개선된 시각적 어텐션 어댑터를 통한 로봇 제어 성능 향상

김지원, 김민성, 박태진, 이재구*
국민대학교 컴퓨터공학과
*jaekoo@kookmin.ac.kr

Enhancing Robot Control with Advanced Visual Attention Adapter

Jiwon Kim, Minsung Kim, Taejin Park, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

본 논문은 추가적인 전문가 시연 데이터 없이, 주어진 시각 정보를 효과적으로 활용해 행동 복제 알고리즘의 로봇 제어 성능을 향상시킬 수 있는 방법을 제안한다. CBAM 은 채널 어텐션과 공간 어텐션을 통해 입력의 중요한 특징을 강조하고, 불필요한 특징을 억제함으로써 제어 정책이 시각 정보를 더욱 효과적으로 이해할 수 있게 한다. 따라서 이미지 인코더에 CBAM 을 추가하여 로봇 제어 정책 학습 시 중요한 시각적 정보를 더 활용할 수 있도록 하였다. 또한 CBAM 을 어댑터 방식으로 사용하여 매개변수 효율성을 높였다. 그 결과, 기존 연구 대비 약 11M 적은 양의 학습 가능 매개변수를 사용하면서도 두 개의 시뮬레이션 과업에서 각각 3%와 6%의 성능 향상을 이루었다.

Ⅰ. 서론

로보틱스 산업의 발전에 따라 효과적인 로봇 제어 정책 학습 방법이 활발하게 연구되고 있다[1]. 그 중 모방 학습(Imitation Learning)에 속하는 행동 복제(Behavior Cloning) 알고리즘은 사전에 수집된 전문가의 행동을 모방하도록 학습된다. 행동 복제 알고리즘의 대표적인 예인 ACT (Action Chunking with Transformers)[2]는 ImageNet1K[3]로 사전 학습된 ResNet[4]을 이미지 인코더(Image Encoder)로 사용하며, 시연 데이터 집합에 대해 미세 조정(Fine Tuning)하여 사용한다. 하지만 미세 조정을 거친 후에도 삽입(Insertion)과 같이 정밀한 조작이 필요한 과업에서는 여전히 낮은 성능을 보인다.

본 논문에서는 이러한 문제를 해결하기 위해 ACT 의 이미지 인코더에 입력의 중요 특징을 강조하고, 불필요한 특징을 억제하는 CBAM(Convolutional Block Attention Module)[5]을 더해주는 방법을 제안한다. 효율적인 학습을 위해 적은 양의 매개변수를 추가하고, 사전 학습된 분포를 최대한 활용하는 어댑터(Adapter)[6] 방식을 통해 CBAM 을 사용하였다. 실험을 통해 기존 모델과의 성공률 및 학습 가능 매개변수 수를 비교함으로써, 제안 방법의 성능 향상과 매개변수 효율성을 입증하였다.

Ⅱ. 방법

본 논문에서 제안한 모델 구조는 [그림 1]과 같다. 기준 모델인 ACT 는 트랜스포머(Transformer) 기반 CVAE(Conditional Variational AutoEncoder) 구조로, [그림 1.(a)]의 인코더(Encoder)와 [그림 1.(b)]의 디코더(Decoder)로 구성되어 있다.

ACT 의 인코더는 [CLS] 토큰(Token)과 현재의 관절(Joint) 값, 위치 임베딩(Positional Embedding)이 더해진 행동 시퀀스(Action Sequence)를 입력 받아 스타일 변수(Style Variable)인 z를 반환한다.

ACT 디코더(Decoder)는 ResNet 기반 이미지 인코더와 트랜스포머 인코더 및 디코더로 구성된다. 디코더는 입력 이미지 특징 맵(Feature Map), 현재 관절 값, 인코더에서 반환된 z를 입력으로 받아 행동 시퀀스를 생성한다. 행동 시퀀스는 t 시간 단계에서 로봇의 행동(Action)을 k 개로 묶은 것을 의미하며 [그림 1.(c)]와 같다.

ACT 의 이미지 인코더는 모델의 입력 이미지로부터 환경의 시각적 특징을 추출하여, 이를 로봇의 행동을 예측하는 데 활용한다. 하지만 추론 시 과업에서 사용되는 객체의 위치는 변화할 수 있으므로, 로봇 제어 정책의 성능 일반화를 위해 시각 정보의 효과적인 활용이 필수적이다. 과업 수행에서 모든 시각적 정보가 동일한 중요도를 갖는 것이 아니므로, 중요한 특징은 강조하고 불필요한 특징은 억제할 필요가 있다.

이를 위해 우리는 이미지 특징 맵 F 에 채널 어텐션(Channel Attention) M_c 와 공간 어텐션(Spatial Attention) M_s 를 순차적으로 적용하는 CBAM 을 추가하여, 모델이 중요한 특징과 위치에 집중하도록 하였다. CBAM 은 식 (1)과 같이 요약된다. 수식의 $\otimes$은 요소별 곱셈을 의미한다.

$$F' = M_c(F) \otimes F \quad (1a)$$
$$F'' = M_s(F') \otimes F' \quad (1b)$$

CBAM 의 채널 어텐션 M_c는 식 (2)과 같이 표현된다. 입력 특징 맵 F에 평균 풀링(Average Pooling)과 최대 풀링(Max Pooling)통해 두 개의 특징 맵을 얻는다. 이를 다층 퍼셉트론(Multi-Layer Perceptron; MLP)에 전달하고, 결과를 더해 시그모이드 함수를 거친다. 이후 식 (1a)를 통해 입력의 '무엇'에 집중해야 하는지 강조된 F'을 얻는다.

$$M_c(F) = \sigma\left(MLP(AvgPool(F)) + MLP(MaxPool(F))\right) \quad (2)$$

이어서 F'에 식 (3)과 같이 표현되는 공간 어텐션 M_s를 적용한다. 평균 풀링과 최대 풀링을 통해 생성된 두 개의 특징 맵을 7×7 의 합성곱 필터(Convolutional Filter)를 통해 합치고, 시그모이드 함수를 통과한다. 이후 식 (1b)를 통해 입력의 '어디'에 집중해야 할지 강조된 F'' 을 얻는다.

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \quad (3)$$

CBAM 의 적용과 더불어 우리는 매개변수 효율적인 학습을 위해 어댑터 방식을 채택하였다. 어댑터는 소규모의 매개변수를 더해주고 해당 부분만 학습을 진행함으로써, 전체 매개변수를 학습하는 기존의 미세

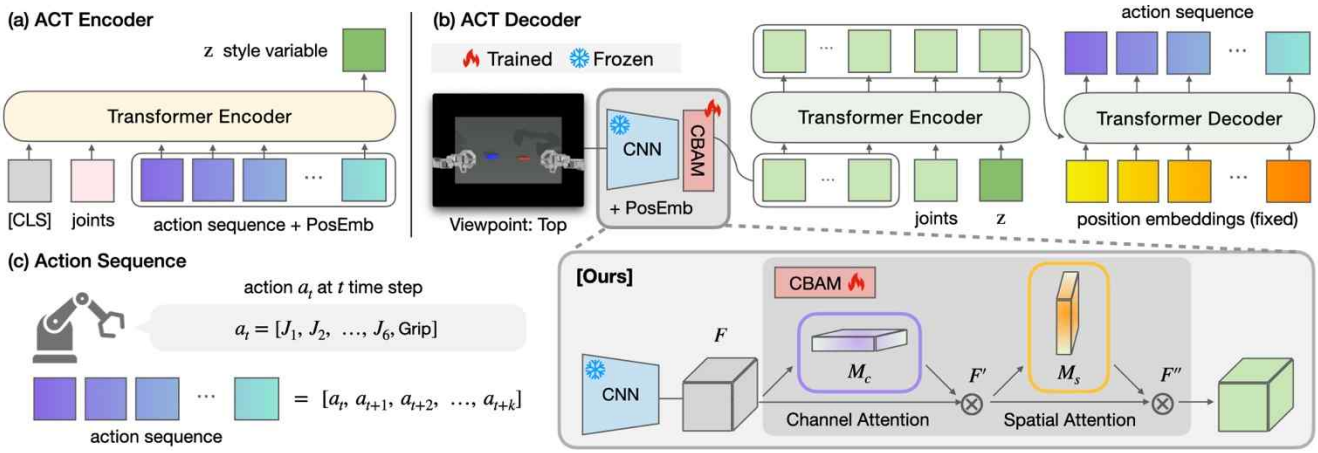

그림 1. 전체 구조도

조정 방식과 유사한 성능을 내면서도 매개변수 효율성이 좋다. 또한 다양한 모델과 쉽게 결합할 수 있어 여러 하위 작업으로의 확장성이 좋다. 본 논문은 이러한 어댑터의 장점에서 착안하여 이미지 인코더 ResNet18 의 모든 매개변수를 학습하는 기존 ACT 와 달리, 이미지 인코더를 동결하고 소규모의 매개변수를 가진 CBAM 만을 학습하는 방식을 택하였다.

Ⅲ. 실험

본 논문에서는 CBAM 의 채널 어텐션과 공간 어텐션이 각각 다른 기능을 수행하므로, 각각의 어텐션을 독립적으로 사용한 경우를 포함하여 실험을 진행하였다. 실험에 사용된 모델은 기존 모델인 ACT, 채널 어텐션만 적용한 ACT+ Channel, 공간 어텐션만 적용한 ACT+ Spatial, 두 어텐션을 모두 사용한 Ours(ACT+ CBAM)으로 총 네 가지이다. [표 1]은 해당 모델들에 대한 실험 결과를 요약한 것으로, 성공률은 3 개의 임의 초기값을 사용하여 학습한 뒤, 각 50 번의 정책 평가를 수행한 평균 값을 나타낸다.

ACT 에서 제시한 두 개의 시뮬레이션 과업에 대해 실험을 진행하였으며, 과업 별로 50 개의 시연 데이터를 사용해 학습하였다. 두 과업은 오른팔이 큐브를 집어 왼팔로 전달하는 큐브 전달(Transfer Cube)과 두 팔이 소켓과 핀을 잡아 결합하는 삽입이다. 평가 시 모델의 일반화 성능 평가를 위해 초기 물체의 위치를 무작위로 변경한다. 각 과업은 전달 또는 결합이 완벽히 이루어지면 성공으로 간주하며, 50 번의 시도 중 성공 횟수로 성공률을 나타낸다.

[표 1]의 실험 결과를 통해 공간 어텐션이 채널 어텐션보다 더 유의미한 성능 향상을 이끌어냄을 확인할 수 있었다. 또한 두 어텐션을 독립적으로 사용하는 것보다 순차적으로 사용하는 것이 더 효과적임을 확인하였다. 기존 방법보다 약 11M 적은 양의 학습 가능 매개변수를 사용하면서도 더 높은 성능을 달성하여 매개변수 효율성을 입증하였으며, 이를 통해 모델에게 '무엇'과 '어디'에 집중해야 할지를 알려주는 우리의 방법이 로봇 제어 정책에 효과적임을 확인하였다.

Ⅳ. 결론

본 논문에서는 이미지 인코더에 CBAM 어댑터를 추가해 모델이 입력 이미지를 더 효과적으로 이해할 수 있도록 함으로써, 로봇 제어 정책의 성능을 향상시키는 방법을 제안하였다. CBAM 어댑터를 통해 중요한 부분에 집중하고, 불필요한 특징을 억제함으로써 추가적인 전문가 시연 데이터 없이도 성능을 개선하였다.

표 1. 두 과업의 성공률 및 학습 매개변수 수 비교

Method	Success Rate (%)		Param (M)
	Transfer Cube	Insertion	
ACT	93	55	83.92
ACT + Channel	84	52	72.79
ACT + Spatial	91	60	72.79
Ours	**96**	**61**	**72.79**

또한 기존 모델과 비교해 약 11M 적은 양의 학습 가능 매개변수를 사용하면서도 두 가지 과업에서 각 3%, 6%의 성능 향상을 달성함으로써, 본 논문의 방법이 매개변수 효율적임을 입증하였다. 하지만 이러한 성능 개선은 특정 환경에서 얻어진 결과이므로, 향후 연구에서는 실제 환경에서의 성능 평가와 다양한 모방 학습 방법과의 결합을 통해 더 넓은 범위에서의 효과를 검증하고자 한다.

ACKNOWLEDGMENT

본 연구는 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167194,미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능).

참 고 문 헌

[1] Ma, Yueen, et al. "A Survey on Vision-Language-Action Models for Embodied AI." arXiv preprint arXiv:2405.14093. 2024.

[2] Zhao, Tony Z., et al. "Learning fine-grained bimanual manipulation with low-cost hardware." arXiv preprint arXiv:2304.13705. 2023.

[3] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2009.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016.

[5] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[6] Chen, Zhe, et al. "Vision transformer adapter for dense predictions." arXiv preprint arXiv:2205.08534. 2022.