다중목표 강화학습에서의 재라벨링 방법에 관한 연구

송원근, 이정우
서울대학교, 서울대학교

swg0110@snu.ac.kr, junglee@snu.ac.kr

A Study on a relabeling method of multi-goal reinforcement learning systems

Song Won Geun, Lee Jung Woo
Seoul Univ., Seoul Univ.

요 약

본 논문은 다중목표 강화학습에서의 재라벨링 방법에 대한 연구를 진행하였다. 우리가 제안하는 방법에서는 우선 전이정보와 목표와의 매칭의 적합도를 정의하고 정의된 측도에 대한 기울기를 구하였다. 그리고 구해진 기울기를 이용하여 전이정보에 매칭된 목표를 변환하는 방식을 사용하였다. 그리고 제안하는 방법의 효과를 로보틱스 환경을 통해서 검증하였다.

Ⅰ. 서 론

다중목표강화학습 문제에서는 하나의 역학모델과 희박한 밀도의 이진의 보상함수를 가지는 과업들의 집합이 주어진다. 이 때 각각의 과업에 맞는 정책의 집합을 학습하는 것이 문제의 목적으로 여러개의 강화학습 문제를 동시에 해결해야 한다. 이런 상황의 특수성 때문에 일반적인 강화학습 환경에서 좋은 성능을 보였던 알고리즘들도 이 환경에 바로 적용하였을 때는 학습에 환경과의 훨씬 많은 상호작용을 필요로 하거나 학습이 되지 않는 등 좋은 결과를 얻지 못해왔다.

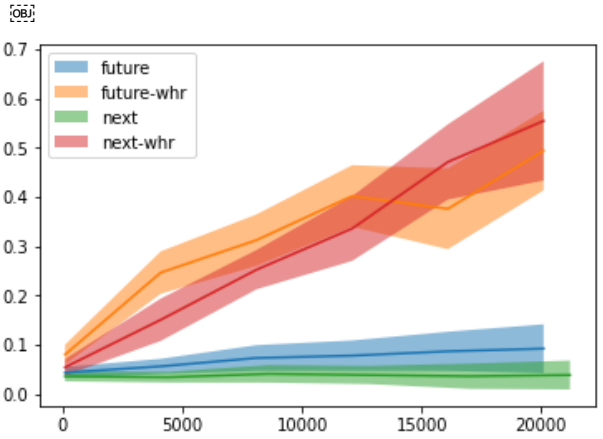
이런 문제를 해결하기 위해서 다양한 시도들이 있어왔다. 이런 시도중 재라벨링은 전이 정보를 학습대상이 되는 정책 집합에 있는 특정 요소와 짝지어주는 문제이다.이 과정을 통해서 전체 학습과정을 가속시키는 것이 재라벨링의 목적이다. 그러므로 합리적인 재라벨링을 위해서는 전이정보와 재라벨링되는 정책사이의 짝지음이 전체 학습과정이 미치는 영향에 대한 평가를 통해 이루어져야 한다. 하지만 이전 연구들은 주어진 전이정보가 어떤 과업에 속하는 전이일 경우 성공적이라 할 수 있을지를 추정할 뿐 짝지음에 대한 평가를 기반으로 재라벨링이 이루어 지지 않았다.

Ⅱ. 본론

우리는 우선 전이 정보와 짝지어진 목표가 얼마나 적합한지에 대한 측도를 정의하였습니다. 우리의 정의는 우선 학습과정에 대한 직관에서 시작한다. 우선 다중목표 강화학습은 여러개의 과업이 주어졌을 때 그 각각의 과업에 맞는 정책의 집합을 학습하는 것을 목표로 한다. 정책의 집합을 학습하는 과정에서 학습을 위한 전이 정보가 주어졌을 때 학습을 가장 효율적으로 시킬 수 있는 방법은 그 것으로부터 얻을 수 있는 정보가 가장 큰 것을 선택하는 것이다. 그런데 우리는 정책의 집합을 Q 학습 알고리즘을 이용하여 학습을 진행한다. Q 학습 알고리즘에서는 TD 오류를 줄이는 방향으로 학습이 일어나고 따라서 우리는 TD 오류를 정책과 전이 정보의 매칭의 적합도로 정의 하였다. 이것은 다르게 해석하면 전이 정보가 주어졌을 때 여기에 전이 정보를 포함하여 계산된 Q 값이 전이 정보를 포함하지 않고 계산된 값에 비해서 얼마나 높은지를 의미하므로 TD 오류가 가장 높은 정책을 선택하는 것은 해당 전이 정보가 가장 필요한 정책을 선택하는 것으로도 해석 할 수 있다.

그리고 우리는 TD 오류 값이 가장 큰 정책을 선택하는 것을 방법을 채택하고자 하였다. 그러나 이 것을 찾는 것은 전체 목표 공간에서의 최적화를 진행해야 한다는 것을 의미하고 이는 실용적이지 못하다. 또한 전체 목표 공간에서의 최적화를 진행하여 최적값을 찾았다고 한다고 해도 해당 정책은 우리가 필요하는 정책의 집합에 포함되어 있지 않을 가능성이 높다. 따라서 원래 정의되었던 정책과의 거리가 그렇게 멀지 않고, 앞서 정의된 값을 감소시키는 정책을 찾는 것을 목적으로 하였다. 따라서 우리는 TD 오류값에 대한 목표의 구배값을 구하고 해당 값을 이용하여 목표를 수정하는 방법을 사용하였다.

또한 우리는 제안하는 방법의 효과성에 대해서 실험적으로 입증하였다. 실험 환경으로 우리는 Fetch Pick and Place 환경을 사용하였다.



여기에서 future와 next는 주어진 전이정보의 미래의 어느 시점의 상태와 다음 타임스텝의 상태를 사용한 방법이고 future-whr와 next-whr는 이전 방법과 우리 방법의 결합된 형태로, 두 가지 방법으로 결정된 목표를 우리가 제안하는

방법으로 수정한 값을 목표로 사용하는 방법이다. 위의 도표에서 알 수 있듯이 우리의 방법을 사용한 것이 더 높은 성공율과 빠른 학습을 보인다라는 것을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 다중목표강화학습에서의 재라벨링 방법에 대해서 연구하였다. 우리는 TD 오류값을 증가시키는 방향으로 이에 대한 구배 정보를 사용하여 목표를 수정하는 방법을 제안하였고 해당 방법의 효과성을 로보틱스 환경인 Fetch Pick and Place 환경에 대해서 검증하였다.

.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (2021R1A4A1030898), Bio-Mimetic Robot Research Center Funded by Defense Acquisition Program Administration, Agency for Defense Development (UD190018ID), INMAC, and BK21-plus.

참 고 문 헌

[1] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. arXiv:1509.02971 [cs, stat], September 2015.

[2] Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine (2018). “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”. In: International Conference on Learning Representations.

[3] Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated super- vised learning. ICLR, 2021.

[4] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. ICRA, 2018. [5] Suraj Nair, Silvio Savarese, and Chelsea Finn. Goal-aware prediction: Learning to model what matters. ICML, 2020. [6] Henry Charlesworth and Giovanni Montana. Plangan: Model-based planning with sparse rewards and multiple goals. NeurIPS, 2020. [7] Menghui Zhu, Minghuan Liu, Jian Shen, Zhicheng Zhang, Sheng Chen, Weinan Zhang, Deheng Ye, Yong Yu, Qiang Fu, and Wei Yang. Mapgo: Model- assisted policy optimization for goal-oriented tasks. IJCAI, 2021.