상태 행동 가치 기반 다중 에이전트 강화학습 알고리즘들의 비교 분석 실험

김주봉, 최호빈, 한연희[1]
한국기술교육대학교 미래융합공학전공/첨단기술연구소
{rlawnqhd, chb3350, yhhan}@koreatech.ac.kr

Comparative Analysis of Multi-Agent Reinforcement Learning Algorithms Based on Q-Value

Ju-Bong Kim, Ho-Bin Choi, Youn-Hee Han[1]
Future Convergence Engineering/Advanced Technology Research Center
Korea University of Technology and Education

요 약

시뮬레이션을 비롯한 많은 다중 에이전트 환경에서는 중앙 집중 훈련 및 분산 수행 (centralized training with decentralized execution; CTDE) 방식이 활용되고 있다. CTDE 방식 하에서 중앙 집중 훈련 및 분산 수행 환경에서의 다중 에이전트 학습을 위한 상태 행동 가치 기반(state-action value; Q-value) 다중 에이전트 알고리즘들에 대한 많은 연구가 이루어졌다. 이러한 알고리즘들은 Independent Q-learning (IQL)이라는 강력한 벤치 마크 알고리즘에서 파생되어 다중 에이전트의 공동의 상태 행동 가치의 분해(Decomposition) 문제에 대해 집중적으로 연구되었다. 본 논문에서는 앞선 연구들에 관한 알고리즘들에 대한 분석과 실용적이고 일반적인 도메인에서의 실험 분석을 통해 검증한다.

1. 서론

최근 다중 에이전트 강화학습 분야의 연구는 복잡한 현실의 문제를 해결하기 위한 노력이 지속되고 있다 [3-6]. 대부분의 복잡한 현실의 문제는 상태 및 행동 공간의 크기(state and action spaces)가 방대하기 때문에 단일 에이전트 문제로 제한하는 것에는 어려움이 따르며, 다중 에이전트 문제로의 전이가 불가피하다. 한편, 다중 에이전트 도메인(domain)에서는 에이전트 간 상호 소통의 제약 및 부분적 관찰(partially observable)의 특성 때문에 분산된 정책의 강화학습이 필수적으로 도입될 필요가 있다.

완전 협력 다중 에이전트 작업(fully cooperative multi-agent task)은 식 $G = \langle S, U, P, Z, O, n, \gamma \rangle$로 정의되는 Dec-POMDP 로 모델링 가능하다 [1]. S는 강화학습 환경에서의 상태 공간에 해당하며, U는 강화학습 에이전트의 행동 공간을 의미한다. 매 타임 스텝 t 마다, 에이전트 $i \in A \equiv \{1, \dots, N\}$ 는 해당 에이전트의 행동 $u^i \in U$을 선택하며, 각 에이전트의 행동을 공동 행동 $\mathbf{u} \in \mathbf{U} \in U^n$ 으로 정의한다. 상태 전이 확률 함수는 $P(s'|s, \mathbf{u}): S \times \mathbf{U} \times S \rightarrow [0, 1]$로 정의한다. 모든 에이전트는 공동의 보상을 받고, 공동 보상 함수는 $r(s, \mathbf{u}): S \times \mathbf{U} \rightarrow \mathbb{R}$로 정의되며, 보상 감쇄 인자는 $\gamma \in [0, 1)$이다. 모든 에이전트는 전지적 시점에서 환경을 내려다보는 경우는 드물며 대부분 현실 세계의 문제에서는 에이전트가 부분적인 상태만을 관찰가능하다. 따라서 에이전트는 환경으로부터 상태에 관한 관찰 $z \in Z$을 행동 제어에 활용하고 관찰 z는 관찰 함수 $O(s, i): S \times A \rightarrow Z$에 의한 결과 값이다. 에이전트 i에 대한 행동-관찰 히스토리(history)는 $\tau^i \in T \equiv (Z \times U)^*$로 정의되며, 여기서 τ는 에이전트의 정책 $\pi^i(u^i|\tau^i): T \times U \rightarrow [0, 1]$을 조정할 수 있다. 에이전트 i 이외의 에이전트들의 행동은 u^{-i}로 정의하며 이와 유사하게 i 이외의 에이전트들의 정책은 π^{-i}로 정의한다. 공동의 정책 π 는 상태 행동 가치 함수: $Q^{\pi}(s_t, \mathbf{u}_t) = \mathbb{E}_{s_{t+1:\infty}, \mathbf{u}_{t+1:\infty}}[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t, \mathbf{u}_t]$ 에 기반을 두며 탐욕적인 공동의 정책은 $\pi(s, \mathbf{u}) := argmax_{a \in A} Q(s, \mathbf{u})$ 로 표현된다. 협력적인 다중 에이

[1] 한연희(Youn-Hee Han, yhhan@koreatech.ac.kr): 교신저자

전트의 최종 목적은 최적의 상태 행동 가치 함수인 Q^*를 찾는 것이다. 중앙 집중 학습에서 알고리즘은 모든 에이전트의 행동-관찰 히스토리에 접근 가능하기 때문에 학습에 활용할 수 있다. 하지만 에이전트 각각은 분산된 행동 추론 중에 자체적인 행동-관찰 히스토리 τ^i에 대해서만 활용가능하다.

본 논문 2 장에서는 다중 에이전트 강화학습 알고리즘들에 대한 설명과 분석 결과를 제시하고, 3 장에서는 대상 알고리즘들에 대한 실험 평가와 함께 경험적 검증을 수행하며, 4 장에서 결론을 맺는다.

2. 상태 행동 가치 기반 협력적 다중 에이전트 강화학습 알고리즘

다중 에이전트 강화학습 알고리즘에서 가장 기본적인 학습 방식은 IQL 과 같이 에이전트 마다 매개 변수화된 상태 행동 가치 함수를 독립적으로 학습하는 것이다 [2]. 이러한 학습 방식은 간단하고 실용적이지만 무한한 탐험의 제한에서도 수렴을 보장하기 어려우며 에이전트 마다 갖는 서로 다른 정책에 대한 비정상성(non-stationarity)에 대한 설명이 불가하다. IQL 의 이러한 한계로 인해 이 후 연구들인 VDN, QMIX, QTRAN, MAVEN 등의 가치 기반(value-based) 다중 에이전트 강화학습 알고리즘들은 CTDE 를 활용한다.

CTDE 특성을 갖는 환경에서 다중 에이전트 강화학습 에이전트 알고리즘은 에이전트에 대한 상태 행동 가치를 개별적인 에이전트의 q_i 로 분해(factorization) 할 수 있다. 이러한 경우 각 에이전트들의 q_i는 기대 누적 보상 값(expected return)을 직접 추정하며 학습을 하는 것이 아니기 때문에 엄밀히 따지면 상태 행동 가치 함수가 아니며 유틸리티(utility) 함수 q_i 이다. CTDE 에서 다중 에이전트의 상태 행동 가치 $Q(s, \mathbf{u})$를 각 에이전트들의 q_i로 분해하여 분산성(decentralisability)을 강조한 표현은 다음 식 (1)이 반드시 성립된다.

$$argmax_u Q^*(s, \mathbf{u}) = \left(argmax_{u^1} q_1(\tau^1, u^1) \dots \ argmax_{u^n} q_n(\tau^n, u^n)\right) \quad (1)$$

가능한 모든 $s, \mathbf{u}$에서 각 에이전트의 유틸리티 함수 q_i 는 식 (1)를 만족하며 $Q(s, \mathbf{u})$ 로부터 분해된다, Individual-Global-Max (IGM) 참고 [5]. 다중 에이전트의 상태 행동 가치 $Q(s, \mathbf{u})$가 식 (1)을 만족하며 분해 가능할 때, 가산성(additivity)과 단조성(monotonicity) 두 가지 특성을 고려할 수 있다, 각각 식 (2), (3) 참조.

$$\text{Additivity} \qquad Q(s, \mathbf{u}) = \sum_{i=1}^{n} q_i(\tau^i, u^i) \quad (2)$$

$$\text{Monotonicity} \qquad \frac{\partial Q(s,\mathbf{u})}{\partial q_i(\tau^i, u^i)} \geq 0, i \in A \equiv \{1, \dots, N\} \quad (3)$$

식 (2)의 경우 분해(factorization)는 다중 에이전트의 유틸리티 함수 값의 합으로 계산되어 제한된다. 한편 식 (3)의 각 에이전트의 추론된 행동 u^i에 대한 유틸리티 함수 값을 결합할 때 활용하는 혼합 네트워크(mixing network)의 가중치가 양의 실수라 가정 할 때, 식 (3)의 왼쪽 변이 항상 양의 값을 갖는다. 이러한 특성은 모든 에이전트의 유틸리티 함수가 추론하는 행동들이 결합된 형태인 공동 행동(joint action) $\mathbf{u}$ 에 관한 상태 행동 가치 함수가 업데이트 되는 방향에 영향을 주어 기여한 에이전트 각각이 업데이트 가능함을 의미한다.

2.1. Value Decomposition Networks (VDN)

VDN 알고리즘은 다중 에이전트의 상태 행동 가치 함수 $Q_{VDN}(s, \mathbf{u})$를 학습하는 것이 목적이다 [3]. 식 (2)와 같이 다중 에이전트의 상태 행동 가치를 각 에이전트들의 개별적 유틸리티 함수 q_i의 합인 $Q_{VDN}(s, \mathbf{u})$으로 구성한다. 때문에 VDN 에서 $Q_{VDN}(s, \mathbf{u})$는 역으로 개별적 유틸리티 함수들로 분해(decomposition) 가능하며, 이는 식 (2)을 보장함과 동시에 식 (1) 또한 만족한다 [3]. VDN 의 이러한 $Q_{VDN}(s, \mathbf{u})$의 분해는 자연스럽게 각 에이전트들의 개별적 q_i가 에이전트의 독립적인 상태 행동 가치로 추상화 할 수 있으며, 훈련 시 에이전트들 공동의 상태 행동 가치에 얼마나 기여를 하는지 추론가능하다.

2.2. QMIX

QMIX 알고리즘은 VDN 과 마찬가지로 다중 에이전트의 상태 행동 가치 함수 $Q_{QMIX}(s, \mathbf{u})$를 학습하는 것이 목적이다 [4]. 하지만 VDN 과 달리 단순히 각 에이전트들의 개별적 유틸리티 함수 q_i의 합계로 구성하지 않는다. QMIX 에서는 $Q_{QMIX}(s, \mathbf{u})$를 에이전트들의 개별적으로 학습되는 유틸리티 함수 q_i의 단조 비선형 조합(monotonic non-linear combination)으로 구성한다. 즉, 모든 개별 에이전트가 선택한 행동들에 관한 유틸리티 함수 값들을 음이 아닌(non-negative) 가중치(weights)를 갖는 혼합 네트워크가 $Q_{QMIX}(s, \mathbf{u})$로 결합한다. 혼합 네트워크는 음이 아닌 성질 (non-negativity) 때문에 식 (3)을 보장한다. 그렇기 때문에 QMIX 에서 혼합 네트워크에 의한 개별 에이전트들의 유틸리티 함수의 결합은 식 (1) 또한 만족한다 [4]. QMIX 와 VDN 각각의 $Q_{QMIX}(s, \mathbf{u})$, $Q_{VDN}(s, \mathbf{u})$ 에서 개별 에이전트의 유틸리티 함수 q_i 로의 분해는 독립적인 에이전트의 구성을 용이하게 하며, IGM 의 특성을 만족하

는 단조로운(monotonic) 협력적 다중 에이전트 작업 (cooperative multi-agent task)에서 수렴 가능하다.

2.3. QTRAN: Learning to Factorize with Transformation

QTRAN 은 VDN, QMIX 와 마찬가지로 가치 기반 다중 에이전트 알고리즘이다 [5]. QTRAN 논문에서는 QTRAN-base, QTRAN-alt 두 알고리즘을 제시한다. VDN 과 QMIX 알고리즘의 제한 사항인 단조성 (monotonicity)에 대한 특별한 언급은 논문 내에 없지만, 반례를 예시로 들어 VDN 과 QMIX 의 단점을 극복한다. QTRAN 은 다중 에이전트의 상태 행동 가치 함수에서 개별적 유틸리티 함수로의 분해 사이에 [5] 의 에 기술된 수식과 같은 선형적인 제한을 걸어 최적의 분해를 목표로 한다.

2.4. MAVEN: Multi-Agent Variational Exploration

MAVEN 역시 본 장에 언급된 다른 알고리즘들과 마찬가지로 가치 기반 다중 에이전트 알고리즘이다 [6]. MAVEN 은 QTRAN 과 마찬가지로 VDN 과 QMIX 의 단조 제약의 개선방안을 제시한다. 하지만 QTRAN 과는 달리 다중 에이전트들의 상태 행동 가치에 대한 비선형적 변형 탐색(variational exploration)을 통해 단조 제약을 극복한다. 한편 MAVEN 에서는 다중 에이전트 작업 관점의 비단조성(non-monotonicity)에 대한 명확한 정의를 제시한다, [6]의 $Definition\ 1$ 참고. 이를 시작으로 MAVEN 에서는 $Theorem\ 1, 2$를 제시하여 QMIX 는 비단조성을 갖는 협력적 다중 에이전트 작업에서 높은 확률로 차선책(suboptimality)에 수렴됨을 증명한다.

3. 실험적 분석

본 장에서는 이 전 장에서 언급된 알고리즘 VDN, QMIX, QTRAN-base, QTRAN-alt, MAVEN 에 대한 경험적 실험 평가를 한다. 실험에 활용되는 도메인(domain)은 payoff matrix game 이며 두 가지 환경(one-step payoff matrix game, m-step payoff matrix game)으로 나누어 사용한다. One-step payoff matrix game 에서는 비단조 특성을 갖는 도메인에서 각각의 알고리즘에 의해 상태 행동 가치 함수 값의 정확한 추론이 가능한지 검증한다 [5, 6]. m-step payoff matrix game 에서는 탐험의 중요성이 강조되며 비단조 특성을 갖는 도메인에서 각각의 알고리즘에 의해 얼마나 신속하게 수렴이 되는지를 검증한다 [6].

3.1. One-step Payoff Matrix Game

One-step payoff matrix game 은 $n = 2, |u_1| = 3, |u_2| = 3$ 이며, 각 에이전트에 의한 공동 보상은 아래 그림 1 (a)에 표현된다. 한 에피소드의 최대 스텝은 1 이고, $t = 0$에서 상태 s에 대해 두 에이전트에 의해 선택된 공동 행동 $\mathbf{u}_t$이 적용되어 공동 보상이 주어지면 에피소드는 종료된다. 두 에이전트의 관찰 o^1, o^2 각각은 s와 동일하며 최대 스텝에 대한 $onehot\ vector$ 형태로 주어진다.

(a) One-step payoff matrix game

u_1 \ u_2	A	B	C
A	10	0	10
B	0	**10.4**	0
C	10	0	10

(b) VDN

u_1 \ u_2	A	B	C
A	7.84	4.32	7.97
B	4.01	**0.49**	4.13
C	7.83	4.31	7.95

(c) QMIX

u_1 \ u_2	A	B	C
A	9.99	1.85	10.00
B	1.85	**1.83**	1.85
C	10.00	1.85	10.00

(d) QTRAN-base

u_1 \ u_2	A	B	C
A	9.72	-0.07	9.71
B	-0.06	**10.28**	-0.06
C	9.71	-0.06	9.70

(e) QTRAN-alt

u_1 \ u_2	A	B	C
A	9.92	-0.03	9.92
B	-0.02	**10.29**	-0.03
C	9.92	-0.03	9.92

(그림 1) One-step payoff matrix game ($\boldsymbol{a}$)과 VDN ($\boldsymbol{b}$), QMIX ($\boldsymbol{c}$), QTRAN-base ($\boldsymbol{d}$), QTRAN-alt ($\boldsymbol{e}$) 알고리즘들의 상태 행동 가치 함수 $\boldsymbol{Q}(\boldsymbol{s}, \mathbf{u})$ 결과 값.

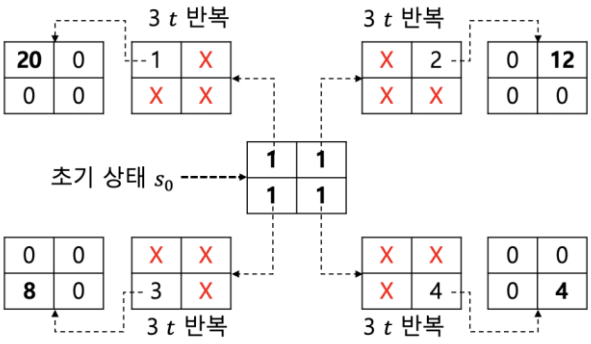
그림 1 은 one-step payoff matrix game 환경의 보상 방식 (a)과 알고리즘 VDN, QMIX, QTRAN-base, QTRAN-alt 에 의해 학습된 각각의 상태 행동 가치 함수 $Q(s, \mathbf{u})$ 평가 값(evaluated value)이다, 그림 1 의 (b~e) 참조. 실험에선 대상 알고리즘들에 대해 각각 10,000 번의 에피소드 훈련을 수행한다.

그림 1 에서 확인할 수 있듯 VDN 과 QMIX 는 정확한 상태 행동 가치 값을 추론하지 못하고 있지만 QTRAN-base , QTRAN-alt 는 비교적 정확히 추론하고 있는 경향을 보인다. 이러한 결과가 도출된 이유는 VDN 의 표현의 한계(representational limitation)인 상태 행동 가치의 분해를 가산성(additivity)의 원리에 근거하여 각 에이전트의 유틸리티 함수를 구성했기 때문이라 분석된다. 그리고 QMIX 가 VDN 보다 비교적 좋은 결과를 보이는 이유는 VDN 의 상태 행동 가치 분해에 활용된 가산성의 원리를 유지하며 비선형 단조 결합을 사용했기 때문이며, 경험적으로 더 신속하게 수렴하는 경향을 보인다. 결과적으로 QTRAN 은 VDN 과 QMIX 의 표현의 한계를 유지하며 상태 행동 가치 업데이트에 선형 제약 조건(linear constraint condition)을 추가하였고 QTRAN-alt 에 활용되는 $L2$ 패널티(penalties)를 사용해 제약을 완화했기 때문에 QMIX 의 한계를 극복했다는 것이 경험적 실험을 통해 증명된다.

3.2. m-step Payoff Matrix Game

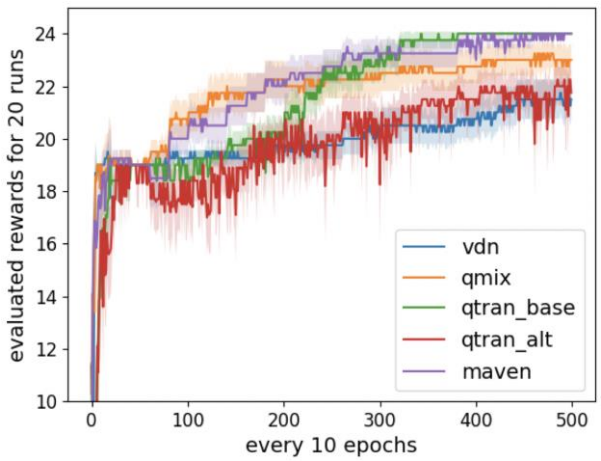
그림 2 에 표현된 m-step payoff matrix game 은 $n = 2, |u_1| = 2, |u_2| = 2$ 이며, 각 에이전트에 의한 공동 보상은 그림의 표 내부에 숫자로 표현된다. 실험에 활

용된 환경의 한 에피소드의 최대 스텝은 $m = 5$이고, $t = 0$에서 초기 상태 s_0에 대해 두 에이전트에 의해 선택된 공동 행동 $\mathbf{u}_t$이 적용되어 어떤 경우든 1의 보상이 주어지고 선택된 행동의 조합에 따라 다음 중간 상태가 결정된다. 중간 상태는 최대 $m-2$번 반복되며 중간 상태의 'X' 값에 해당되는 공동 행동이 결정되면 에피소드가 종료되며 0의 보상이 각 에이전트에게 주어진다. 두 에이전트의 관찰 o^1, o^2 각각은 s와 동일하며 $4m+5$에 대한 *onehot vector* 형태로 주어진다.



(그림 2) $\boldsymbol{m}$-step payoff matrix game, $\boldsymbol{m = 5}$.

실험은 평가 대상 알고리즘 VDN, QMIX, QTRAN-base, QTRAN-alt, MAVEN 에 대한 20 번의 반복 훈련으로 구성되며, 1 번의 훈련 당 5,000 번의 에피소드 학습이 수행된다. 그림 3 은 대상 알고리즘 각각 10 에피소드 마다 평가된 평균 에피소드 보상에 대한 결과 그래프이다.



(그림 3) 20 번의 반복 훈련 실험 평가에서 대상 알고리즘 각각 10 에피소드 마다 평가된 평균 에피소드 보상, 음영 부분은 반복 실험의 표준편차를 평균 값에 가감한 상한과 하한에 대한 표현.

그림 3 에 관한 실험에서 MAVEN 이 가장 안정적인 $Q_{MAVEN}(s, \mathbf{u})$의 학습을 보였고, QMIX 는 학습 초기 빠른 속도로 수렴되는 듯 하나 $Q_{QMIX}(s, \mathbf{u})$의 학습이 최적에 보상 값에 수렴이 되지 않는다. 한편, QTRAN-base 는 최적 보상에 근접한 $Q_{QTRAN-base}(s, \mathbf{u})$의 학습이 이루어짐이 확인된다.

4. 결론

본 논문은 협력적 다중 에이전트 강화학습 알고리즘의 대표적 벤치 마크인 IQL 로부터 파생된 가치-기반에 중점을 둔 알고리즘 VDN, QMIX, QTRAN, MAVEN 에 대한 분석과 경험적 실험 평가로 각 알고리즘에 대한 검증을 수행했다. 언급된 알고리즘들은 연구된 시대순으로 나열되어있기 때문에 갈수록 발전하는 듯한 경향을 보였으나 각 알고리즘이 중점을 두어 시사하는 바에 명확한 차이가 존재하기 때문에 각 알고리즘의 한계점에 대한 깊은 연구가 지속되어야 할 필요가 있음이 경험적 실험을 통해 드러난다.

Acknowledgement

이 논문은 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업 (No. NRF-2020R1I1A3065610) 이며, 또한 2018 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2018R1A6A1A03025526).

참고문헌

[1] Oliehoek, F. A. and Amato, C, "A Concise Introduction to Decentralized POMDPs," SpringerBriefs in Intelligent Systems. Springer, 2016.

[2] Tan, M., "Multi-agent reinforcement learning: Independent vs. cooperative agents," In Proceedings of the Tenth International Conference on Machine Learning, pp. 330–337, 1993.

[3] Sunehag, P. et al., "Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward," In Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, 2017.

[4] Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S., "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," In Proceedings of the 35th ICML, 2018.

[5] Kyunghwan, S., Daewoo K., Wanju K., David H., and Yung Y., "QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning," In Proceedings of the 36th ICML, 2019.

[6] Mahajan, A., Rashid, T., Samvelyan, M., Whiteson, S., "MAVEN: Multi-Agent Variational Exploration," NeurIPS, pp. 7611-7622, 2019.