사족보행 로봇 보행 제어를 위한 특권 정보 기반 교사 학생 학습 기법 비교

최요한[1], 김민준[2], 눌란벡 키즈 아셀[1], 김진성[2], 한연희[1†]

[1]한국기술교육대학교 컴퓨터공학과 미래융합공학전공

[2]한국기술교육대학교 컴퓨터공학과

{yoweif, june573166, aselbaekki, kjs0820k, yhhan}@koreatech.ac.kr

Comparison of Privileged Information-based Teacher Student Learning Methods for Quadrupedal Robot Locomotion Control

Yohan Choi[1], Min Jun Kim[2], Asel Nurlanbek kyzy[1], Jin Sung Kim[2], Youn-Hee Han[1]

[1]Future Convergence Engineering, Dept. of Computer Science and Engineering, KOREATECH

[2]Dept. of Computer Science and Engineering, KOREATECH

요 약

강화학습 기반의 로봇 보행 제어에서 특권 정보의 활용은 정책의 학습 효율성과 일반화 성능을 향상시키는 데 기여한다. 그러나 실제 로봇 배포 환경에서는 이러한 정보에 접근이 어려우므로, 교사-학생 학습 기법을 활용한 지식 증류 기법이 널리 연구되고 있다. 본 논문에서는 사족보행 로봇의 보행 제어를 위한 두 가지 주요 증류 방식인 잠재 표현 증류와 행동 복제를 비교 분석한다. 실험은 NVIDIA IsaacLab 시뮬레이션 환경과 Unitree Go1 로봇을 사용하였으며, 특권 정보의 재구성 정확도와 보행 제어 성능(에피소드 보상)을 기준으로 각 모델을 평가하였다. 그 결과, 잠재 표현 증류 방식은 특권 정보의 암묵적 재구성에서 더 높은 정확도를 보였으나, 행동 복제 방식이 보행 제어 성능에서는 더 우수한 결과를 나타냈다. 이를 통해 로봇이 특권 정보를 얼마나 정확하게 복원하는지와 실제 제어 성능 사이의 관계가 단순히 정비례하지 않음을 명확히 보여준다.

Ⅰ. 서론

최근 강화학습 기반의 로봇 보행 제어 분야에서 상당한 발전이 이루어지고 있으나 [1, 2], 로봇이 실제 환경에 효과적으로 배치되기 위해서는 환경 변화에 대한 높은 적응성과 강건성이 요구된다. 이를 달성하기 위해, 로봇이 실시간으로 인지하거나 센서로 직접 측정하기 어려운 물리적 파라미터(로봇 질량, 무게중심 위치, 지면 마찰 계수 등)와 같은 특권 정보(Privileged Information)를 시뮬레이터 내에서 활용하여 학습 효율성과 정책의 일반화 능력을 향상시키는 접근법이 주로 사용된다.

그러나 실제로 로봇이 배치될 때는 특권 정보가 명시적으로 주어지지 않기 때문에, 이를 명확하게 내재화하고 간접적으로 활용할 수 있는 방법이 필수적이다. 이를 위해 최근 교사-학생 학습 기법 [3]이 많이 사용되고 있다. 지식 증류(Knowledge Distillation) 기법 중 하나인 이 방법은 특권 정보에 접근할 수 있는 교사 모델에서 이에 접근할 수 없는 학생 모델로 지식을 전달하는 것을 목표로 한다. 사족보행 로봇의 보행 제어에서 두 가지 주요 접근법으로써 교사 모델의 특권 정보가 인코딩된 잠재 표현(Latent Representation)을 학습하는 잠재 표현 증류(Latent Distillation) 방식, 교사 모델의 행동을 직접적으로 모방하는 행동 복제(Behavior Cloning) 방식이 널리 사용되고 있다. 기존 연구들은 각 방식의 제어 성능이나 재구성 능력을 개별적으로 분석한 바 있으나, 두 방식 간의 상관관계를 정량적으로 비교한 연구는 드물다. 본 논문에서는 이 두 접근 방식 간의 특권 정보 재구성 정확도와 로봇 보행 제어 성능 간의 상관관계를 분석함으로써, 실제 환경에서의 로봇 배치를 고려할 때보다 효과적인 학습 전략이 무엇인지 파악하고자 한다.

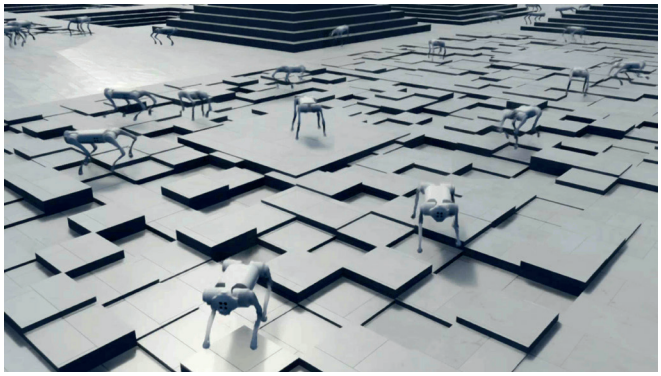
Ⅱ. 교사-학생 학습 기법 및 평가 방법

본 논문에서는 사족보행 로봇의 보행 제어를 위한 교사-학생 학습 기법의 학습 방법을 비교한다. 교사 모델과 학생 모델은 유사한 모델 구조를 가지지만, 입력 정보 처리 메커니즘에서 차이점이 존재한다. 교사 모델은 특권 정보에 직접 접근 가능한 모델로, 모델의 입력 정보로 고유수용성 정보(Proprioception)와 특권 정보를 모두 받아 사용한다. 반면, 학생 모델은 특권 정보에 직접 접근이 불가하며, 고유수용성 이력(Proprioception History)만을 통해 마치 특권 정보가 있는 것처럼 환경 정보를 추론해야 한다.

교사-학생 학습 기법에는 크게 두 가지 학습 방법이 있다. 첫 번째는 잠재 표현 증류 방법으로, 학생 모델이 교사 모델의 환경 잠재 표현을 모방하도록 최적화하는 접근법이다. 이 방법은 교사 모델에서 특권 정보의 인코딩 결과인 환경 잠재 표현과 학생 모델에서 고유수용성 이력의 인코딩인 환경 잠재 표현 간의 오차를 최소화하며, DAgger 알고리즘 [4]을 통해 학습된다. 이는 지식 증류의 일종으로 볼 수 있는데 교사의 잠재적 지식을 학생에게 전달하는 메커니즘이다. 두 번째는 행동 복제 방법으로, 학생 모델이 교사 모델의 최종 행동을 직접 모방하도록 최적화하는 접근법이다. 교사 모델의 행동과 학생 모델의 행동 간의 차이를 최소화하며, 동일하게 DAgger 알고리즘을 통해 학습된다. 이는 모방 학습(Imitation Learning)의 일종으로, 잠재 표현보다는 최종 행동 자체에 초점을 둔다.

각 학생 모델의 환경 잠재 표현이 특권 정보를 얼마나 효과적으로 인코딩하는지 검증하기 위해, 추가적인 디코더 네트워크를 설계하여 잠재 표현으로부터 원래의 특권 정보를 재구성하는 능력을 정량적으로 평가한다. 다층 퍼셉트론 구조의 디코더가 환경 잠재 표현을 입력

† 교신저자: 한연희 (yhhan@koreatech.ac.kr)



[그림 1] NVIDIA IsaacLab에서 로봇 보행 제어를 학습하는 모습

받아 원본 특권 정보를 예측하고, 원본 특권 정보와 재구성된 특권 정보 간의 오차를 평균 절대 오차로 성능을 측정한다. 디코더 학습은 학생 모델 학습과 동시에 진행되어, 실시간으로 잠재 표현의 품질을 평가할 수 있다. 이러한 접근을 통해, 특권 정보의 재구성 정확도와 로봇 제어 성능 간의 관계를 분석한다.

Ⅲ. 실험

모든 실험은 NVIDIA IsaacLab 시뮬레이션 환경 [5, 6]에서 수행되었으며, 사용된 로봇은 Unitree Go1 사족보행 로봇이다. 그리고 도메인 무작위화(Domain Randomization) 항목이자 특권 정보는 표 1의 범위에서 샘플링되었다. 이 외에도 특권 정보로 주지 않지만, 로봇의 실제 배치 성능을 높이기 위한 도메인 무작위 항목들로 PD 컨트롤러의 Damping, Stiffness, 모터의 마찰 계수, 센서 지연시간 등이 있다.

교사 모델은 PPO 알고리즘 [7]을 이용해 보행 안정성, 에너지 효율성, 그리고 목표 속도 추종 성능을 종합적으로 고려한 보상 함수에 따라 학습되었다. 두 학생 모델은 DAgger 알고리즘을 활용하여 잠재 표현 증류 방식과 행동 복제 방식으로 각각 학습되었다.

모델의 성능 평가는 잠재 표현 디코딩을 통한 특권 정보의 재구성 정확도와 보행 제어 성능을 나타내는 에피소드 보상을 기준으로 수행하였다. 표 1은 각 평가 지표에 대한 실험 결과를 제시하며, 모든 수치는 10개의 에피소드 평균으로 계산되었다. 교사 모델(Teacher)은 오라클 기준선으로서 특권 정보 재구성 성능과 보행 제어 성능의 상한을 나타낸다. 주목할 만한 점은 학생 모델 간의 성능 비교이다. 잠재 표현 증류를 통해 학습된 학생 모델(Student-L)은 행동 복제 방식으로 학습된 학생 모델(Student-A)에 비해 모든 특권 정보 항목에서 더 낮은 재구성 오차를 기록하였다. 이는 잠재 표현 증류 방식이 특권 정보의 암묵적 표현을 효과적으로 학습하는 데 유리함을 시사한다. 그러나 실질적인 제어 성능 지표인 에피소드 보상에서는 오히려 Student-A 모델이 더 우수한 성능을 보였다. 이러한 결과는 특권 정보 재구성 정확도와 실제 보행 제어 성능 사이에 반드시 정비례 관계가 존재하지 않음을 보여준다. 다시 말해, 특권 정보를 보다 정확히 재구성할 수 있는 모델이 항상 더 나은 제어 성능을 발휘하는 것은 아님을 시사한다.

Ⅳ. 결론

본 논문은 사족보행 로봇의 보행 제어 성능 향상을 위한 교사-학생 학습 기법 내 두 가지 접근법, 즉 잠재 표현 증류와 행동 복제의 성능을 정량적으로 비교하였다. 실험 결과, 잠재 표현 증류 방식은 특권 정보를 잠재 공간에 효과적으로 내재화하는 데 유리하며, 재구성 정확도 측면에서 우수한 성과를 나타냈다. 반면, 행동 복제 방식은 특권 정보의 직접적인 내재화 없이도 실제 보행 제어 성능, 즉 에피소드 보상에서 더 뛰어난 결과를 보였다.

[표 1] 각 모델의 특권 정보 예측 오차 및 에피소드 보상

	샘플링 범위	Teacher	Student-L	Student-A
질량	-1kg~6kg	<1g	620g	740g
무게중심(x)	-10cm~10cm	<1mm	1.6cm	2.1cm
무게중심(y)	-5cm~5cm	<1mm	0.8cm	1.1cm
몸체 선속도	0m/s~1m/s	<0.002m/s	0.03m/s	0.07m/s
지면 마찰 계수	0.2~2.1	<0.004	0.2	0.28
에피소드 보상	-	41.6±1.2	34.2±4.7	39.4±3.1

이는 로봇이 특권 정보를 얼마나 정확하게 복원하는지와 실제 제어 성능 사이의 관계가 단순히 정비례하지 않음을 명확히 보여준다. 특히, 실제 운용 환경에서는 특권 정보에 직접 접근할 수 없는 상황이 일반적이므로, 행동 복제와 같은 접근이 실용적 관점에서 더욱 효과적일 수 있음을 시사한다. 따라서 향후 로봇 보행 제어 학습에서는 내재화된 특권 정보의 품질뿐만 아니라, 최종 제어 성능과의 연계를 고려한 균형 잡힌 학습 전략 수립이 중요함을 강조한다.

이러한 결과는 로봇 제어 분야에서 지식 증류 기법의 적용 가능성을 재조명하고, 실용적 강화학습 기반 제어 시스템 설계에 있어 방향성을 제시한다.

ACKNOWLEDGMENT

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2018R1A6A1A03025526).

참 고 문 헌

[1] Zhuang, Z., Fu, Z., Wang, J., Atkeson, C., Schwertfeger, S., Finn, C., and Zhao, H. "Robot parkour learning," arXiv preprint arXiv:2309.05665, 2023.

[2] Cheng, X., Shi, K., Agarwal, A., and Pathak, D. "Extreme parkour with legged robots," 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 11443-11450, May 2024.

[3] Hinton, G., Vinyals, O., and Dean, J. "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[4] Ross, S., Gordon, G., and Bagnell, D. "A reduction of imitation learning and structured prediction to no-regret online learning," Proc. of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 627-635, June 2011.

[5] Rudin, N., Hoeller, D., Reist, P., and Hutter, M. "Learning to walk in minutes using massively parallel deep reinforcement learning," Conference on Robot Learning, pp. 91-100, Jan. 2022.

[6] Mittal, M., Yu, C., Yu, Q., Liu, J., Rudin, N., Hoeller, D., Yuan, J. L., Singh, R., Guo, Y., Mazhar, H., Mandlekar, A., Babich, B., State, G., Hutter, M., and Garg, A. "Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments," IEEE Robotics and Automation Letters, vol. 8, no. 6, pp. 3740-3747, 2023. (doi:10.1109/LRA.2023.3270034)

[7] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.