최신 제로-샷 강화학습 기술 고도화 탐구

박주영, 김태환, 박정호, 이주원
고려대학교 제어계측공학과
{parkj, kteaw0110, seanpark0107, saero94j}@korea.ac.kr

Investigations of Advancing Modern Zero-Shot Reinforcement Learning Methods

Jooyoung Park, Taehwan Kim, Jeongho Park, Juwon Lee,
Department of Control & Instrumentation Engineering, Korea University

요 약

본 논문에서는 미리 정의된 외부의 보상 함수를 중점적으로 다루는 고전적인 강화학습 패러다임으로부터, 초기 무보상 학습 단계 완료 후 환경이 처한 상황에 따라 보상이 다양하게 정의되는 새로운 강화학습 작업을 추가적인 플래닝 등의 계산 집약적인 과정을 거치지 않고 짧은 시간에 즉각적으로 풀 수 있는 "제어 가능한(controllable)" 에이전트를 구하는 형태의 강화학습 패러다임으로의 전환에 대한 고도화를 고려한다. 이같은 제어 가능 에이전트를 다루는 제로-샷 강화학습에 대한 탐구는 최근에 후속 특징(successor features) 등을 기반으로 한 다양한 접근 방식들이 시도되고 있다. 본 논문은 이러한 시도들 중 가장 대표적인 방안 중 하나인 딥마인드의 VISR(Variational Intrinsic Successor FeatuRes) 알고리즘의 개선을 다룬다. 보다 구체적으로, 기본 상태 특징(fundamental state features)과 후속 특징(successor features)의 학습을 위한 VISR의 방법론의 기본 구조와 잠재적인 문제점을 검토하고 이에 관한 특정 방향의 개선을 추구함으로써, 제로-샷 강화학습 관련 기술의 고도화를 추구한다. 아울러, 관련 예제를 통한 실험 결과를 소개함으로써 본 논문의 방법론이 갖는 효과를 간단히 살펴본다.

Ⅰ. 서 론

최근 들어, 심층 강화학습(deep reinforcement learning, deep RL)은 첨단 인공지능 기술의 혁신에 있어서 자연어 처리 및 비전 분야와 함께 매우 중요한 선도적인 역할을 수행하고 있다. 그러나, 타겟 레이블이 미리 지정되지 않은 대규모 데이터 세트에 대한 비지도 방식의 사전 학습(unsupervised pre-training)을 통해 광범위한 종류의 작업들을 제로-샷 방식[1]으로 처리할 수 있는 역량에 초점을 맞춘다면, 제로-샷 강화학습은 최근에 괄목할 만한 성공을 계속적으로 거두고 있는 제로-샷 자연어처리 및 제로-샷 비전 기술에 비해 상대적으로 더디게 발전하고 있는 것도 일정 부분 사실이다. 이러한 더딘 발전은 기본적으로 강화학습이 다루는 세팅이 상태 전이, 보상 신호의 변화 및 제어 입력을 산출하는 계산 절차 간의 결합이 복잡하게 얽혀있는 측면에 기인한다고 볼 수 있다. 그리고, 전형적인 강화학습의 기본적인 패러다임 자체가 정해진 하나의 보상 함수 또는 그에 가깝게 정의될 수 있는 소규모 관련 작업군들에 대한 보상 함수들의 집합에 대해서만 학습하는 형태에 의존하기 때문이기도 하다[2]. 최근에 비지도 방식의 사전 학습을 강화 학습에 적용하는 접근 방식들 중 가장 주목을 끌고 있는 방법들 중 하나로 소위 후속 특징(successor features)을 활용하는 딥마인드의 VISR(Variational Intrinsic Successor FeatuRes)[1]를 들 수 있다. 후속적인 표현을 활용하는 전략들은 중장기적인 상태 전이 과정을 다루기 쉽게 요약함으로써, 모델-기반 RL및 모델-프리 RL의 중간 정도의 위치에서 강화 학습 문제 풀이를 진행하는 전략이라고 할 수 있다. VISR는 상태-입력 가치함수인 Q-함수와 리워드 시그널(reward signal)의 표현을 위하여 작업 벡터로 불리는 task vector w를 내적의 공통 항으로 공유하는 형태를 취할 수 있는 매개변수화 전략에 의존한다. 본 논문에서는, VISR[1]의 알고리즘 절차와 실험 결과에 대한 관찰을 바탕으로, 비지도 사전 학습의 효과가 증대될 수 있는 표현 학습을 위한 방안을 탐구하여 보다 효율적이고 실용적인 제로-샷 심층 강화학습을 구현하는 방안에 대한 고찰을 수행한다.

본 논문은 참고문헌[3]-[5]와 같이 다양한 방법으로 제로-샷 강화학습을 고도화시키기 위한 최근의 연구 흐름의 일환으로 볼 수 있으며, 그 기여는 "VISR[1]의 연장선상에서 유니버설 후속 특징 근사기(universal successor features approximator) ψ와 기본 상태 특징(fundamental state features) 추출기 ϕ의 학습이 실용적인 의미를 가지기 위한 방안을 생각해보고 이를 구현하기 위하여 새로운 개선 방안을 탐구함"으로 요약될 수 있다.

본 논문의 순서는 다음과 같다. 1장에서는 본 논문이 다루는 문제에 대한 설명과 함께 주요 관련 연구를 인용한다. 2장에서는 본 논문의 방법론을 기술하기 위하여 주요 탐구 대상인 VISR 방법을 간단히 설명하고 추가적인 고도화를 위해 시도한 방법을 제시한다. 그리고, 간단한 실험을 통하여 본 논문이 제시한 방법이 가져다주는 효과를 관찰한다. 마지막으로, 3장에서는 본 논문의 결론 등을 언급한다.

Ⅱ. 본론

후속 표현을 이용하는 제로-샷 강화학습은 비지도 방식의 무보상(no reward) 학습 단계를 통하여 얻은 후속 특징(successor features)과 기본 특징(fundamental features)이 응용 단계에서 보상 함수로 정의되는 새로운 작업에 대해 별도의 강화 학습이나 플래닝(planning) 과정 없이 최적 제어 정책을 즉각적으로 제공할 수 있는 편리하고 강력한 일반화 능력을 가지므로 근래에 많은 관심을 집중적으로 받고 있다[1-5]. Finite MDP의 경우, 상태-액션 쌍 (s_0, a_0)의 후속 표현(successor representation)[6] $M^\pi(s_0, a_0 s)$은 정책 π 하에서 각 상태의 미래 발생 횟수의 할인된 합계로 정의된다. 즉,

$$M^\pi(s_0, a_0, s) = E[\sum_{t \geq 0} \gamma^t \, 1_{\{s_{t+1} = s\}} | (s_0, a_0), \pi], \ \forall s \in S. \quad (1)$$

후속 특징 $\psi^\pi(s,a)$는 (1) 식의 후속 표현의 정의에 사용된 indicator function $1_{\{s_{t+1}=s\}}$을 기본 상태 특징 맵(fundamental state feature map) $\phi(s_{t+1})$로 대체하는 형식으로 구한 표현 양식이다. 그리고, VISR[1]에서 주로 사용한 후속 특징 ψ은 소위 "유니버설 후속 특징 근사(universal successor feature approximation, USFA)로 부르는 특수한 경우로써, 보상 함수가 기본 후속 특징과 작업 벡터(task vector) w의 내적으로 표현(혹은 근사)될 때 그 핵심적인 사용 방식이 다음과 같이 요약될 수 있다.

$$Q^\pi(s,a) = <\psi^\pi(s,a), w> = <\psi(s,a,\tilde{w}), w> \quad (2)$$

VISR의 전개과정에서, 전형적인 Q-함수를 구할 때에는 (2) 식의 $\tilde{w}$는 w와

같게 사용되고, GPI(generalized policy improvement) 적용 과정에서는 w를 중심(location)으로 하고 적절한 집중도(concentration)를 갖는 Von Mises 분포로부터 취한 샘플들이 $\tilde{w}$를 위하여 사용된다. 그리고, VISR [1]의 기본 특징 ϕ을 구하기 위한 학습 부분에서는 작업 변수($p(w)$)와 상태 방문($p(s)$) 간의 행동 상호 정보(behavioral mutual information, BMI)를 최대화하여 구별 가능한 작업 조건 행동을 학습하도록 장려하는 손실함수를 변분 추론(variational inference)과 함께 사용하는 전략을 사용한다. VISR 알고리즘을 해당 논문 [1]의 8장 Appendix에서 고려한 예제에 적용하면, 다음과 같은 5차원의 기본 특징(fundamental features) $\phi(s)$ 얻게 된다(VISR 알고리즘 및 해당 예제에 관련된 상세한 설명은 각각 [1]의 Algorithm 1과 8장을 참고하면 됨).

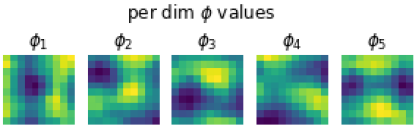

그림 1. VISR 방법론[1]에 따른 기본 특징 예시.

그림 2. VISR[1] 적용 결과.

이러한 시뮬레이션 결과는 참고문헌 [1]의 그림 3에 보고된 해당 결과와 대동소이함을 관찰할 수 있다. 본 논문에서 주목하는 측면은 그림 1이 보여주듯이, VISR 알고리즘[1]을 통해 학습된 기본 특징 벡터의 각 성분 $\phi_i(s)$는 비교적 전역적으로 분산된 영역에서 큰 값을 갖는 비국소적 성질을 갖는다는 점이다. 이러한 측면은 결과적으로 적용 단계에서 보상 함수가 특정 목표 지점을 추구하는 경우에 설명가능성이 약해지는 현상을 낳게 된다. 해당 예제에 대한 VISR 알고리즘 응용 결과를 전체적으로 보여주는 다음의 그림 2에서 이러한 현상을 뚜렷하게 관찰할 수 있다. 즉, 참 보상함수를 기본 특징의 일차결합으로 fitting한 predicted reward가 비국소적 성질을 갖는 관계로, 알고리즘 적용 결과로 얻어지는 visitation과 policy등이 특정 목표 지점을 추구하는 성질을 확실히 갖는가에 대한 설명 가능성이 약해지는 지점을 낳게 된다.

본 논문에서는, 이와 같이 기본 특징 벡터의 각 성분 $\phi_i(s)$이 전역적으로 분산된 영역 대신 비교적 국소적 특징을 갖도록 하는 방안을 기본 특징 ϕ를 위한 손실함수에 직접 강제하지 않고 전체적인 VISR 알고리즘 틀 안에서의 변형을 통해 추구해보기로 한다. 만일 실제로 다루어야 하는 응용 분야가 국소적 기본 특징에 대한 필요성을 절대적으로 갖는 경우라면, 본 논문에서 고려한 VISR의 변형 방안에 손실함수를 통해 이러한 필요성을 강제하는 측면을 추가적으로 반영할 수 있을 것이므로 손실함수에 국소적 특징을 직접 강제하는 방안을 추가적으로 융합하는 부분은 추후 연구 과제를 위한 숙제로 미루어 두기로 한다. 본 논문에서 탐구하는 VISR 알고리즘 틀 안에서의 변형은, [1]에 제시된 VISR 알고리즘 ([1]의 Algorithm 1)에서 유니버설 후속 특징 ψ의 학습을 위한 손실함수가 ϕ_i와 ψ_i에 관한 항들을 반영함에 착안하여 전체적인 VISR 알고리즘 틀 안에서 이들이 명시적으로 반영될 수 있도록 이들에 대응하는 작업 벡터들을 학습과정에 직접 포함하는 전략과 [2]에서 언급된 바 있는 기본 특징에 관한 orthonormality $E[\phi(s)\phi^T(s)] \approx Id$에 관한 regularization 항을 ϕ를 위한 손실함수에 포함하는 전략 등을 사용한다. 이러한 전략들을 채용한 결과들은 그림 3과 4에 예시되었으며 이들은 그림 1과 2에 비교하여 설명 가능성의 강화 등의 바람직한 측면을 보임을 관찰할 수 있다. 이와 관련된 향후 연구로, 추가적인 전략들에 대한 전반적인 고려를 통하여 보다 우수한 버전의 관련 고도화를 달성할 수 있도록 종합적인 탐구가 진행될 예정이다

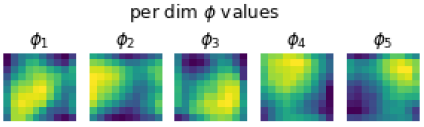

그림 3. 본 논문 방법론에 따른 기본 특징 예시.

그림 4. 본 논문 방법론 적용 결과.

Ⅲ. 결론

본 논문에서는 초기 무보상 학습 단계 완료 후 환경에서 상황에 따라 다양하게 보상이 정의되는 새로운 강화학습 작업을 추가적인 플래닝 등의 고비용 계산을 필요로 하지 않고 단시간에 즉각적으로 풀 수 있는 "제어 가능한(controllable)" 에이전트를 학습하는 형태의 강화학습 패러다임으로의 전환을 위한 고도화를 위하여, 최근에 발표된 딥마인드의 VISR 알고리즘 [1]의 개선을 다루었다. 보다 구체적으로, 기본 특징과 후속 특징(successor features)의 학습을 위한 VISR의 방법론의 학습 절차와 손실함수를 검토하고 이에 관한 특정 방향의 개선을 추구함으로써 관련 기술의 고도화를 추구하였다. 본 논문에서 주목하는 측면은 VISR 알고리즘[1]을 통해 학습된 기본 특징 벡터의 각 성분이 비교적 전역적으로 분산된 영역에서 큰 값을 갖는 비국소적 성질을 갖는다는 점이다. 이러한 측면은 결과적으로 적용 단계에서 보상 함수가 특정 목표 지점을 추구하는 경우에 설명가능성이 약해지는 결과를 낳을 수 있으므로, 본 논문에서는 VISR 알고리즘 틀 안에서의 변형을 통한 개선을 위하여 개선 방안을 고려하고, 관련 예제를 통한 실험 결과를 분석함으로써 본 논문이 제시하는 방안들이 갖는 효과를 간단히 검증하고 향후 연구 방향을 제시하였다. 본 논문의 고찰이 장차 제로-샷 강화 학습에 관한 추가적인 이론 개발 및 구현을 위한 의미 있는 방향이 될 수 있기를 기대한다.

ACKNOWLEDGMENT

감사의 글: 본 연구는 과학기술정보통신부의 재원으로 한국연구재단(NRF-2020R1F1A1072772)의 지원을 받아 수행되었음.

참 고 문 헌

[1] Hansen, S., Dabney, W., Barreto, A., Van de Wiele, T., Warde-Farley, D., & Mnih, V. (ICLR2020). Fast Task Inference with Variational Intrinsic Successor Features.

[2] Touati, A., Rapin, J., & Ollivier, Y. (2022). Does Zero-Shot Reinforcement Learning Exist? *arXiv preprint arXiv:2209.14935*.

[3] Touati, A., & Ollivier, Y. (2021). Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, *34*, 13-23.

[4] Liu, H., & Abbeel, P. (2021). APS: Active pretraining with successor features. In *International Conference on Machine Learning* (pp. 6736-6747). PMLR.

[5] Ren, T., Zhang, T., Lee, L., Gonzalez, J. E., Schuurmans, D., & Dai, B. (2022). Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*.

[6] Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural computation*, *5* (4), 613-624.