지능형 무인 창고에서 QMIX 를 이용한 무인운반기 협력 제어

최호빈, 김주봉*, 김찬명**, 한연희*
한국기술교육대학교 컴퓨터공학과, *한국기술교육대학교 미래융합공학전공,
**한국기술교육대학교 첨단기술연구소

{chb3350, *rlawnqhd, **cmdr, *yhhan}@koreatech.ac.kr

Cooperative Control of Automated Guided Vehicles Using QMIX in Intelligent Unmanned Warehouse

Ho-Bin Choi, Ju-Bong Kim*, Chan-Myung Kim**, Youn-Hee Han*
Department of Computer Science and Engineering, Korea University of Technology and Education,
*Future Convergence Engineering Major, Korea University of Technology and Education,
**Advanced Technology Research Center, Korea University of Technology and Education

요 약

온라인 쇼핑의 주문이 늘어남에 따라 물류창고의 규모가 커지게 되었고 대규모의 물류창고를 효과적으로 운영하기 위해 무인운반기가 도입되고 있다. 다수의 무인운반기가 독립적으로 움직이는 물류창고에서 시스템 전체의 성능을 극대화하기 위해 멀티 에이전트 강화 학습을 사용하여 각 무인운반기를 학습시킬 수 있다. 각 무인운반기는 다른 무인운반기와의 충돌을 피하며 자신의 임무를 가능한 한 빠르게 완료해야 한다. 본 논문에서는 강화 학습 환경을 설계하고 대표적인 멀티 에이전트 강화 학습 알고리즘 QMIX 를 사용하여 실험을 통해 무인운반기의 성공적인 제어가 가능함을 입증한다.

Ⅰ. 서 론

무인운반기(Automated Guided Vehicle, AGV)는 대규모 물류창고 무인화의 핵심이며 연구 및 적용이 가속화되고 있다 [1]. 무인운반기로 물류창고의 모든 과정을 무인화 할 수는 없지만 운반을 기능을 완벽하게 수행할 수 있어 인력 부족 문제를 해결할 수 있다 [2]. 즉, 사람이 피킹(Picking), 포장(Packing) 및 배송(Shipping) 등의 다른 업무에만 집중할 수 있도록 한다. 특히, 무거운 물건들을 빠르게 운반해야 하는 물류창고에서는 그 필요성이 부각된다.

강화 학습은 미리 준비된 학습 데이터 없이 에이전트가 환경을 탐험하며 학습 데이터를 만들어 학습하는 방식이다. 자세하게, 강화 학습 에이전트는 현재 State 에서 자신의 정책을 기반으로 한 Action 을 취한 후, 환경으로부터 Reward 를 받고 이 세 가지의 State, Action, Reward 정보를 활용해 기대 Reward 를 최대화 하는 방향으로 정책을 업데이트한다. 이 과정을 반복하며 에이전트의 정교한 정책이 수립된다. 한 개의 에이전트만이 존재하는 강화 학습은 싱글 에이전트 강화 학습이라 부르며 비교적 어렵지 않게 학습을 진행할 수 있다. 하지만, 여러 개의 에이전트가 존재하는 강화 학습인 멀티 에이전트 강화 학습은 싱글 에이전트 강화 학습과 다르게 학습에 여러 어려움이 따른다. 대표적인 멀티 에이전트 강화 학습 알고리즘으로 QMIX 가 널리 알려져 있으며 여러 변형이 발표되고 있다 [3]. 무인운반기 창고와 같은 환경은 시스템의 전체적인 성능을 위해 모든 에이전트들이 협력하여 문제를 풀어야 하며 QMIX 는 그러한 목적에 적합하다. QMIX 는 각 에이전트의 정책과 더불어 모든 에이전트들의 출력 값 및 전역 State 를 입력으로 받는 공동 정책이 존재한다. 이 공동 정책을 모든 에이전트 정책들의 합이라 가정하고 학습시키며 각 에이전트의 정책에 영향을 미친다.

Ⅱ. 본론

본 절에서는 강화 학습 환경의 시나리오와 State, Action, Reward 에 대해 자세히 설명한다. 각 무인운반기는 독립적인 에이전트가 되며 공유되는 하나의 정책을 통해 행동을 결정한다. 해당 정책과 더불어 공동 정책의 기대 보상을 최대화하므로 시스템 전체의 성능 향상을 유도한다.

1. 시나리오

물류창고에는 무인운반기, 무인운반기 대기 장소, 선반, 선반 저장소, 피킹 스테이션 및 노동자가 존재한다 ([그림 1] 참조). 본 연구의 초점은 무인운반기의 제어에 맞춰져 있으므로 무인운반기의 제어를 제외한 다른 작업들은 무작위 방식으로 작동하도록 하였다. 즉,

필요한 선반 할당, 해당 선반을 운반할 무인운반기 할당, 피킹 작업이 완료된 선반을 위한 선반 저장소 선택 그리고 임무를 마친 무인운반기가 대기할 무인운반기 장소 선택은 가능한 범위에서 무작위로 결정된다. 나머지 모든 이동 또는 운반 작업들은 QMIX 를 사용한 멀티 에이전트 강화 학습으로 제어되며 이동 또는 운반 작업마다 출발지와 목적지가 설정된다. 모든 무인운반기는 배터리가 충분하다고 가정하며 중앙시스템에 동기화되어 일정 시간 단위로 동시에 제어된다.

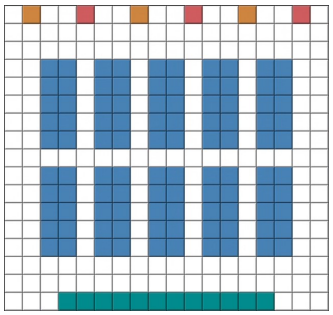

그림 1. 레이아웃 예제

2. State

각 에이전트의 State 는 자신을 중심으로 한 5x5 크기의 지역 정보로 구성된다. 세 개의 채널로 이루어져 3x5x5 의 Shape 으로 표현되며 각 픽셀에는 해당 위치를 표현하는 정보가 들어있다. 해당 정보는 0.0 에서 1.0 사이의 실수로 표현된다. 첫 번째 채널에는 무인운반기를 제외하고 어떤 모듈이 어느 위치에 존재하는지의 정보가 들어있다. 이 정보를 통해 기본적인 길과 장애물의 위치를 판단할 수 있다. 두 번째 채널에는 에이전트인 무인운반기들이 어느 위치에 존재하는지의 정보가 들어있다. 이 정보를 통해 에이전트들 간의 충돌을 방지할 수 있다. 세 번째 채널에는 목적지까지 1.0 에 가까워지는 경사 정보가 들어있다. 이 정보를 통해 나아가야 하는 방향을 알 수 있다. 각 에이전트는 타임 스텝마다 이러한 세 채널의 함축된 정보를 사용하여 자신의 정책을 따르는 Action 을 선택한다.

3. Action

각 에이전트는 타임 스텝마다 자신의 정책을 따르는 행동을 수행한다. Action Space 는 {Stop, Move Forward, Move Right, Move Back, Move Left}이며 이 중 한 개를 선택하여 수행한다. 정지를 제외한 나머지 네 가지 Action 은 현재 에이전트가 바라보는 방향을 기준으로 한다. 이 바라보는 방향은 이전 타임 스텝의 Action 에 의해 결정되며 Stop 의 경우는 변경되지 않는다. 예를 들어, 에이전트가 직전 타임 스텝에 의해 바라보는 방향이 절대적 방향으로 오른쪽인 상태에서 Move Right Action 을 수행할 경우 움직이는 방향은 절대적 방향으로 아래쪽이 된다. 또, 에이전트는 자신의 바라보는 방향을 기준으로 회전된 State 를 사용한다. 즉, 에이전트는 바라보는 방향을 기준으로 State 를 인식하고 정책을 따르는 Action 을 선택한 후 바라보는 방향을 기준으로 Action 을 수행한다.

4. Reward

각 에이전트는 목적지에 도달하면 +1.0 Reward 를 받는다. 반대로 Action 이 장애물 또는 다른 에이전트의 위치나 Action 등으로 인해 충돌이 발생하는 경우 Action 이 Stop 으로 변경되고 -0.3 Reward 를 받는다. 그 외의 Reward 는 모두 0 이다.

이러한 설계를 통해 각 에이전트는 자신에게 임무가 할당되면 최단 타임 스텝 내에 목적지에 도달하는 것이 가장 높은 기대 Reward 를 갖는다. 따라서, 모든 에이전트들은 자신의 목적지에 빠르게 도달하기 위해서 장애물을 피해 가장 빠른 경로로 이동해야 한다. 하지만, 다른 에이전트와의 충돌을 고려하지 않는다면 목적지에 도달하는 타임 스텝이 더 길어질 수 있다. 즉, 타임 스텝마다 장애물과 다른 에이전트들을 고려하여 Action 을 결정해야 한다. 공동 정책은 모든 에이전트들의 정책 출력 값과 전역 State 를 종합하여 모든 에이전트들의 기대 Reward 합을 최대화하는 방향으로 업데이트 된다. 그 과정에서 에이전트들의 정책에도 영향을 주어 에이전트가 공동의 목표를 위해 행동하도록 개입한다. 즉, 공동 정책은 환경의 전체 보상을 안정적으로 높일 수 있도록 개별 정책을 제어한다.

Ⅲ. 결론

본 연구의 실험에 사용된 레이아웃은 그림 1 과 같다. QMIX 알고리즘을 사용하였고 총 12 개의 에이전트가 존재한다. 에이전트는 녹색, 선반은 파란색, 피킹 스테이션의 입구는 빨간색, 피킹 스테이션의 출구는 황색으로 표현된다.

그림 2 는 학습 그래프이며 x 축, y 축은 각각 episode, episode 의 총 Reward 를 나타낸다. 약 4000 episodes 동안 학습이 진행되었으며 약 42 의 Reward 에 수렴하여 다수의 무인운반기를 성공적으로 제어할 수 있음을 확인할 수 있다.

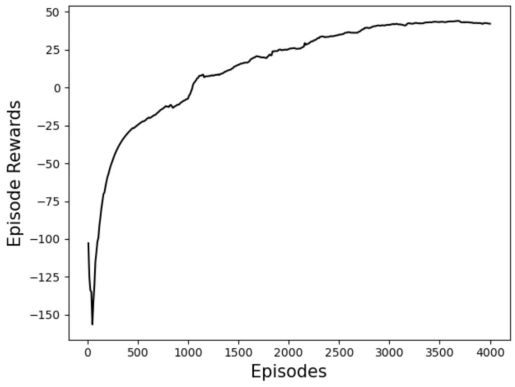

그림 2. 학습 그래프

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2020R1I1A3065610).

참 고 문 헌

[1] Han, Zengliang, et al. "Multi-AGV path planning with double-path constraints by using an improved genetic algorithm." PloS one, 2017.

[2] Yan, Xuejun, Canrong Zhang, and Mingyao Qi. "Multi-AGVs collision-avoidance and deadlock-control for item-to-human automated warehouse." 2017 International Conference on Industrial Engineering, Management Science and Application (ICIMSA). IEEE, 2017.

[3] Rashid, Tabish, et al. "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning." International Conference on Machine Learning. PMLR, 2018.