Transactions of KSAE, Vol. 31, No. 12, pp.949-957 (December, 2023)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2023.31.12.949

강한 Data Augmentation과 Contrastive Learning을 통한 이미지 기반 강화학습 일반화 성능 향상 연구

박 상 훈[1] · 유 진 우*[2]

국민대학교 자동차공학전문대학원[1] · 국민대학교 자동차IT융합학과[2]

Improved Generalization Performance of Image-based Reinforcement Learning through Strong Data Augmentation and Contrastive Learning

Sanghoon Park[1] · Jinwoo Yoo*[2]

[1]Graduate of Automotive Engineering, Kookmin University, Seoul 02707, Korea
[2]Department of Automobile and IT Convergence, Kookmin University, Seoul 02707, Korea
(Received 4 August 2023 / Revised 6 September 2023 / Accepted 6 September 2023)

Abstract : In this paper, we are proposing a convolutional contrastive learning method that can improve the generalization performance of image-based reinforcement learning. To do this, methods on augmenting input images were mainly used. However, strong augmentation hinders the stability of reinforcement learning. Thus, by gradually increasing the random image mixing ratio during training, a reinforcement learning agent is not affected by strong data augmentation. At the same time, the effect on generalization performance is maximized. Experiments on DM Control test environments have shown that the proposed method outperforms the existing studies on the generalization of image-based reinforcement learning.

Key words : Deep learning(딥러닝), Reinforcement learning(강화학습), Data augmentation(데이터 증강), Generalization (일반화), Contrastive learning(대조학습), Self-supervised learning(자기지도학습)

1. 서 론

알파고의 등장 이후 심층 강화 학습의 가능성이 입증됨에 따라, 자율주행, 자동화 로봇 등 다양한 분야에 강화학습이 활발히 적용되고 있다.[1,2] Fig. 1과 같은 강화 학습과 심층 신경망의 조합은 이미지와 같은 고차원 데이터를 사용하여 강화학습을 수행할 수 있게 한다.[3] 영상으로부터 다양한 게임(보드 게임[4]과 비디오 게임[5,6])을 하는 방법을 배우는 것, 가상 환경의 카메라 프레임을 통한 차량 제어,[7] 그리고 실제 세계에서의 물체를 줍는 로봇[8] 등이 그 예시이다. 그러나 이미지와 같은 고차원 입력 데이터를 사용하면 데이터 효율성이 상대적으로 낮다.[9,10] 즉, 동일한 수의 데이터로 학습해도 저차원 상태 벡터를 사용할 때보다 고차원 이미지를 사용할 때 더 낮은 학습 성능이 나타난다. 많은 연구 중 CURL[11]은 대조학습을 통해 입력 프레임 간의 유사성을 학습하여 이미지 데이터 효율성을 높였으며, 이는 쿼리와 키를 대조하면서 이미지에서 더 풍부한 Feature를 추출하는 자기지도학습을 사용한 방법이다.

그러나 학습 환경에서의 과적합으로 인해 테스트 환경의 사소한 배경 변화에도 강화학습 성능이 저하되는 문제가 존재한다. 즉, 학습 환경과 유사하지만 Action 선택에 영향을 주지 않는 주위 배경이 다른 테스트 환경에서는 대조학습을 통해 향상시킨 데이터 효율성이 보장되지 않으며, 이를 이미지 기반 심층 강화학습에서는 일반화(Generalization) 문제라고 한다.[12,13]

강화학습에서 입력 이미지 데이터는 일반적으로 모델이 관측하지 않은 테스트 환경에서도 강건한 성능을 보이기 위해 증강(Augmentation)된다.[14] 증강을 통해 다양하게 변화된 입력 이미지를 사용하면 학습 환경에 대한 과적합을 방지하는 데 도움이 될 수 있다. 또한 데이터

*Corresponding author, E-mail: jwyoo@kookmin.ac.kr

증강은 대조학습에 필수적으로 사용되는 방법이다. 데이터 증강 강도가 강할수록 더 효과적으로 대조학습이 진행되며 강화학습의 강건한 성능에 대한 보조효과가 커지지만, 입력 이미지의 큰 변화가 강화학습 자체를 방해하기 때문에 강한 강도의 증강은 사용이 제한적이다.[15] 강력한 증강으로 인한 입력 데이터의 큰 변화가 강화학습을 방해하지 않는다면, 대조학습이 강화학습에 대한 보조 효과를 극대화하여 일반화 성능을 크게 향상시킬 수 있다.[16]

본 논문에서는 이미지 기반 강화학습의 일반화 성능을 위해 대부분의 강화학습 프레임워크에 추가할 수 있는 간단한 컨볼루션 대조학습 아키텍처를 제안한다. 또한 증강 강도와 관련된 Trade-off를 극복하여 강화학습과 대조학습 모두에 강한 증강을 사용하기 위한 학습 방법을 제안한다. (i) 학습 초기에는 강한 데이터 증강 없이 원본 이미지 데이터만으로 강화학습과 대조학습이 수행된다. (ii) 이후 랜덤 컨볼루션과 같은 강한 데이터 증강이 적용된 이미지를 원본 이미지와 혼합하여 강화학습 및 대조학습에 사용한다. (iii) 학습이 진행됨에 따라 랜덤 이미지의 혼합 비율을 점점 더 크게 하여 강화학습 에이전트는 점점 더 강한 증강 이미지로부터 학습한다. 제안한 방법을 통해 대조학습은 점진적으로 강해지는 증강 이미지를 사용함으로써 이미지 기반 강화학습에 더 큰 보조 효과를 유도하여 일반화 성능을 크게 향상시킬 수 있다.

본 논문의 가장 큰 기여 중 하나는 학습 전반에 걸쳐 동일하게 강한 데이터 증강이 입력 이미지에 일관되게 적용될 때보다 제안한 방법에서 더 효과적으로 강한 증강 이미지를 사용한다는 것이다. 또한, 본 논문에서는 강화학습에서 이미지 데이터를 효율적으로 사용하는 방법에 대한 새로운 방법을 제시한다. 일반화 성능 검증을 위해 Fig. 2와 같이 DMControl(Deep Mind Control) 시뮬레이션 환경의 두 가지 모드(Color-hard, Video-easy)에서 실험을 진행하였다. 실험 결과, 제안한 학습 방법은 배경이나 색상이 정적 및 동적으로 변화하는 테스트 환경 모두에서 기존의 강화학습 일반화 성능을 위한 연구들을 크게 능가함을 확인할 수 있다.

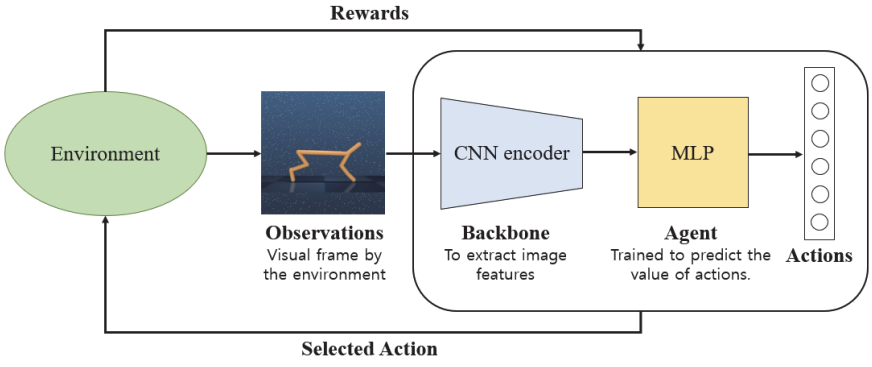

Fig. 1 Image-based reinforcement learning architecture

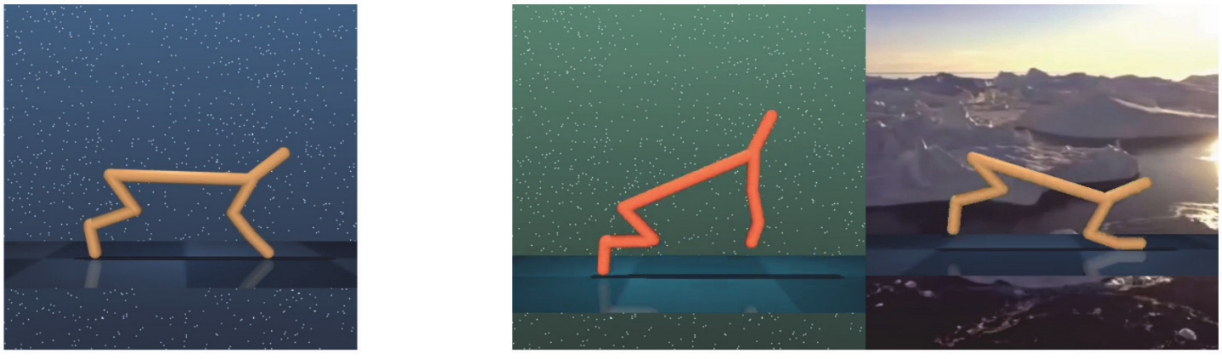

Fig. 2 Left: Training environment of DMControl. Right: Test environments of DMControl generalization benchmark[17] (Color-hard and Video-easy mode)

2. 관련 연구

강화학습에서 이미지 데이터를 효과적으로 사용하기 위해 기존 이미지 기반 딥러닝 기법을 강화학습에 융합하는 연구가 활발하게 진행되고 있다. 본 논문에서는 강화학습에서 사용하는 이미지 데이터의 효율을 향상시키기 위해 기존 강화학습 아키텍처에 자기지도학습 방법 중 하나인 대조학습을 융합하고, 동시에 일반화 성능을 위해 입력 이미지에 랜덤 컨볼루션 레이어를 통과시켜 입력 이미지를 강하게 증강시킨다. 본 절에서는 제안한 아키텍처를 구성하고 있는 강화학습 및 대조학습 방법과 입력 이미지 증강을 위한 랜덤 컨볼루션 방법을 설명한다.

2.1 SAC(Soft Actor Critic)

이미지 기반 강화학습 알고리즘으로 보상의 합에 대한 기댓값을 최대화하는 Off-policy actor-critic 강화학습 알고리즘인 SAC를 사용한다.[18] 에이전트는 입력 이미지로부터 최적의 Action을 선택할 수 있도록 학습되며, 출력 Action은 보상과 함께 Replay buffer D에 Transition으로 저장된다. SAC의 매개 변수는 가치 함수의 Ψ, Q 함수의 Θ, 정책 함수의 ϕ로 구성된다. Critic 매개 변수는 Replay buffer D에서 샘플링된 Transition을 사용하여 벨만 오류를 최소화함으로써 학습된다.

$$J_{Q_\theta} = E_{(o_t,a_t)\sim D}\left[(Q_\theta(o_t,a_t) - (r_t + \gamma V_\psi(o_{t+1})))^2\right] \quad (1)$$

현재 정책에 따라 Action을 샘플링하여 소프트 상태가치함수를 다음과 같이 추정할 수 있으며 $\bar{Q}_\theta$는 Critic network의 Exponential moving average를 의미한다.

$$V_\psi(o_{t+1}) = E_{a'\sim\pi_\phi}\left[(\bar{Q}_\theta(o_{t+1},a') - \alpha\, log\pi_\phi(a'|o_{t+1})\right] \quad (2)$$

정책 매개 변수는 Q 함수의 Exponential과의 차이를 최소화하도록 학습된다.

$$J_{\pi_\phi} = -E_{a_t\sim\pi_\phi}\left[(Q_\theta(o_t,a_t) - \alpha\, log\pi_\phi(a_t|o_t)\right] \quad (3)$$

SAC는 고차원 이미지 데이터를 사용할 때 효과적인 강화학습 알고리즘이며 특히 연속적인 Action space에서 좋은 성능을 나타낸다.

2.2 Self-Supervised Learning

비지도학습(Unsupervised Learning) 방법의 일종인 자기지도학습(Self-supervised Learning)은 최종적으로 성능을 향상시키고자 하는 Downstream task를 위해 구실 작업(Pretext task)을 학습하는 것을 목표로 한다.[19] 분류, 객체 감지 또는 강화학습과 같은 Downstream task의 성능을 향상시킬 수 있는 적절한 구실 작업을 학습한 모델은 라벨이 지정되지 않은 고차원 이미지 데이터로부터 유효한 Feature를 추출할 수 있으며, 전이 학습을 통해 이를 활용할 수 있다. 최근 MoCo,[20] SimCLR,[21] BYOL,[22] BERT[23]와 같은 자기지도학습 알고리즘은 자연어 처리 및 컴퓨터 비전 분야에서 큰 발전을 이루었으며 이미지 기반 강화학습에도 적극적으로 적용되고 있다.

자기지도학습은 구실 작업에 따라 여러 유형으로 나눌 수 있다. 그 중에서도 대조학습(Contrastive learning)은 Positive 이미지 쌍 간의 유사성을 높이고 Negative 이미지 쌍 간의 유사성을 줄이는 것을 목표로 하는 자기지도학습 방법이다.[24] Fig. 3에 나타난 바와 같이, Positive 쌍과 Negative 쌍을 정의하기 위해, 입력 이미지를 두 번 Random crop 하여 쿼리와 키 이미지를 생성한다. 쿼리를 기준으로 동일한 이미지에서 증강된 키를 Positive 쌍으로 정의하고, 다른 이미지에서 생성된 키를 Negative 쌍으로 정의한다. 대조학습을 통해 쿼리 인코더는 라벨이 지정되지 않은 이미지에서 풍부한 표현 벡터를 추출하

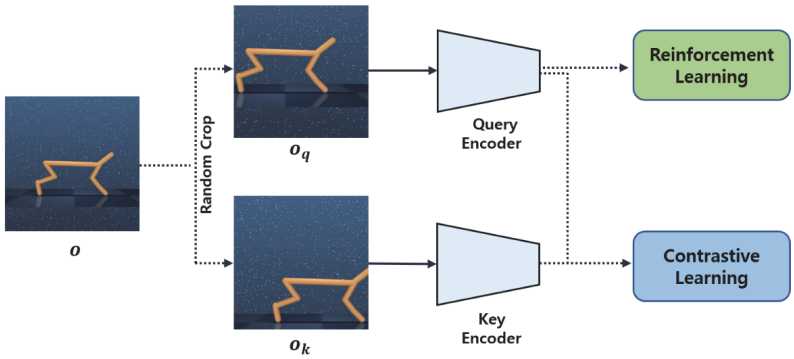

Fig. 3 Conventional reinforcement learning architecture with contrastive learning(CURL[11])

여 강화 학습의 이미지 효율성을 향상시킬 수 있다. 본 논문에서는 대조 학습의 손실 함수로 InfoNCE를 사용한다. 식 (4)에서, q는 이미지로부터 50차원으로 임베딩된 쿼리 벡터이고, k_+와 k_i는 각각 동일한 차원의 Positive 키와 Negative 키 벡터이며, W는 벡터 내적 연산을 위한 행렬이다.[25)]

$$L_{NCE} = log \frac{exp(q^T W k_+)}{exp(q^T W k_+) + \sum_{i=0}^{K-1} exp(q^T W k_i)} \quad (4)$$

2.3 Random Convolution

랜덤 네트워크는 이미지 기반 심층 강화학습의 다양한 관점에서 성능을 개선하기 위해 사용되었다. 예를 들어, 앙상블 기반 접근법에 초점을 맞춘 연구에서는 심층 강화학습의 불확실성 추정과 탐색을 개선하기 위해 무작위 네트워크를 사용했다.[26)] 또한, 미탐색 상태 인식 작업에서 무작위로 초기화된 신경망을 사용하여 미탐색 상태 방문에 대한 본질적 보상을 정의했다.[27)] 본 연구에서는 이미지 기반 강화학습의 일반화 성능을 개선하기 위해 랜덤 네트워크를 사용한다. 입력 이미지는 커널 크기가 3인 단일 레이어 CNN에 의해 무작위로 변화한다. 또한 출력 이미지의 차원은 입력 차원과 동일하도록 패딩 과정을 거친다. 학습 시 매 Iteration마다 랜덤 네트워크의 가중치는 Xavier 정규 분포[28)]를 따라 무작위로 초기화된다.

입력 이미지가 학습의 모든 Iteration마다 무작위로 초기화되는 컨볼루션 레이어를 통과할 때 강화학습 에이전트는 학습 환경과 같지만 사소한 배경이나 색상과 같은 것들이 변화한 테스트 환경에 더 강건하도록 학습될 수 있다. 즉, Fig. 4와 같이 랜덤 컨볼루션을 통과한 입력 이미지는 시각적 패턴이 다양하게 변화하고 색상, 모양 또는 질감과 같은 다양하고 교란된 Low level feature을 제공할 수 있도록 강력하게 증강되기 때문에 강화학습의 일반화를 크게 개선할 수 있다.[28)] 랜덤 컨볼루션과 같은 강력한 데이터 증강은 일반화에 대한 보조 효과를 향상시킬 수 있지만, 입력 이미지의 정보를 크게 변화시켜 강화학습의 불안정성과 성능 저하를 초래하기 때문에 독립적인 사용이 제한된다.[18)]

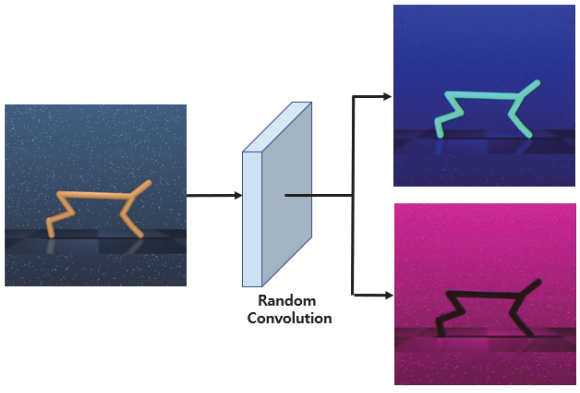

Fig. 4 Example of a random convolution process

3. Proposed Convolutional Contrastive Learning for Reinforcement Learning

이 섹션에서는 기존 강화학습 프레임워크에 추가할 수 있는 간단한 컨볼루션 대조학습 아키텍처를 제안한다. 먼저, 이미지 기반 강화학습의 일반화성능을 향상시키기 위한 컨볼루션 대조학습을 소개한다. 이후 랜덤 컨볼루션을 통한 강한 증강으로 인해 강화학습 성능이 저하되는 것을 방지함과 동시에 학습 환경과 배경이 다른 다양한 테스트 환경에서의 일반화 성능 향상을 극대화할 수 있도록 하는 학습 방법을 소개한다.

3.1 Randomized Input Observation

강화학습 에이전트는 랜덤하게 증강된 입력 이미지를 사용하여 학습된다. 입력 이미지를 랜덤화하기 위해 Feature extractor 전면에 단층의 컨볼루션 신경망을 랜덤 네트워크로 추가한다. 각 반복마다 랜덤 네트워크의 가중치는 Xavier 정규 분포를 따라 무작위로 초기화된다.[29)] 랜덤 네트워크를 통과한 출력 이미지는 입력과 동일한 차원을 가지며, 여러 패턴의 다양한 랜덤 이미지가 생성된다.

입력 이미지의 과도한 변화로 인한 시각적 정보의 손실을 방지하기 위해 Fig. 5와 같이 랜덤 컨볼루션 레이어를 통과한 이미지 o_{random}과 원본 이미지 o_{origin}을 일정 비율로 혼합하여 최종 입력 o_{blend}로 사용한다. 영상 혼합 비율은 식 (5)와 같이 파라미터 α를 통해 설정된다.

$$o_{blend} = \alpha * o_{origin} + (1-\alpha) * o_{random} \ldots (0 < \alpha < 1) \quad (5)$$

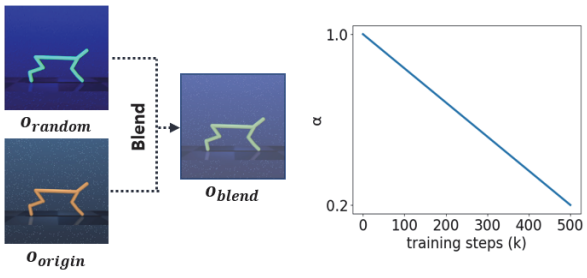

Fig. 5 Principle of blending original and randomized images

3.2 Strong Convolutional Contrastive Learning

α가 1에 가까울수록 원본 이미지의 혼합 비율이 크므로 컨볼루션 대조학습이 일반화 성능에 대한 충분한 보조 효과를 얻을 수 없다. 반대로 α가 0에 가까운 값일 때는 랜덤 이미지 혼합 비율이 크기 때문에 일반화 효과는 크지만 학습 자체를 혼란스럽게 할 수 있다. 본 절에서는 데이터 증강 강도와 관련된 Trade-off를 극복하고 강력한 데이터 증강을 효과적으로 활용하는 학습 방법을 소개한다.

학습 시작 단계에서는 랜덤 컨볼루션이 입력 이미지에 적용되지 않는다. 즉 CURL[11]와 유사한 방식으로, 인코더를 통해 생성된 쿼리 및 키 벡터는 강화학습 및 대조학습에 사용된다. Fig. 3에 나타난 바와 같이, 랜덤 컨볼루션 레이어는 추가되지 않으며, 인코더는 대조학습을 위한 약한 데이터 증강(Random crop)만을 사용하여 훈련된다.

이후, Fig. 6과 같이 랜덤 컨볼루션 레이어가 인코더 전면에 추가되어 강력한 데이터 증강을 유도한다. 생성된 랜덤 이미지는 원본 이미지와 혼합하여 최종 입력 데이터로 사용되며 학습이 진행될수록 랜덤 이미지의 혼합 비율은 커진다. 실험 결과, α가 0.2 아래의 값으로 내려갈 경우, 원본 이미지의 정보를 거의 사용하지 않게 되어 학습이 매우 불안정하게 진행되었다. 최종적으로, α를 1에서 0.2로 선형적으로 감소하도록 설정했을 때 전반적으로 가장 나은 일반화 성능을 나타냈다. 제안한 학습 방식과 같이 랜덤 이미지 혼합 비율을 동적으로 변화시킨 경우가 학습의 시작부터 끝까지 같은 강도의 강한 증강을 사용하는 학습 방법보다 더 효과적으로 강하게 증강된 이미지를 사용할 수 있다.

4. Experimental Results

본 연구의 목표는 강한 증강으로 인한 강화학습 성능 저하를 방지하여 일반화 성능에 대한 보조 효과를 극대화하는 것이다. 일반화 성능 평가를 위해 먼저 Google DeepMind Control(DMControl) generalization benchmark의 Train 환경에서 500 k개의 Frame을 통해 강화학습 에이전트를 학습시킨다.[14] 그 후 PAD[30]의 검증 환경을 따라 정적으로 변화하는 배경(Color-Hard 모드)과 동적으로 변화하는 배경(Video-Easy 모드)의 테스트 환경에서 일반화 성능을 측정한다.

실험은 CPU: AMD Ryzen 7 3700X / RTX 3090Ti / RAM: 32 GB 환경에서 진행되었다. 네트워크 학습을 위해 Optimizer = ADAM, Learning Rate = 0.0001을 사용하였으며, 입력 받은 이미지를 84 × 84로 Crop하여 사용하였다. 검증은 다음과 같이 두 가지 방식으로 이루어진다. 먼저, 제안한 컨볼루션 대조학습 방법의 증강 강도에 따른 일반화 성능을 비교하여 최적의 이미지 증강 방법에 대해 연구하고, 해당 방법을 이미지 기반 강화학습의 일반화 성능에 대한 여러 기존 알고리즘들과 비교한다.

4.1 증강 강도에 따른 일반화 성능 비교

본 절에서는 제안한 컨볼루션 대조학습 방법에서의 이미지 혼합 파라미터 α가 일반화 성능에 미치는 영향을 연구한다. Table 1은 DMControl generalization benchmark 환경의 Color-Hard 및 Video-easy 모드에 대한 테스트 점수를 나타낸다. 먼저 제안한 컨볼루션 대조학습 방법 자체의 일반화 효과를 검증하기 위해 기본 강화학습 알고리즘인 SAC의 테스트 점수를 측정한다. 그 후, 랜덤 이미지와 원본 이미지의 혼합 파라미터를 각각 1, 0.8, 0.2로 고정하고 학습을 진행한다. Fig. 7에서 볼 수 있듯이, 랜덤 네트워크를 통과하는 이미지의 혼합 비율이 클 수록

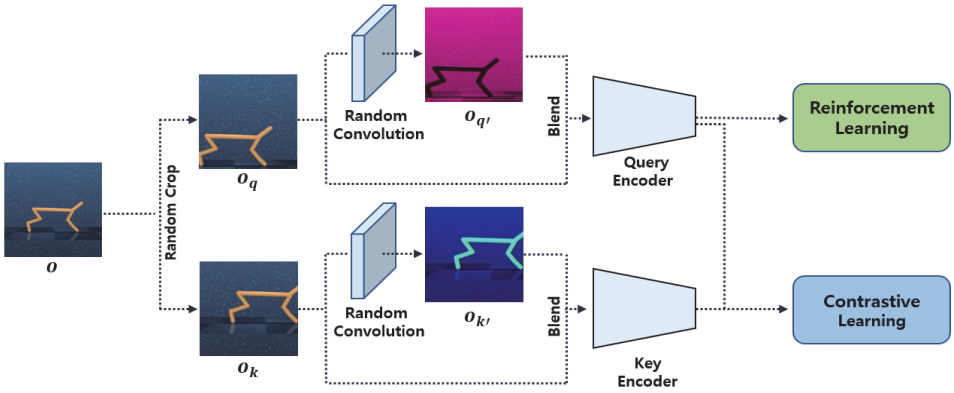

Fig. 6 Reinforcement learning and contrastive learning with the random convolution ($0 < \alpha < 1$)

Table 1 Test scores for different augmentation strength of proposed convolutional contrastive learning in the DMControl color-hard mode and video-easy mode

Proposed convolutional contrastive methods	CURL (SAC + Contrastive) α = 1	CURL with random convolution α = 0.8 (fixed)	CURL with random convolution α = 0.2 (fixed)	CURL with random convolution α: 1→0.2 (decreasing)
Color-hard mode				
Walker, walk	445 ± 99	707 ± 43	617 ± 46	**794 ± 18**
Walker, stand	662 ± 54	874 ± 46	912 ± 27	**939 ± 61**
Cartpole, swingup	454 ± 110	**790 ± 59**	375 ± 39	**766 ± 30**
Cartpole, balance	782 ± 13	921 ± 15	970 ± 22	**971 ± 27**
Ball in cup, catch	231 ± 92	713 ± 166	713 ± 93	**804 ± 62**
Finger, turn_easy	202 ± 32	438 ± 95	454 ± 133	**473 ± 104**
Cheetah, run	202 ± 22	251 ± 33	274 ± 13	**287 ± 8**
Reacher, easy	325 ± 32	317 ± 67	212 ± 91	**330 ± 46**
Video-easy mode				
Walker, walk	556 ± 133	784 ± 34	689 ± 46	**849 ± 55**
Walker, stand	852 ± 75	766 ± 47	891 ± 35	**921 ± 22**
Cartpole, swingup	404 ± 67	589 ± 44	415 ± 38	**636 ± 29**
Cartpole, balance	850 ± 91	926 ± 13	942 ± 18	**944 ± 25**
Ball in cup, catch	316 ± 119	692 ± 85	643 ± 93	**723 ± 99**
Finger, turn_easy	248 ± 56	**461 ± 188**	367 ± 154	384 ± 171
Cheetah, run	154 ± 50	**287 ± 21**	234 ± 32	256 ± 29

Fig. 7 Learning curves of proposed convolutional contrastive reinforcement learning according to blending parameter α in DMControl Walker-walk environment, color-hard mode. (a) uses only original image. (b) and (c) use blended image with blending parameter α is 0.8 and 0.2 respectively. (d) uses decreasing blending parameter α (1 → 0.2)

일반화 성능에 대한 대조학습의 보조 효과가 증가하기 때문에 학습 점수와 테스트 점수 사이의 차이는 작지만 학습 점수 자체가 낮아지기 때문에 비약적인 성능 향상을 기대할 수 없다.

이러한 증강 강도에 따른 강화학습의 안정성과 일반화 성능 사이의 Trade-off를 극복하기 위해 학습 동안 이미지 혼합 파라미터 α가 1에서 0.2로 일정하게 감소하도록 설정하고 학습한 뒤 테스트 결과를 비교하였다. 학습 초기에는 원본 이미지의 비율을 높게 설정하고 학습이 진행함에 따라 랜덤 이미지 혼합 비율을 점점 더 증가시키는 방법을 통해 강화학습 에이전트는 점점 더 강해지는 증강 이미지를 사용하게 된다. 해당 학습 방법을 통해 강한 증강으로 인한 학습 성능 저하를 방지함으로써 정적 및 동적으로 배경이 변화하는 테스트 환경에서 더 높은 점수를 얻을 수 있다. 제안된 접근 방식은 훈련 과정 전반에 걸쳐 동일한 증강을 사용하는 방법보다 이미지 기반 강화학습의 일반화 성능을 효과적으로 향상시킬 수 있다.

Table 2 Test scores for proposed method (convolutional contrastive learning with dynamic image blending ratio) and SOTA (state-of-the-art) studies in DMControl color-hard mode and video-easy mode

Comparison with SOTA studies	CURL	RAD	DrQ	PAD	Proposed (Convolutional contrastive)
Color-hard mode					
Walker, walk	445 ± 99	400 ± 61	520 ± 91	468 ± 47	**794 ± 18**
Walker, stand	662 ± 54	644 ± 88	770 ± 71	797 ± 46	**939 ± 61**
Cartpole, swingup	454 ± 110	590 ± 53	586 ± 52	630 ± 63	**766 ± 30**
Ball in cup, catch	231 ± 92	541 ± 29	365 ± 210	563 ± 50	**804 ± 62**
Video-easy mode					
Walker, walk	556 ± 133	606 ± 63	682 ± 89	717 ± 79	**849 ± 55**
Walker, stand	852 ± 75	745 ± 146	873 ± 83	**935 ± 20**	921 ± 22
Cartpole, swingup	404 ± 67	373 ± 72	485 ± 105	521 ± 76	**636 ± 29**
Ball in cup, catch	316 ± 119	481 ± 26	318 ± 157	436 ± 55	**723 ± 99**

4.2 기존 알고리즘과의 일반화 성능 비교

다음으로, 앞서 제안한 이미지 혼합 파라미터를 변화시키는 컨볼루션 대조학습 방법을 기존 SOTA 알고리즘들과 비교한다. CURL[11]: 제안한 방법에서 이미지 혼합 파라미터 α를 1로 설정하는 것과 동일한 방법으로, 대조학습을 위한 약한 증강(Random crop)만을 사용하는 방법; RAD[31]: 랜덤 Translation 및 랜덤 진폭 스케일이라는 두 가지 새로운 데이터 증강을 도입하여 일반화 성능을 향상시킨 방법; DrQ[32]: 데이터 증강을 통한 가치함수 정규화를 사용한 방법; PAD[30]: 테스트 시 정책 적응(Policy Adaptation)을 위한 자기지도학습(Self-supervised Learning)을 적용한 방법. Table 2에서 확인할 수 있듯이, DMControl의 정적 및 동적 배경 변화를 포함하는 모든 테스트 환경에서 제안한 동적 이미지 혼합 파라미터를 통한 컨볼루션 대조학습을 통한 강화학습은 기존 연구들의 일반화 성능을 크게 능가한다.

5. 결 론

본 논문에서는 이미지 기반 강화학습 에이전트가 강한 증강 이미지를 입력 데이터로 사용할 수 있도록 하는 새로운 자기지도학습 방법을 제안한다. 학습 초기에는 원본 이미지만 사용하여 강화학습과 대조학습을 진행하고, 학습이 진행됨에 따라 랜덤 이미지의 혼합 비율이 점차적으로 증가하는 방식을 통해 증강 강도를 높인다. 이를 통해 강화학습 모델은 점점 더 강력한 증강 이미지로부터 학습하게 되며 다양한 Test 환경의 정적 및 동적 배경 변화에 대해 강건해진다. 동시에, 학습의 처음부터 끝까지 같은 강도의 강한 이미지 증강을 사용하는 방법에 비해 학습 안정성이 크게 감소하지 않으므로 강화학습 성능 자체도 안정적으로 보장된다.

Google DMControl generalization benchmark에서의 다양한 환경에서의 실험 결과를 통해, 제안한 학습 방법을 사용하면 훈련 전체에 일관되게 강한 데이터 증강을 적용할 때보다 더 효율적으로 강하게 증강된 입력 이미지를 활용할 수 있다는 것을 확인할 수 있다. 또한, 제안한 방법은 고차원 이미지로부터 강건한 Feature을 추출하는 데 있어 기존 연구들의 성능을 능가한다. 이러한 연구 결과는 이미지 기반 강화학습에서 데이터 증강의 이점은 극대화하면서 증강으로 인해 발생하는 학습 안정성에 대한 문제는 상쇄하여 강화학습 모델의 성능과 일반화 능력을 동시에 향상시킬 수 있다는 점에서 의의가 있다.

후 기

이 논문은 2023년도 정부(국토교통부)의 재원으로 국토교통과학기술진흥원의 지원을 받아 수행된 연구임(과제번호: 23AMDP-C162182-03).

References

1) C. Song, G. B. Gu, W. Lim, S. C. Park and S. W. Cha, “A Energy Management Strategy for Hybrid Electric Vehicles Using Deep Q-Networks,” Transactions of KSAE, Vol.27, No.11, pp.903-909, 2019.
2) S. Kim, Y. Kim, H. Jeon, D. Kum and K. Lee, “Autonomous Driving Technology Trend and Future Outlook: Powered by Artificial Intelligence,” Transactions of KSAE, Vol.30, No.10, pp.819-830,

2022.
3) V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, “Playing Atari with Deep Reinforcement Learning,” arXiv preprint arXiv:1312.5602, 2013.
4) D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan and D. Hassabis, “A General Reinforcement Learning Algorithm that Masters Chess, Shogi and Go through Self-play,” Science, Vol.362, No.6419, pp.1140-1144, 2018.
5) V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, “Human-level Control through Deep Reinforcement Learning,” Nature, 518, pp.529-533, 2015.
6) O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, J. Quan, S. Gaffney, S. Petersen, K. Simonyan, T. Schaul, H. van Hasselt, D. Silver, T. Lillicrap, K. Calderone, P.l Keet, A. Brunasso, D. Lawrence, A. Ekermo, J. Repp and R. Tsing, “Starcraft ii: A New Challenge for Reinforcement Learning,” arXiv preprint arXiv: 1708.04782, 2017.
7) T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver and D. Wierstra, “Continuous Control with Deep Reinforcement Learning,” arXiv preprint arXiv:1509.02971, 2015.
8) D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke and S. Levine, “Scalable Deep Reinforcement Learning for Vision-based Robotic Manipulation,” Proceedings of the Conference on Robot Learning, PMLR, Zürich, Switzerland, pp.651-673, 2018.
9) B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman, “Building Machines that Learn and Think Like People,” Behavioral and Brain Sciences, Vol.40, e253, 2017.
10) L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, R. Sepassi, G. Tucker and H. Michalewski, “Model-based Reinforcement Learning for Atari,” arXiv preprint arXiv:1903.00374, 2019.
11) M. Laskin, A. Srinivas and P. Abbeel, “CURL: Contrastive Unsupervised Representations for Reinforcement Learning,” International Conference on Machine Learning, PMLR, Virtual, pp.5639-5650, 2020.
12) C. Zhang, O. Vinyals, R. Munos and S. Bengio, “A Study on Overfitting in Deep Reinforcement Learning,” arXiv preprint arXiv:1804.06893, 2018.
13) K. Cobbe, O. Klimov, C. Hesse, T. Kim and J. Schulman, “Quantifying Generalization in Reinforcement Learning,” International Conference on Machine Learning, PMLR, Long Beach, CA, USA, pp.1282-1289, 2019.
14) G. Ma, Z. Wang, Z. Yuan, X. Wang, B. Yuan and D. Tao, “A Comprehensive Survey of Data Augmentation in Visual Reinforcement Learning,” arXiv preprint arXiv:2210.04561, 2022.
15) N. Hansen and X. Wang, “Generalization in Reinforcement Learning by Soft Data Augmentation,” 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi’an, China, IEEE: Piscataway, NJ, USA, pp.13611-13617, 2021.
16) S. Park, J. Kim, H. Y. Jeong, T. K. Kim and J. Yoo, “C2RL: Convolutional-Contrastive Learning for Reinforcement Learning Based on Self-Pretraining for Strong Augmentation,” Sensors, Vol.23, No.10, Paper No.4946, 2023.
17) Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. D. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap and M. Riedmiller, “Deepmind Control Suite,” arXiv preprint arXiv:1801.00690, 2018.
18) T. Haarnoja, A. Zhou, P. Abbeel and S. Levine, “Soft Actor-critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor,” International Conference on Machine Learning, PMLR, Stockholm, Sweden, pp.1861-1870, 2018.
19) C. Doersch, A. Gupta and A. A. Efros, “Unsupervised Visual Representation Learning by Context Prediction,” Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, pp.1422-1430, 2015.
20) K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp.9729-9738, 2020.
21) T. Chen, S. Kornblith, M. Norouzi and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” International Conference

on Machine Learning, PMLR, Virtual, pp.1597-1607, 2020.
22) J. B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos and M.l Valko, "Bootstrap Your Own Latent-a New Approach to Self-supervised Learning," Advances in Neural Information Processing Systems, Vol.33, pp.21271-21284, 2020.
23) J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv: 1810.04805, 2018.
24) Z. Wu, Y. Xiong, S. X. Yu and D. Lin, "Unsupervised Feature Learning Via Non-parametric Instance Discrimination," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp.3733-3742, 2018.
25) A. V. Oord, Y. Li and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv preprint arXiv:1807.03748, 2018.
26) I. Osband, J. Aslanides and A. Cassirer, "Randomized Prior Functions for Deep Reinforcement Learning," Advances in Neural Information Processing Systems, Vol.31, pp.8626-8638, 2018.
27) Y. Burda, H. Edwards, A. Storkey and O. Klimov, "Exploration by Random Network Distillation," arXiv preprint arXiv:1810.12894, 2018.
28) K. Lee, K. Lee, J. Shin and H. Lee, "Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning," arXiv preprint arXiv:1910.05396, 2019.
29) X. Glorot and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Sardinia, Italy, pp.249-256, 2010.
30) N. Hansen, R. Jangir, Y. Sun, G. Alenyà, P. Abbeel, A. A. Efros, L. Pinto and X. Wang, "Self-supervised Policy Adaptation during Deployment," arXiv preprint arXiv:2007.04309, 2020.
31) M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel and A. Srinivas, "Reinforcement Learning with Augmented Data," Advances in Neural Information Processing Systems, Vol.33, pp.19884-19895, 2020.
32) I. Kostrikov, D. Yarats and R. Fergus, "Image Augmentation is All You Need: Regularizing Deep Reinforcement Learning from Pixels," arXiv preprint arXiv:2004.13649, 2020.