글로벌 최적 솔루션을 위한 설명 가능한 심층 강화 학습 지식 증류

이봉준 [1], 조인휘 [2]
[1] 한향대학교 컴퓨터소프트웨어학과 대학원생
[2] 한양대학교 컴퓨터소프트웨어학과 교수
lfj.6295@hanyang.ac.kr, iwjoe@hanyang.ac.kr

Explainable Deep Reinforcement Learning Knowledge Distillation for Global Optimal Solutions

Fengjun Li[1], Inwhee Joe[2]
[1]Dept. of Computer Science, Hanyang University

요 약

설명 가능한 심층 강화 학습 지식 증류 방법(ERL-KD)이 제안하였다. 이 방법은 모든 하위 에이전트로부터 점수를 수집하며, 메인 에이전트는 주 교사 네트워크 역할을 하고 하위 에이전트는 보조 교사 네트워크 역할을 한다. 글로벌 최적 솔루션은 샤플리 값과 같은 해석 가능한 방법을 통해 얻어진다. 또한 유사도 제약이라는 개념을 도입하여 교사 네트워크와 학생 네트워크 간의 유사도를 조정함으로써 학생 네트워크가 자유롭게 탐색할 수 있도록 유도한다. 실험 결과, 학생 네트워크는 아타리 2600 환경에서 대규모 교사 네트워크와 비슷한 성능을 달성하는 것으로 나타났다.

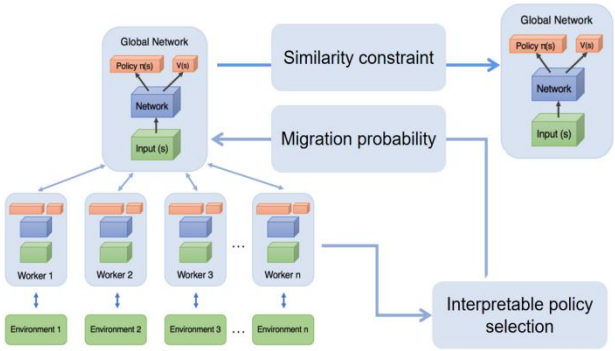
1. 서론

강화학습은 순차적 의사결정 문제를 해결하기 위한 유력한 접근법으로 부상했으며 다양한 영역에서 괄목할 만한 성과를 거두었다. 예를 들어, AlphaGo 프로그램은 바둑 게임에서 최고의 인간 플레이어를 이겼고[1], 생산 자동화[2]에 적용되었으며, 자율 주행 분야[3]에서도 광범위하게 활용되고 있다. Mnih 등[4]이 소개한 A3C 알고리즘은 액터-크리틱 방식과 비동기 훈련을결합한 주목할 만한 접근법이다. 하지만 강화 학습은 샘플링 과정에서 몇 가지 한계가 있다.

(1)로컬 최적 솔루션으로 글로벌 모델이 최적의 상태에 도달하지 못할 가능성이 있다[5]. (2)심층 강화 학습 알고리즘은 일반적으로 학습 과정에 상당한 컴퓨팅 리소스와 시간이 필요한다[6]. (3)복잡한 내부 구조와 매개 변수 간의 상호 작용으로 인해 알고리즘의 의사 결정 과정을 상대적으로 이해하기 어렵다.

2. ERL-KD

해석 가능한 심층 강화학습 지식 증류 방법을 제안한였다. 흐름도는 그림 1 에 나와 있다. 먼저 사전 학습된 병렬 네트워크의 하위 에이전트의 성능을 평가한다. 그런 다음 Shapely 값을 사용하여 환경 내 모든 하위 에이전트의 성능을 메인 에이전트의 파라미터 전송 확률로 변환한다. 마지막으로, 유사성 제약을 활용하여 학생 네트워크는 지식 수용과 환경 탐색 간의 최적의 균형을 달성함으로써 학습 성과를 개선한다.



(그림 1) ERL-KD 프레임워크 다이어그램.

A. 해석 가능한 정책 선택

병렬 네트워크에 있는 모든 K 개의 하위 에이전트 네트워크의 파라미터를 고정한다. 전이 확률은 shapely 값의 공식 (1)를 통해 구할 수 있다:

$$P_k = \frac{e^{S_k}}{\sum_{k=1}^{K} e^{S_k}} \quad (1)$$

여기서 P_k는 각 라운드에서 에이전트가 선택될 확률이다. 공식 (1)는 특정 시나리오에서 각 하위 에이

전트의 성능을 메인 에이전트를 업데이트하기 위한 확률 표현식으로 변환하는 데 사용된다.

B. 네트워크 간 유사도

ERL-KD 는 유사성 개념을 도입하여 학생 네트워크가 교사 네트워크로부터 지식을 받을 수 있을 뿐만 아니라 환경에 대한 효과적인 탐색을 장려할 수 있도록 하며, 정책 업데이트 과정에서 다음과 같은 제약 조건이 도입된다:

$$L_{\pi}^{\eta} = E[\, \| \eta(\pi_{tea}(\cdot|s), \pi_{stu}(\cdot|s)) - \varphi_1 \| + \| \eta(\pi_{tea}(\cdot|s), \pi_{stu}(\cdot|s)) - \varphi_2 \| \,] \quad (2)$$

여기서 φ_1과 φ_2는 경사 최적화를 통해 정책 간의 유사성을 $[\varphi_1, \varphi_2]$ 구간 내로 제한하기 위한 유사성 상수이다. 알고리즘의 안정성을 높이기 위해 φ_1 및 φ_2의 값은 학습 프로세스 중에 0 을 향해 동적으로 조정된다. 다음은 φ_1 및 φ_2에 대한 업데이트 공식:

$$\varphi_1 = \varphi_1^{'}(1 - \frac{n}{M}), \varphi_2 = \varphi_2^{'}(1 - \frac{n}{M}) \quad (3)$$

여기서 n은 현재 반복 횟수, M은 최대 반복 횟수, φ_1 및 φ_2는 훈련 시작 전에 초기화된 고정 값이다.ERL-KD 목표 달성을 위한 전체 손실 함수는 다음과 같다:

$$J(\theta_{stu}, s, t) = \lambda_k J_{RL}(\theta_{stu}, s, t) + (1 - \lambda_k) L_{\pi}^{\eta} \quad (4)$$

이 맥락에서 k는 모델의 학습 반복 횟수를 나타낸다. 교사 네트워크 에이전트의 제어 정책은 학생 네트워크 에이전트에 영향을 미친다.

3. 실험결과

우리는 네트워크 파라미터의 압축률을 달리하여 소규모 학생 네트워크를 구성하고 Atari 2600 환경에서 실험을 진행했다. 적절한 하이퍼파라미터를 선택하기 위해 다음과 같은 실험적 검증 방법을 사용했다.

Atari 2600 에서 세 가지 고전적인 이미지 기반 환경을 선택했다: Pong, Breakout, SpaceInvaders. 이 세 가지환경의 '교사 네트워크'와 '학생 네트워크'의 구조는 표 1 와 표 2 에 나와 있다.

<표 1> 학생 네트워크 1 구조

Layer	Input	Convolutional kernel	Neuron	Activation formula	Ooutput
Conv1	84×84×4	3×3	32	ReLU	42×42×32
Conv2	42×42×32	3×3	32	ReLU	21×21×32
Conv3	21×21×32	3×3	32	ReLU	11×11×32
FC1	3872		465	ReLU	465
FC_P	465		4	Softmax	4

<표 2> 학생 네트워크 2 구조

Layer	Input	Convolutional kernel	Neuron	Activation formula	Ooutput
Conv1	84×84×4	3×3	32	ReLU	42×42×32
Conv2	42×42×32	3×3	32	ReLU	21×21×32
Conv3	21×21×32	3×3	32	ReLU	11×11×32
Conv4	11×11×32	3×3	64	ReLU	6×6×64
FC1	2304		340	ReLU	340
FC_P	465		4	Softmax	4

Atari 2600 환경의 피드백 곡선은 그림 2~4 에 나와 있으며, Pong 환경에서는 '학생 네트워크'의 파라미터를 20%만 압축했을 때 '교사 네트워크' 에이전트와 비슷한 성능을 달성했다. Breakout 환경에서는 두가지 압축률로 구성된 '학생 네트워크'의 성능이 '교사네트워크' 에이전트의 성능을능가하여 제안한 방법의 효과를 검증했다. SpaceInvaders 환경에서는 압축률이 45%인 학생 네트워크가 교사 네트워크보다 우수한 성능을 보였지만, 압축률이 20%인 경우 학생 네트워크의 성능이 교사 네트워크보다 떨어진다.

표 3 과 같이 학생 네트워크가 45%와 20%의 압축률에서 교사 네트워크와 일관된 학습 성능을 달성하는 아타리 2600 환경에서 제안된 접근법을 검증한다. 또한 Breakout 및 SpaceInvaders 이미지 환경에서 '교사 네트워크'를 능가하는 성능을 보여 A3C 모델의 압축을 위해 제안한 방법의 효과를 확인한다.

원본 A3C 모델에 비해 45% 와 20% 로 압축되고 Breakout 환경과 Pong 환경에서 A3C 모델과 비슷한 성능을 얻을수 있다.

<표 3> 네트워크 성능 비교

Environment	Teacher Network	Student Network 1	Student Network 2
Breakout	21.0	21.0	21.0
PONG	314.0	384.0	329.0
SpaceInvaders	705.0	800.0	450.0

참고문헌

[1] SILVERD, SCHRITTWIESERJ, SIMONYANK, et al. Master ringthe game of go without human knowled ge [J]. Feature, 2017550 (7676): 354-359

[2] Shi Daming, Fan Wenhui, Xiao Yingying, et al. Intel ligent scheduling of discrete automated production lin e via deep reinforcement leading [J] International Jou rnal of Production Research, 2020,58 (11):3362-3380

[3] Duan Jingliang, Li Shengbo, Guan Yang, et al. Hiera rchical information learning for self driving decision making without relationship on labeled driving data [J] IET Intelligent Transport Systems, 2020,14 (5): 2 97-305

[4] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deferred force learning [C]//Proc of the 33rd Int Conf on MachineLearning New York, NY: ACM, 2016:1928-1937

[5] Mao Hongzi, Venkatakrishnan S B, Schwarzkopf M, et al. Variancereduction for reinforcement learning in input-driven environments [J]. arXiv print, arXiv: 18 07.02264.2018

[6] Gamrian S, Goldberg Y. Transfer learning for related reinforcement learning tasks via image to image tra nslation [C]//Proc of the 36th IntConf on Machine L earning New York: ACM, 2019: 2063-2072