주의 집중 통신을 사용한 다중 에이전트 딥 강화학습

최진영[0] 이범진 장병탁
서울대학교 바이오지능 연구실
(jychoi, bjlee)@bi.snu.ac.kr, btzhang@snu.ac.kr

Attentive Communication for Multi-agent Deep Reinforcement Learning

Jin-young Choi[1○], Beom-Jin Lee[2], Byoung-Tak Zhang[12]
[1] Interdisciplinary Program in Cognitive Science, [2] School of Computer Science and Engineering,
Seoul National University

요 약

딥 강화학습은 여러가지 과제에서 인간 수준의 학습 능력을 선보였다. 그러나 인간의 지각과는 달리 딥 강화학습 모델은 감각 정보에서 필요한 부분을 추출하여 사용하는 능력을 갖추지 못해 낮은 단계의 감각 정보 전체를 상태-행동 가치 함수에 직접 연결한다. 이런 차이 때문에 딥 강화학습 모델은 학습에 엄청난 양의 경험 샘플이 필요하다. 이에 우리는 다중 초점 주의 집중 네트워크를 제안한다. 제안되는 모델은 감각정보에서 과제 해결에 필요한 물체들을 찾아내고 주의를 집중시킬 수 있는 인간의 능력을 모방할 수 있다. 제안되는 모델은 기존의 딥 강화학습 모델과 비교하여 높은 성능과 약 3.5배의 샘플 효율성을 보여주었다.

1 서 론

딥 강화학습은 여러 가지 과제를 사람 수준으로 학습하는 등 뛰 어난 성능을 보여주고 있다. DQN[1]은 Atari 2600게임들을 인간 수준으로 학습하는 것에 성공했고, 로봇 팔의 제어에도 사용되어 높은 성능을 보였다[2]. 그러나 이러한 연구들은 보통 하나의 에이전트가 활동하는 상황을 가정하였고, 여러 에이전트가 서로 의사소통을 하며 협동하여 과제를 처리해야 하는 다중 에이전트 상황에서의 연구는 상대적으로 많이 이루어지지 않았다. 이에 우리는 주의 집중 통신을 사용한 다중 에이전트 딥 강화학습 모델을 제안한다. 제안되는 모델은 각각의 에이전트가 보내는 메시지 중 어떤 메시지를 받아들일 것인지 결정하는 주의 집중 층(layer)을 네트워크에 삽입하여 상황에 맞게 중요한 정보를 가진 에이전트와 통신을 주고 받을 수 있도록 하였다. 제안되는 모델은 우선 각각의 에이전트의 감각정보에서 특성(feature)을 추출한 뒤 어떤 에이전트의 특성이 문제를 푸는데 중요한 요소인지를 End-to-End로 학습한다. 제안되는 모델은 일반적인 DQN에 비해 월등히 빠른 학습 속도와 성능을 보였고, 최신 모델[3]에 비해서도 속도와 성능의 향상이 있었다.

2 주의 집중 통신을 사용한 딥 강화학습 모델

제안되는 모델은 그림 1에서 보듯 3개의 모듈로 구성되어 있다. 첫 번째 특성 추출 (feature extraction)모듈로, 각각의 에이전트들의 감각정보로부터 특성을 추출해낸다. 두 번째 모듈은 주의 집중으로, 각각의 에이전트가 어떤 에이전트의 특성을 참고해야 협동문제를 푸는데 적합한지 판단한다. 마지막 모듈은 상태-행동 가치 평가로, 주의 집중을 통해 파악한 현재의 상태로부터 어떤 행동의 가치 (누적된 보상)이 가장 큰 지 판단한다. 이 장에서는 우리 모델의 기초가 되는 딥 강화학습을 간략히 소개하고 각각의 모델에 대해 자세히 다루도록 한다.

2.1 딥 강화학습

우리는 딥 Q-학습[1]의 프레임워크를 사용하였다. 딥 Q-학습은 상태-행동 가치 함수를 근사하기 위해 인공신경망을 사용한다. s_t를 시간 t에서의 감각 정보, a_t를 시간 t에서 취한 행동, r_t를 시간 t에서 얻은 보상으로 정의할 때, 상태-행동 가치 함수를 근사하기 위한 신경망의 학습 식은 식 1과 같다.

$$L = \{r_t + \gamma * \max_{a'} Q'(s_{t+1}, a') - Q(s_t, a_t)\}^2 \quad (1)$$

식 1에서 γ는 디스카운트 계수를, a'는 시간 t+1에서

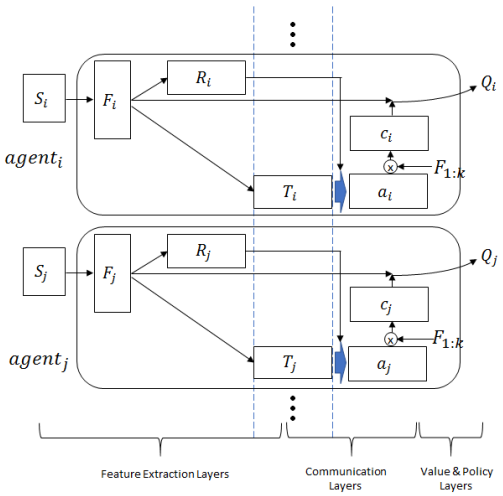

그림 1. 제안되는 모델의 구조

선택할 수 있는 행동들을, Q와 Q'은 각각 시간 t와 시간 t+1에서 네트워크를 사용해 계산한 가치 함수의 값을 뜻한다. 우리는 [1]과 마찬가지로 target network와 replay memory 테크닉을 활용하였다.

2.2 특성 추출 모듈

특성 추출 모듈은 각각의 에이전트의 감각정보 로부터 'Receiver'와 'Transmitter'로 정의된 특성을 추출한다. 'Receiver'는 어떤 정보가 필요한 지 결정하기 위한 특성이고 'Transmitter'는 각각의 에이전트가 어떤 정보를 가지고 있는지를 나타내도록 학습된다. 이는 아래 식으로 나타낼 수 있다.

$$\begin{aligned} \mathrm{R_i} &= \sigma_R(\mathrm{F_i}) \\ \mathrm{T_i} &= \sigma_T(\mathrm{F_i}) \end{aligned} \tag{2}$$

식 2에서 F_i는 i번째 에이전트의 감각정보(state)의 encoding, R_i 는 i번째 에이전트의 Receiver feature, T_i 는 i번째 에이전트의 Transmitter feature, σ_R , σ_T 는 추출 나타낸다. 본 연구의 실험에서 우리는 깊은 신경망을 σ_R , σ_T 로 사용하였다. 추출함수의 activation함수로는 ReLU[8]가 사용되었다.

2.3 주의 집중

추출된 특성으로부터 각각의 에이전트의 주의 집중 층이 식 3을 통해 어떤 에이전트의 정보에 얼마나 가중치를 줄 것인지 결정한다.

$$\mathrm{a_j^i} = \frac{\exp(R_i \cdot T_j^T)}{\sum_{j'} \exp(R_i \cdot T_{j'}^T)} \tag{3}$$

식 3에서 N은 주의 집중 층의 개수, a_j^i 는 i번째 에이전트에 대한 j번째 에이전트의 가중치이다.

2.4 상태-행동 가치 평가

주의 집중 층으로부터 얻은 주의 가중치들을 사용해 가중치가 반영된 'Communication Vector'를 식 4와 같이 정의한다.

$$\mathrm{C_i} = \sum_{j=1}^{k} F_j \cdot a_j^i \tag{4}$$

식 4에서 C_i 는 i번째 에이전트가 가치함수 계산에 참조할 Communication vector이다.

마지막으로, 계산된 특성을 식 5의 신경망 층에 입력하여 각 에이전트의 상태-행동 가치 함수를 계산한다.

$$Q_\mathrm{i} = \sigma_Q(\{F_i, C_i\}) \tag{5}$$

식 5에서 Q_i 는 식(1)에서 사용한 가치함수이고, σ_Q 는 신경망을 뜻한다. 실험에서는 fully-connected 층 하나와 이어진 linear 층 하나를 σ_Q로 사용하였다. Q의 마지막 층은 선택할 수 있는 행동의 개수 만큼의 출력 노드를 가지고 있어 모든 행동에 대한 가치 함수를 동시에 계산할 수 있다.

3 실험

3.1. 전투 시뮬레이션

우리는 모델의 성능을 검증하기 위하여 전투 시뮬레이션 게임을 실험 도메인으로 디자인했다. (그림 2) 게임은 15x15 크기의 격자공간 내에서 진행된다. 각각 5개체의 에이전트로 구성된 두 개의 팀이 전투를 벌이며, 피카츄 팀은 신경망 모델에 의해 조종되고 파이리 팀은 가장 가까이에 있는 지각 가능한 피카츄를 공격하도록 사전에 작성된 스크립트대로 조종된다. 각각의 에이전트는 주변 2칸까지의 범위에 무엇이 있는지 지각할 수 있다. 각각의 에이전트는 1칸 이내에 있는 적을 공격할 수 있으며, 각각의 에이전트는 3번의 공격을 받으면 게임에서 퇴장한다. 한 팀이 전멸하게 되면 게임이 끝나며, 80 time-step동안 결판이 나지 않을 경우 신경망이 조종하는 팀이 패배한 것으로 간주한다. 파이리 팀은 모든 지각 정보를 공유하기 때문에, 적절한 통신 프로토콜을 학습하지 못할 경우 피카츄 팀은 게임에서 승리할 수 없다. 이 환경에서 우리가 제안하는 모델, 최근 발표된 State-Of-The-Art 모델[3], 통신이 없는 일반적인 DQN, 모든 에이전트의 state를 concatenation하여 한번에 모든 에이전트의 Q value를 출력하는 DQN의 4가지 모델을 학습하여 성능(100판당 몇 판을 승리하는지의 승률)을 비교했다.

3.2. 정량적 분석

학습의 결과는 그림 3과 같다. 제안되는 모델은 학습 속도와 성능 면에서 모두 큰 향상을 보였다. 제안되는 모델은 State-Of-The-Art 모델[3]보다 약 20% 먼저 수렴하였고, 이를 통해 주의 집중 층이 에이전트 간 통신 프로토콜을 학습하는데 보다 효율적임을 입증하였다. 통신이 없는 DQN은 느린 학습속도와 현저히 낮은 승률을 보였으며, concatenation모델은 학습에 실패하였다.

3.3. 정성적 분석

게임 화면과 상응하는 주의 가중치를 분석한 결과 적을 감지한 에이전트에 많은 주의 집중이 이루어졌고, 5개체의 에이전트가 3개체, 2개체로 된 두 그룹으로 나뉘어 집중공격을 통해 적을 효율적으로 공격하는 것을 관찰할 수 있었다.

4. 고찰

우리는 효과적으로 다중 에이전트의 협동 과제를 해결하는데 주의 집중을 통한 통신 프로토콜 학습을 제안했다. 제안되는 모델은 실험에서 기존의 모델들에 비해 큰 성능 향상과 학습속도 향상을 보였다. 향후 연구에서 우리는 서로 다른 state-action space를 가진 에이전트 간에도 통신이 가능한 모델을 구현할 계획이다.

Acknowledgement

본 연구는 국방생체모방 자율로봇 특화연구센터를 통한 방위사업청과 국방과학연구소 연구비 지원으로

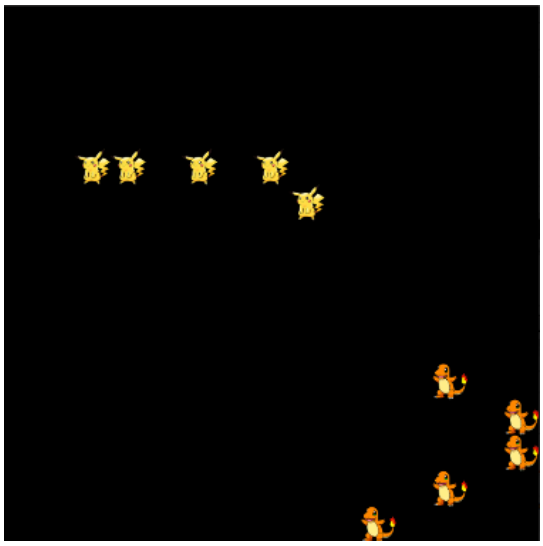

그림 2. 전투 시뮬레이션

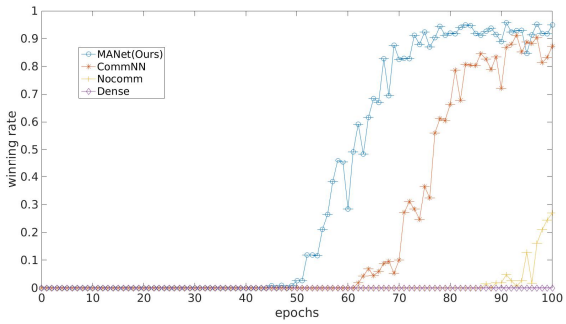

그림 3. 학습의 결과

수행되었습니다 (UD130070ID)

5. 참고문헌

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bel-lemare, M., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 02 2015.

[2] Zhang, F., Leitner, J., Milford, M., Upcroft, B., and Corke, P. 2015. Towards Vision-Based Deep Reinforcement Learning for Robotic Motion Control. Paper presented at the Australasian Conference on Robotics and Automation (ACRA) 2015

[3] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. In Advances in Neural Information Processing Systems, pages 2244–2252, 2016.