 요구되는 GymCards 의 NumberLine-v0 환경에서 진행한다. 에이전트는 카드 이미지를 보고 목표 숫자에 도달하기 위해 '+' 또는 '-' 행동을 선택해야 한다.

실험 대상 모델은 3 가지로 구성한다. SmolVLM 에 CoT 프롬프팅과 LoRA 를 적용해 PPO 로 파인튜닝한 제안 모델(SmolVLM-CoT-LoRA), 성능 비교를 위해 제안 모델에서 CoT 프롬프팅을 제외하고 LoRA 로만 파인튜닝한 SmolVLM-LoRA (w/o CoT) 모델, 그리고 별도의 RL 파인튜닝 없이 사전 학습된 SmolVLM-500M 모델의 SmolVLM-Zero-Shot Eval 모델을 사용한다. 이를 통해 RL 파인튜닝의 효과, CoT 적용의 이점, 그리고 LoRA 기반 학습의 효율성을 종합적으로 검증한다.

4.2. 성능 평가

평가 지표로는 학습 효율성(최대 GPU 메모리 사용량), 최종 추론 성능(에피소드 성공률, 평균 에피소드 보상, 평균 에피소드 스텝 수), 추론 효율성(추론 지연 시간, 추론 FPS, 최대 GPU 메모리 사용량)을 사용한다.

<표 1> 최종 추론 성능

모델	에피소드 성공률 (%)	평균 에피소드 보상	평균 에피소드 스텝 수
SmolVLM-CoT-LoRA (제안)	71.5	1.03	5.26
SmolVLM-LoRA (w/o CoT)	54.0	-4.06	5.82
SmolVLM (Zero-Shot Eval)	43.0	-5.27	20.45

표 1 은 학습 완료 후 최종 추론 성능을 비교한 결과이다. RL 파인튜닝을 적용한 SmolVLM-LoRA 모델은 Zero-Shot 모델 대비 성공률이 11% 향상되었고 (54.0% vs 43.0%), 평균 보상도 소폭 개선되었으며(-4.06 vs -5.27), 특히 평균 에피소드 스텝 수가 20.45 에서 5.82 로 크게 단축되어 문제 해결 효율성이 높아졌음을 확인했다. 제안 모델(SmolVLM-CoT-LoRA)은 SmolVLM-LoRA 모델보다 성공률이 17.5% 더 높았으며 (71.5% vs 54.0%), 평균 에피소드 보상도 양수 값을 기록하고 평균 스텝 수도 단축되어 CoT 의 긍정적인 효과를 보여주었다.

4.3. 요소별 효과 분석

RL 파인튜닝의 효과: Zero-Shot 모델 은 43.0%의 성공률을 보였으나, 평균 보상(-5.27)이 낮고 평균 스텝 수(20.45)가 길어 비효율적이었다. 반면, 동일 모델을 LoRA 로 파인튜닝한 모델(SmolVLM-LoRA)은 성공률(54.0%), 평균 보상(-4.06), 평균 스텝 수(5.82) 등 모든 지표에서 명확한 개선을 보였다. 이는 NumberLine 과 같은 특정 순차적 의사결정 문제에서 환경과의 상호작용을 통한 RL 파인튜닝이 성능 향상에 중요한 부분임을 입증했다.

CoT 의 효과: 제안모델(SmolVLM-CoT-LoRA)은 CoT 프롬프팅을 추가하여 명시적인 추론 과정을 거치도록 설계되었다. 그 결과, CoT 가 없는 SmolVLM-LoRA 모델 (성공률 54.0%) 대비 최종 추론 성공률이 17.5% 향상된 71.5%를 달성했다. 또한, 학습 과정에서도 CoT 를 통해 더 빠른 학습 속도와 높은 샘플 효율성을 확보하여, 같은 스텝 수의 동일한 학습 조건에서 더 높은 성능에 도달했다. 이는 CoT 가 에이전트의 추론 능력과 학습 효율성을 동시에 개선함을 나타낸다.

4.4. 효율성 분석

<표 2> 학습 및 추론 효율성 지표 비교

모델	추론 지연 시간(ms/step)	추론 속도(FPS)	최대 GPU 메모리(GB)
SmolVLM-CoT-LoRA (제안)	195.0	5.1	13.6(학습) / 1.54(추론)
SmolVLM-LoRA (w/o CoT)	170.4	5.9	13.4(학습) / 1.45(추론)
SmolVLM (Zero-Shot Eval)	141.0	7.09	1.44(추론)

학습 효율성: 표 2 에서 볼 수 있듯이, LoRA 를 사용한 SmolVLM-LoRA 모델과 SmolVLM-CoT-LoRA 모델은 학습 중 각각 약 13.4 GB, 13.6 GB 의 최대 GPU 메모리를 사용했다. 이는 전체 모델 파인튜닝(Full FT)

시 요구되는 막대한 메모리(수십 GB 이상 예상)에 비해 현저히 낮은 수치로, LoRA 가 제한된 자원 하에서 VLM 기반 RL 에이전트의 학습을 실현 가능하게 함을 보여준다.

추론 효율성: SmolVLM-Zero-Shot Eval 모델이 가장 빠른 추론 속도(141.0 ms/step, 7.09 FPS)를 보였으나, 이는 4.2 절에서 확인된 낮은 성능과의 Trade-off 이다. SmolVLM-LoRA 모델은 추론 시 약 170.4 ms/step (5.9 FPS)의 속도를 보였으며, Zero-Shot 모델과 유사한 약 1.45 GB 의 낮은 추론 메모리를 사용했다. 이는 LoRA 가중치를 병합한 후에도 추론 오버헤드가 크지 않음을 의미한다. 제안모델(SmolVLM-CoT-LoRA)은 CoT 프롬프팅을 통합함에 따라 SmolVLM-LoRA 모델보다 추론 지연 시간이 약 25ms 증가한 195.0 ms/step (5.1 FPS)을 기록했으나, 이는 4.2 절에서 확인한 상당한 성능 향상으로 충분히 상쇄될 수 있는 수준이다.

5. 결론 및 향후 연구

5.1. 결론

본 연구는 경량 VLM 인 SmolVLM 을 기반으로, CoT 추론과 LoRA 파인튜닝을 결합하여 시각적 추론 능력이 필요한 강화학습(RL) 환경인 GymCards 에서 효과적으로 작동하는 에이전트를 개발하는 것을 목표로 했다. 실험 결과는 제안된 SmolVLM-CoT-LoRA 접근 방식이 기존 방식들에 비해 여러 측면에서 우수함을 입증했다.

주요 연구 결과는 다음과 같다. 첫째, RL 파인튜닝은 사전 학습된 VLM 을 특정 순차적 의사결정 작업에 적응시키는 데 필수적이며, 파인튜닝 없이 Zero-Shot 으로 평가했을 때는 매우 낮은 성능을 보였다. 둘째, CoT 프롬프팅을 정책 생성 과정에 통합함으로써 에이전트의 성능(최종 성공률)과 샘플 효율성(학습 속도)을 유의미하게 향상시킬 수 있음을 확인했다. 이는 CoT 가 명시적 추론 과정을 통해 복잡한 의사결정 문제 해결에 기여함을 시사한다. 셋째, LoRA 를 적용하여 파라미터 효율적인 파인튜닝을 수행함으로써, 제한된 GPU 메모리 환경에서도 경량 VLM 기반 RL 에이전트의 학습을 성공적으로 완료할 수 있었다. 이는 전체 파인튜닝 시 메모리 부족으로 학습이 불가능했던 것과 대조적으로, LoRA 가 VLM 기반 RL 연구의 실용성과 접근성을 크게 높이는 핵심 기술임을 보여준다.

결론적으로, 본 연구는 경량 VLM, CoT, LoRA, 그리고 PPO 를 효과적으로 결합함으로써 자원 제약 하에서도 시각적 추론 능력을 갖춘 효율적인 RL 에이전트를 개발할 수 있음을 실증적으로 보여주었다.

5.2. 향후 연구

향후 연구에서는 본 접근법을 다른 복잡한 과제를 통해 더욱 확장 · 개선하고자 한다. 예를 들어, 본 논문에서 다루지 않은 GymCards 환경의 Points24 나 Blackjack 과제 혹은 Embodied AI 환경에 SmolVLM 기반 에이전트를 적용하여 성능을 평가할 수 있다.

더불어, 경량 VLM 의 장점인 추론 속도와 배포 용이성을 활용하여 실제 어플리케이션에 적용하는 연구가 필요하다. 제안한 방식으로 학습된 SmolVLM 에이전트는 적은 수의 GB 메모리로도 동작 가능하여 로컬 디바이스에서 수행되는 다양한 태스크에 투입될 잠재력이 있다. 이러한 응용을 염두에 두고, 소형이면서도 우수한 성능과 효율성을 가진 멀티모달 에이전트를 현실화하는 것이 본 연구의 장기적인 목표이다.

참고문헌

[1] Yuexiang Zhai et al., “Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning”, arXiv:2405.10292, 2024

[2] Andrés Marafioti et al., “SmolVLM: Redefining small and efficient multimodal models”, arXiv preprint arXiv:2504.05299, 2025

[3] HuggingFace, “HuggingFaceTB/SmolVLM-500M-Instruct Model Card,” Hugging Face, 2024.

[4] Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” NeurIPS, 2022

[5] Brohan et al., “RT-1: Robotics Transformer for Real-World Control at Scale”, arXiv:2212.06817, 2022

[6] Reed et al., “A Generalist Agent”, arXiv:2205.06175, 2022

[7] Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models”, arXiv:2106.09685, 2021

[8] Schulman et al., “Proximal Policy Optimization Algorithms”, arXiv:1707.06347, 2017

[9] Schulman et al., “High-Dimensional Continuous Control Using Generalized Advantage Estimation”, arXiv:1506.02438, 2016