다중 에이전트의 효율적인 행동을 위한 마스터-슬레이브 정책 네트워크

김기범 O[1,4], 김은솔 [2], 최진영 [3], 장병탁 [1,2,3,4]

서울대학교 뇌과학 협동과정 [1], 서울대학교 컴퓨터공학부 [2],
서울대학교 인지과학 협동과정 [3], Surromind Robotics[4]
{kbkim, eskim, jychoi, btzhang}@bi.snu.ac.kr

Master-Slave Policy Network for Learning Multi-Agent's Behavior

Kibeom Kim O[1,4], Eun-Sol Kim[2], Jin-young Choi[3], Byoung-tak Zhang[1,2,3,4]

Interdisciplinary Program in Neuroscience, Seoul National University[1]
School of Computer Science and Engineering, Seoul National University[2],
Interdisciplinary Program in Cognitive Science[3], Surromind Robotics[4]

요 약

본 논문에서는 두 개 이상의 에이전트가 협동하면서 높은 보상을 얻을 수 있는 정책을 학습하는 새로운 강화학습 알고리즘을 소개한다. 전통적인 다중 에이전트 시스템과 비교하여 제안하는 알고리즘은 master-slave 구조를 새롭게 제안한다. 해당 구조는 각 에이전트들이 가지고 있는 부분적으로 관측가능한 상황들을 종합함으로써 궁극적인 정책을 학습하는 master 정책 네트워크와 행동을 하는 slave 정책 네트워크로 구성된다. 특히 기존의 다중 에이전트 시스템과 비교하여 에이전트의 수가 유동적인 상황에서도 쉽게 적용이 가능하며, 전체적인 상황을 고려하여 종합적인 판단을 할 수 있다는 점에서 장점을 보인다. 제안하는 알고리즘을 스타크래프트2의 미니 게임에 적용한 사례를 실험으로 보이고, 전반적인 상황 판단으로 각각의 에이전트가 효율적으로 행동함으로써 보상을 극대화 하는데 중요한 역할을 함을 보인다.

1. 서 론

DQN[1] 이후 신경망 구조와 강화 학습 알고리즘을 결합하는 연구가 진행되어 왔다. 특히 Atari와 같은 고전 게임을 자동으로 수행할 수 있는 방법들이 소개되면서 많은 주목을 받았다[2-3]. 또한 바둑 문제에서 사람보다 뛰어난 승률을 보이는 연구가 제시되면서[4], 게임 도메인에서의 강화학습 알고리즘에 대한 연구가 급속도로 진행되고 있다. 하지만 제안된 대부분의 연구들이 하나의 에이전트가 학습을 진행한다는 한계가 있다.

과거부터 연구가 꾸준히 이루어지고 있는 다중 에이전트 강화학습(MARL) 문제는 둘 이상의 에이전트가 효율적인 행동을 통해 하나의 에이전트로는 해결하기 힘든 문제를 효과적으로 해결하기 위한 연구로 이루어지고 있다. 이런 유형은 게임이론에서 협조적 게임 유형에 해당하며, 둘 이상의 에이전트들이 상호협력 없이는 문제를 해결할 수 없거나, 상호협력을 통해 보상을 극대화하는 상황을 가리킨다. 대표적인 예로 딥마인드와 블리자드가 함께 공개한 스타크래프트2 학습 환경이 있다[5].

최근 다중 에이전트 강화 학습을 위해 다양한 연구가 이루어지고 있다. 그 중 대표적인 예로는 communication network[6] 가 있다. 각 에이전트의 파라미터와 전체 에이전트 파라미터의 평균을 입력 받음으로써 파라미터를 공유하여 에이전트 간의 행동을 위한 의사소통 과정이 발생한다. 파라미터를 공유함으로써 에이전트 간의 정보가 공유될 수 있지만, 각 에이전트의 제한적인 상황공유와 에이전트의 숫자가 고정적이라는 한계가 있다.

본 논문에서는 다중 에이전트 강화학습에서 에이전트 수가 고정인 한계를 벗어나 에이전트의 숫자가 유동적인 상황에서도 전체적인 상황을 공유함으로써 효율적인 행동을 학습할 수 있는 마스터 정책 네트워크를 제안한다.

2. Master-Slave Policy Network

일반적인 다중 에이전트 문제에서는 각 에이전트의 상

황 정보가 분산되어 있으며, 서로 공유되고 있지 않다. 따라서 각 에이전트들은 제한적인 상황만을 갖고 있으며, 그로 인해 효율적인 행동을 하기가 쉽지 않다. 본 논문에서는 각 에이전트의 행동이 모여 전체 시각에서 시너지 효과를 발휘할 수 있는 행동을 위해 서로 다른 에이전트들의 상황 정보를 종합하여, 각 에이전트의 상황에 적절한 행동을 취할 수 있는 방법을 제안한다.

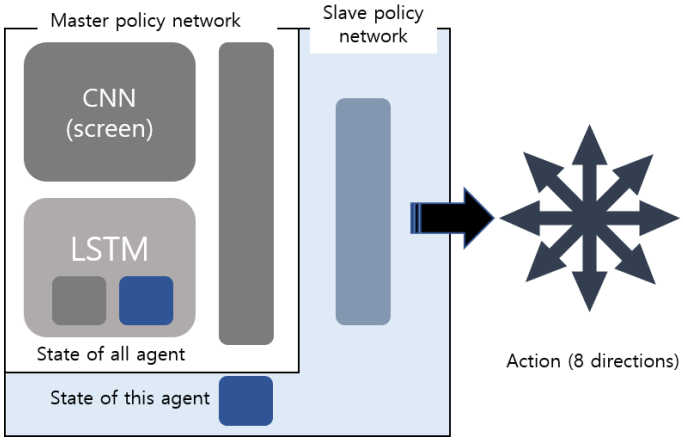

그림 1 마스터-슬레이브 정책 네트워크 구조

그림 1에 정책 네트워크 학습을 하기 위한 Master-Slave Network 구조도를 보인다. 해당 네트워크는 master 네트워크와 slave 네트워크로 구분된다.

우선 master 네트워크는 환경으로부터 정보를 받아 전체적인 상황을 판단한다. 본 논문에서 제안하는 master network는 실험 환경을 따라 환경으로부터 map정보와 각 에이전트의 state 값을 입력으로 한다. 이 때 map 정보를 취합하기 위하여 CNN (Convolutional Neural Network)을 사용하고, 에이전트의 상황정보는 LSTM으로 유동적일 수 있는 다중 에이전트들로부터 각 정보를 축약한다. 이후 추출한 두 정보를 취합하여 MLP(Multi-Layer Perceptron)에 입력하고, 학습함으로써 마스터 정책 네트워크가 구성되며, 256개의 연속 값으로 된 히든 노드 값이 마스터 네트워크의 출력이 된다.

이후 슬레이브 정책에서는 마스터 정책의 출력과 행동을 취할 에이전트의 상황정보를 이어 붙임으로써 입력으로 처리한다. 이후 MLP를 통해 학습을 하게 되고, 행동을 위해 가고자 하는 위치에 대한 정보를 출력으로 하기 위해 8방향 중 하나의 방향을 선택한다. 선택된 방향에 현재 위치를 반영하여 해당 방향의 일정 거리만큼의 좌표 값을 산출하여 해당 위치로 이동하는 행위를 하게 된다. 이로써 전반적인 상황을 인지하고 그에 적절한 행동을 유도 할 수 있게 된다. 따라서 종합적인 상황처리와 개별 에이전트들의 정보로 적절한 행동을 하도록 유도할 수 있다.

3. 실 험

마스터-슬레이브 정책을 위해 스타크래프트2 학습 환경에서 제안하는 미니게임 중 하나인 “Collect Mineral Shards”를 실험을 위한 가상 환경으로 선택하였다. 그림 2 에서 보는 것과 같이 이 환경은 두 마리의 유닛이 무작위로 흩어져 있는 여러 개의 미네랄 파편에 도달함으로써 보상을 얻는 게임으로, 두 마리의 유닛이 흩어져 움직임으로써 함께 다니는 것 보다 더욱 보상을 극대화 할 수 있다.

그림 2 스타크래프트2 학습환경 미니 게임 예시

실험에서 마스터 정책에서 처리하는 정보 중 CNN으로 처리를 하는 스크린 정보에는 64x64 크기의 특징 지도에서 미네랄 파편의 정보만 담긴 특징 지도를 추출한다. 그리고 각각의 에이전트의 위치 좌표를 추출 후, 특징지도는 3개 층으로 이루어진 CNN으로 입력하여 처리하고, 에이전트의 위치 좌표는 LSTM을 통해 처리하였다. 이후 마스터 정책에서 256개의 연속적인 상태 값을 출력하고, 슬레이브 정책 네트워크에서는 256 개의 마스터 정책 네트워크의 출력 값과 행동하고자 하는 2개의 해당 유닛의 위치 정보(x, y좌표)를 결합한 정보를 입력으로 한다. 이후 3 개 층의 히든 레이어를 가진 MLP로 네트워크를 구성하고, Action은 8 방향 중 하나의 방향을 선택, 해당 유닛의 위치 정보를 반영하여 해당 방향으로 일정한 거리가 떨어진 위치의 좌표 값을 산출한다. 해당 유닛은 산출한 위치 좌표로 이동하며, 이동하는 동안 미네랄 파편에 도달하거나 닿으면 +1의 보상을 얻게 된다.

실험에서는 학습을 위해 DQN을 사용하였으며, 데이터 간의 상관관계를 낮추고, 강건한 학습을 하기 위해 4개의 cpu 코어를 병렬처리를 통해 학습을 위한 replay memory를 함께 저장하고, 0.2초의 간격을 두고 계속 학습을 진행하는 프로세스와 테스트하는 프로세스를 따로 두고 학습하였다. 행동을 위해 각 유닛을 번갈아 가며 선택해주도록 설정하였으며, 행동에 대한 복잡도를

낮추기 위해 이동 좌표를 선택하는 것 외의 행동은 네트워크를 학습할 때 배제하였다. 테스트 동영상은 아래 링크를 통해 확인할 수 있다.
https://youtu.be/gr1Tlr8Zcso

표1. 스타크래프트2학습환경에서 제시하는 인간 및 에이전트의 기준치 및 마스터-슬레이브정책 네트워크 실험 결과

Agent	Max Reward
Random policy	35
Random Search	57
DeepMind Human player	142
StarCraft2 GrandMaster	179
FullyConv LSTM [5]	134
Master-Slave policy network (DQN)	117

4. 결론 및 향후 연구

효율적인 다중 에이전트 강화학습을 위해 마스터 정책에서 종합적인 상황 정보를 축약하고, 슬레이브 정책에서 그 정보와 행동을 취하는 에이전트의 정보를 통해 행동을 결정하는 마스터-슬레이브 정책을 제안하였다. 실험을 위해 스타크래프트2 학습 환경을 활용하였으며, 두 개의 에이전트의 효율적인 행동을 통해 보상을 극대화 하는 방안을 실험해보았다. 상대적으로 간단한 네트워크를 구축 후 실험 결과는 딥마인드에서 제안하는 기준치와 비슷한 수준에 다다른 것을 확인할 수 있었다.

향후 실험을 보완하기 위해 실험하지 않은 다른 스타크래프트2의 미니 게임에서도 실험을 할 예정이며, DQN 이외에도 추가로 다양한 학습 알고리즘으로 학습을 할 것이다. 또한 조금 더 어려울 수 있는 협력 게임을 위한 다중 에이전트 실험 환경을 새롭게 세팅할 예정이다.

감 사 의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업[2017-0-00162, 고령사회에 대응하기 위한 실환경 휴먼케어 로봇 기술개발]과 정보통신·방송 기술개발 사업[2015-0-00310-SW스타랩, 웨어러블센서기반 실생활 학습 자율지능 인지에이전트 SW]의 일환으로 수행하였음.

참 고 문 헌

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602. (2013).

[2] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In International Conference on Machine Learning (pp. 1928-1937). (2016, June).

[3] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. (2017).

[4] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489. (2016).

[5] Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., ... & Quan, J. Starcraft ii: A new challenge for reinforcement learning. arXiv preprint arXiv:1708.04782. (2017).

[6] Sukhbaatar, S., & Fergus, R. Learning multiagent communication with backpropagation. In Advances in Neural Information Processing Systems (pp. 2244-2252). (2016).