KIPS Trans. Comp. and Comm. Sys.
Vol.11, No.9 pp.281~288
ISSN: 2287-5891 (Print), ISSN 2734-049X (Online)

https://doi.org/10.3745/KTCCS.2022.11.9.281

Collision Avoidance Path Control of Multi-AGV Using Multi-Agent Reinforcement Learning

Ho-Bin Choi† · Ju-Bong Kim† · Youn-Hee Han†† · Se-Won Oh††† · Kwi-Hoon Kim††††

ABSTRACT

AGVs are often used in industrial applications to transport heavy materials around a large industrial building, such as factories or warehouses. In particular, in fulfillment centers their usefulness is maximized for automation. To increase productivity in warehouses such as fulfillment centers, sophisticated path planning of AGVs is required. We propose a scheme that can be applied to QMIX, a popular cooperative MARL algorithm. The performance was measured with three metrics in several fulfillment center layouts, and the results are presented through comparison with the performance of the existing QMIX. Additionally, we visualize the transport paths of trained AGVs for a visible analysis of the behavior patterns of the AGVs as heat maps.

Keywords : Fulfillment Center, Warehouse, AGV, Path Control, MARL

다중 에이전트 강화학습을 이용한 다중 AGV의 충돌 회피 경로 제어

최 호 빈† · 김 주 봉† · 한 연 희†† · 오 세 원††† · 김 귀 훈††††

요 약

산업 응용 분야에서 AGV는 공장이나 창고와 같은 대규모 산업 시설의 무거운 자재를 운송하기 위해 자주 사용된다. 특히, 주문처리 센터에서는 자동화가 가능하여 유용성이 극대화된다. 이러한 주문처리 센터와 같은 창고에서 생산성을 높이기 위해서는 AGV들의 정교한 운반 경로 제어가 요구된다. 본 논문에서는 대중적인 협력 MARL 알고리즘인 QMIX에 적용될 수 있는 구조를 제안한다. 성능은 두 종류의 주문처리 센터 레이아웃에서 세 가지의 메트릭으로 측정하였으며, 결과는 기존 QMIX의 성능과 비교하여 제시된다. 추가적으로, AGV들의 행동 패턴에 대한 가시적인 분석을 위해 훈련된 AGV들의 운반 경로를 시각화한 히트맵을 제공한다.

키워드 : 주문처리 센터, 창고, 무인운반차, 경로 제어, 다중 에이전트 강화학습

1. 서 론

주문처리 서비스를 제공하는 대표적인 기업인 Amazon은 fulfillment center (주문처리 센터)를 통해 고객의 요구에 빠른 응답을 선보이고 있다. 온라인 유통 산업에서 주문처리 서비스는 주문에 따라 창고로부터 고객에게 제품을 피킹, 포장 및 배송하는 일련의 프로세스이다. 이러한 프로세스는 고객의 요구를 신속하게 충족시킬 뿐만 아니라 배송 대행, 재고관리, 보안, 화재보험 등 기업에 많은 이점을 제공한다.

주문처리 센터 내에서는 수많은 재고를 동시다발적으로 운반해야 하므로 체계적인 재고관리가 필요하다. 피킹, 포장, 배송과 같은 작업은 오직 사람만 수행할 수 있다. 그러나 무인운반차(Automated Guided Vehicle, AGV)는 무거운 재고를 쉽게 운반할 수 있으므로 단순한 재고 운반은 AGV가 맡는 것이 효율적이다. 따라서 AGV는 주문처리 센터의 자동화를 위한 필수 구성 요소이며, 이들의 조직적인 제어는 효율적인 재고관리로 이어지며 창고의 생산성을 높인다.

다중 AGV 창고를 포함한 대부분의 현실 문제들은 단일 개체가 아닌 여러 개체가 서로 협력하거나 경쟁할 때 발생한다 [1]. 다중 에이전트 강화학습(Multi-Agent Reinforcement Learning, MARL)은 [2-4]와 같이 다양한 분야에서 좋은 성과를 얻었으며 크게 3가지의 프레임워크로 분류된다. 1) 완전 중앙집중 학습(Fully centralized learning)은 단일 에이전트 강화학습에서 일반적으로 사용되는 프레임워크이지만, 에이전트의 수가 증가함에 따라 액션 공간이 기하급수적으로 증가하는 치명적인 문제가 있다. 2) 완전 분산 학습(fully decentralized learning)은 액션 공간이 기하급수적으로 늘어나는 단점은 없지만, 이 프레임워크에서는 비정상성 문제가 에이전

※ 이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2020R1I1A3065610). 또한, 이 논문은 2020년도 한국기술교육대학교 교수 교육연구진흥과제 지원에 의하여 연구되었음.
† 준 회 원 : 한국기술교육대학교 컴퓨터공학과 미래융합공학전공 박사과정
†† 종신회원 : 한국기술교육대학교 컴퓨터공학과 미래융합공학전공 교수
††† 비 회 원 : 한국전자통신연구원 책임연구원
†††† 비 회 원 : 한국교원대학교 인공지능융합교육전공 교수
Manuscript Received : April 15, 2022
Accepted : April 22, 2022
* Corresponding Author : Youn-Hee Han(yhhan@koreatech.ac.kr)

트 간의 통신 부족으로 인해 더욱 악화된다. 3) 중앙집중 학습 및 분산 실행(centralized training with decentralized execution, CTDE)은 위의 두 프레임워크가 가진 단점들에 강건하며 일반적인 분산화 시스템에 적합하다[5].

본 논문에서는 AGV 창고 내에서 자율적으로 이동하는 수많은 개별 AGV의 경로를 체계적으로 제어하여 창고의 생산성을 높이기 위해 CTDE 프레임워크 기반 MARL을 채택하였다. 이를 위해 AGV 창고에서 AGV의 경로 제어 작업을 분산 부분 관찰 가능한 마르코프 결정 프로세스(decentralized partially-observable Markov decision process, Dec-POMDP)로 형식적으로 정의한다[6]. 또한 시스템을 구성하는 모듈에 대한 설명 및 전체적인 시나리오와 함께 상태(state) 및 관찰(observation) 표현, 행동(action) 표현, 보상(reward) 함수를 제시한다.

AGV 창고에서 여러 AGV가 동시에 이동하기 위해서는 세밀한 경로 제어가 필요하다. 또한 각 AGV의 경로는 서로 간의 충돌이 엄격하게 고려되어야 한다. AGV의 모든 경로가 탐욕스럽게 제어된다면 충돌 가능성은 기하급수적으로 증가하고 시스템은 치명적인 성능 저하를 겪을 것이다. 그렇지만, 모든 AGV는 충돌을 너무 두려워하지 않아야 하며 혼잡과 병목 현상을 일으키지 않고 가능한 한 짧은 운반 경로로 이동해야 한다.

QMIX는 MARL 분야에서 널리 사용되는 알고리즘으로 협력 문제 해결에 적합한 구조로 되어 있다. QMIX를 통해 각 AGV를 개별 에이전트로 설정할 수 있으며 협업을 고려한 운반 경로를 실시간으로 제어할 수 있다. 그러나 QMIX는 단순히 팀의 보상을 극대화하는 것을 목표로 하기 때문에 개별 에이전트의 희생이 무시될 수 있다. 따라서 본 논문에서는 개별 에이전트에 대한 피드백을 향상시키기 위해 추가적인 기술을 적용하여 QMIX의 한계를 개선하는 구조를 제안한다.

실험을 위해 두 종류의 주문처리 센터 레이아웃을 생성하였고 세 가지 메트릭을 기반으로 결과를 제시한다. 제안하는 구조의 조직적 성능 우수성을 입증하기 위해 기존의 QMIX를 사용한 결과와 비교한다. 동시에, 주어진 동일한 주문처리 센터 레이아웃에서 AGV의 수만 변경하여 실험함으로써 조직적 성능 차이를 더욱 강조한다. AGV는 서로 충돌할 수 있다는 치명적인 단점이 있다. 따라서 일부 AGV로 인해 좁은 지역에서 병목 현상이 발생하거나 정체될 수 있으므로 각 AGV는 자신의 이동 경로를 신중하게 결정해야 한다. 이 분석을 위해 훈련된 AGV들의 움직임을 시각화한 히트맵들을 제시한다.

본 논문의 주요 기여들은 다음과 같이 요약된다:

1) 협력 MARL로 문제를 해결하기 위해 AGV 창고를 Dec-POMDP로 형식적으로 정의한다.
2) 순차적 액션 마스킹을 도입함으로써 가능한 모든 충돌 경우를 제거한다.
3) 구체적인 개별 피드백을 통해 모든 AGV가 체계적으로 행동하는 것을 보장한다.

본 논문은 다음과 같이 구성된다. 2장에서는 강화학습을 AGV에 적용한 관련 연구들을 소개하고, QMIX의 기본 아이디어를 기술한다. 3장에서는 AGV 창고 시스템의 모델 설계 내용을 기술하며, 4장에서는 제안하는 QMIX 기반 다중 에이전트 강화학습 구조를 활용한 AGV 충돌 회피 경로 제어 기법에 관해 기술한다. 5장에서는 실험 결과들에 대해 논의하며, 6장에서는 결론과 함께 본 논문을 마무리한다.

2. 관련 연구

2.1 무인운반차

Amazon Robotics는 불확실한 상황에 놓인 AGV의 의사결정, 경로 계획, 스케줄링을 비롯한 리소스 할당 문제를 다루며 AGV의 조직적인 자율성과 분산 의사 결정을 위한 다중 AGV 창고 시나리오를 제공한다[7]. 다중 AGV의 경로 계획 문제는 다양하게 해석될 수 있으며, 이를 해결하기 위해 많은 알고리즘이 제안되었다[8-10]. 최근에는 단일 에이전트 강화학습을 적용하여 실시간으로 다중 AGV의 경로를 제어하는 연구가 시도되어 기존 알고리즘을 능가하는 좋은 결과를 얻었다[11,12]. 본 논문에서는 기존의 방식들과는 다르게, CTDE 프레임워크를 기반한 다중 에이전트 강화학습을 사용하여 분산 의사 결정을 간단하게 실현하며 효과적인 자율성을 달성한다.

2.2 다중 에이전트 심층 강화학습

단일 에이전트 강화학습에서 샘플 효율이 높은 대표적인 값 기반 알고리즘은 Q-learning 기반 DQN이다[13,14]. DQN은 다중 에이전트 강화학습의 적용을 위해 IQL로 간단히 확장될 수 있다[15]. 그러나 IQL은 에이전트 간의 통신 부재로 인해 비정상성 문제가 악화되는 것을 방지할 수 없다. VDN은 단순히 모든 에이전트의 액션 가치를 합산하여 전체 팀의 합동 액션 가치를 계산한다[16]. 이 가산성 제약은 QMIX에서 단조성 제약으로 완화되었다[17]. QMIX는 VDN과 같은 선형 방식이 아닌 하이퍼 네트워크를 활용한 정교한 비선형 방식으로 팀 전체의 합동 액션 가치를 추정한다. 하이퍼 네트워크는 상태 정보를 추가적으로 사용하여 합동 액션 가치 함수의 더 풍부한 표현 능력을 끌어낸다. 혼합 네트워크의 단조성 제약은 하이퍼 네트워크에 의해 생성된 가중치를 음이 아닌 수로 제한함으로써 보장된다.

3. 시스템 모델

본 논문에서는 에이전트 간의 협력이 필요한 작업을 해결하기 위한 다중 에이전트 환경을 고려한다. 이 작업은 튜플 $< A, S, O, Z, U, P, R_g, R_l, \gamma >$로 표현되는 Dec-POMDP에 의해 형식적으로 정의될 수 있다[6]. 각 타임 스텝 t마다 환경은 글로벌 상태 $s_t \in S$와 관찰 $o_t^a = O(s_t, a) \in Z$를 출력한다. 여기서, S는 글로벌 상태 공간을 의미하며 Z는 각 에이전트 $a \in \{1, ..., \mathrm{n}\} \equiv \mathrm{A}$에 대한 관찰 공간을 나타내고 $O: S \times A \rightarrow Z$는 관찰 함수이다. 각 에이전트는 자신의 관찰을 사용하여 액션 $u_t^a \in U$를 선택한다. 여기서, U는 각 에이전트 $a \in \{1, ..., \mathrm{n}\} \equiv \mathrm{A}$에 대한 액션 공간을 의미한다. 환경은

모든 에이전트에 대한 합동 액션 $\boldsymbol{u}_t \in \boldsymbol{U} \equiv U^n$를 수행하고 상태 전이 함수 $P(s_{t+1}|s_t, \boldsymbol{u}_t) : S \times \boldsymbol{U} \times S \rightarrow [0,1]$에 따른 전이를 초래한다. 이 전이는 글로벌 보상 $r_t = R_g(s_t, \boldsymbol{u}_t)$과 개별 보상 $r_t^a = R_l(s_t, u_t^a)$이 포함되어 완성된다.

이 프로세스에서 각 에이전트는 글로벌 보상으로 합산되는 자신의 개별 보상을 높이려 노력하며 궁극적으로는 감가된 누적 글로벌 보상을 최대화하는 것을 목표로 한다. 각 에이전트는 확률적 정책 $\pi^a(u^a|\tau^a) : T \times U \rightarrow [0,1]$를 조건으로 하여 자신의 관찰-액션 히스토리 $\tau^a \in T \equiv (Z \times U)^*$를 생성한다. 여기서, $(Z \times U)^*$는 가능한 모든 관찰-액션 히스토리들의 집합을 나타낸다. 각 에이전트의 정책 π^a로 구성되는 합동 정책 π는 합동 액션 가치 함수를 가지며 다음과 같이 공식화된다:

$$Q^{\pi}(s_t, \boldsymbol{u}_t) = \mathbb{E}_{s_{t+1:\infty}, \boldsymbol{u}_{t+1:\infty}}[\sum_{i=0}^{\infty} \gamma^i r_{t+i} | s_t, \boldsymbol{u}_t]. \quad (1)$$

3.1 문제 정의

본 절에서는 주문처리 센터에서 요구하는 AGV 창고를 모델링한 강화학습 환경을 제시한다. Fig. 1은 설계한 시스템을 시각적으로 기술한다. 시스템의 전체 프로세스는 다음과 같이 요약된다: 1) 시스템이 작동을 시작하면 모든 AGV가 충전소에서 할당을 대기한다; 2) 각 피킹 스테이션의 작업자는 필요한 선반을 AGV에 할당한다; 3) 할당된 AGV가 요청된 선반으로 이동하여 들어 올린다; 4) AGV는 선반을 피킹 스테이션의 피킹 스테이션 입구(PSE)로 운반한다; 5) 피킹 스테이션에서 피킹 작업이 완료되면 선반을 사용 가능한 선반 보관소로 운반하여 내려놓는다. 그 후, AGV는 2)부터 프로세스를 반복한다.

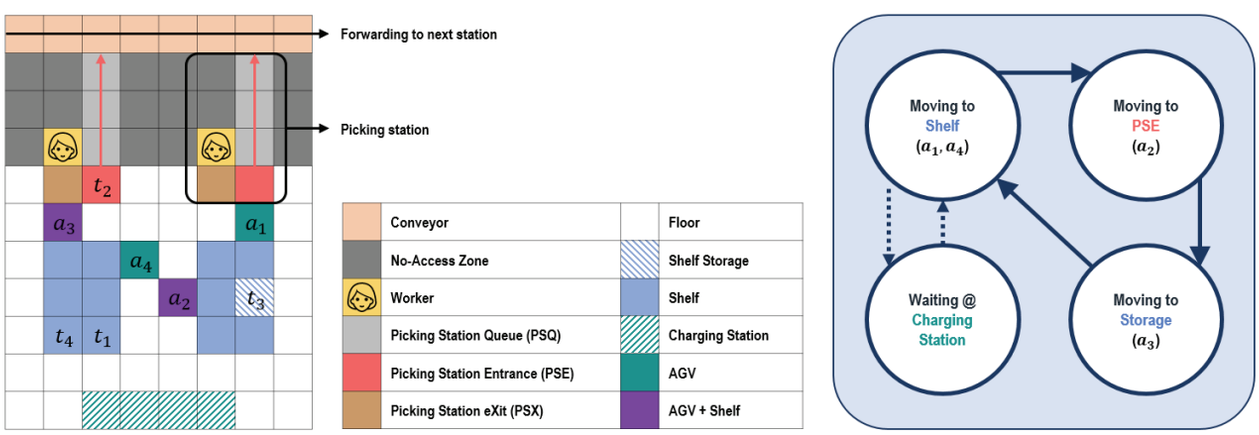
본 연구에서는 전체 프로세스를 모두 다루지 않고 AGV의 운반 경로 제어에만 중점을 둔다. AGV 운반 경로 제어 이외의 작업은 다음과 같이 가정한다. 모든 피킹 스테이션의 작업자는 지속적으로 AGV에게 선반을 가져오도록 요청한다. 각 요청은 이용 가능한 임의의 AGV에 할당된다. 마찬가지로, 각 AGV는 피킹 작업이 완료된 선반을 임의의 이용 가능한 선반 저장소로 반환한다. 전체 과정에서 AGV의 운반 경로 제어와 관련된 작업은 Fig. 1(c)와 같은 상태 전이 주기로 구성된다. 각 AGV는 서로 다른 상태에 있을 수 있지만, 한 AGV는 동시에 여러 상태에 있을 수 없다. 실제 AGV 창고에서는 AGV, 선반 및 선반 보관소가 무작위로 선택되지 않는다.

3.2 상태 및 관찰 표현

분산 의사 결정을 준수하기 위해 각 AGV를 독립적인 에이전트로 취급한다. 각 타임 스텝마다 각 에이전트는 환경에서 생성된 자신의 관찰을 수신한다. 각 에이전트의 관찰은 자신을 중심으로 주변 9×9 영역에 대한 2채널 정보로 구성되어 $2 \times 9 \times 9$의 크기를 형성한다. 첫 번째 채널은 에이전트가 해당 위치로 이동할 수 있는 경우 1.0, 그렇지 않으면 0.0으로 구성된다. 두 번째 채널은 각 위치에서 에이전트의 타겟 위치까지의 남은 맨해튼 거리에 대해 0.0에서 1.0의 범위로 정규화된 값으로 구성된다. 관찰과 동시에 환경은 주문처리 센터 레이아웃의 전체 영역에 대한 정보인 상태도 생성한다. 유일한 첫 번째 채널은 모든 에이전트에 대해 0.0에서 1.0의 범위로 정규화된 각 위치의 경로 사용량으로 구성된다. 상태의 크기는 주문처리 센터 레이아웃의 그리드 크기에 따라 결정된다. 상태에 비해 관찰의 크기가 너무 작거나 너무 크면 에이전트가 주변 상황을 인식하는 능력이 손상될 수 있다.

3.3 액션 표현

각 에이전트는 환경에서 제공하는 자신의 관찰을 사용하여 해당 위치에서 하나의 그리드 셀을 이동할 수 있는 액션을 선택한다. 선택된 액션은 에이전트가 바라보는 방향에 따라 수행되며, 모든 에이전트의 액션 공간은 {stop, move forward, move right, move left, move backward}와 같이 정의된다. 에피소드 시작 시 각 에이전트의 초기 시선 방향은 임의로 설정되며, 그 후에는 선택하는 액션에 따라 결정된다. 이 메커니즘을 위해 관찰은 에이전트의 바라보는 방향을 기준으로 회전된다.



(a) Snapshot (b) Color-based module representation (c) AGV state transition diagram

Fig. 1. Description of the Proposed Multi-agent Reinforcement Learning Environment Modeling AGV Warehouse, Where the Target Position of Agent a_i is t_i for Each Agent $i \in \{1, ..., 4\}$.

3.4 리워드 함수

환경은 각 에이전트가 선택한 액션에 따라 타임 스텝마다 개별 보상을 반환한다. 개별 보상은 에이전트가 선택한 액션을 수행한 후 에이전트와 타겟 위치 사이의 남은 맨해튼 거리가 감소하는 경우 $+1/n$로 결정되며, 남은 맨해튼 거리가 유지되거나 증가하는 경우 $-1/n$로 결정된다. 여기서 n은 에이전트 수이다. 에이전트가 정지 액션을 선택했을 때도 음의 보상을 받는 것은 에이전트에게 가혹한 패널티 일 수 있다. 하지만, 에이전트들이 이 보상 시스템을 통해 더 효율적인 운반 경로로 이동하기를 기대한다. 글로벌 보상은 모든 개별 보상의 합계이며 범위는 -1.0에서 $+1.0$이다. 이러한 보상 설계는 에이전트의 높은 일반화 성능을 끌어낼 수 있다는 점에서 유리하다.

4. 제안하는 구조

4.1 순차적 액션 마스킹

AGV는 벽에 부딪히거나 서로 충돌할 수 있다는 큰 단점이 있다. 이러한 단점으로 인해 생성된 경험은 학습에 유용하지 않으며 훈련을 방해할 수도 있다. 따라서 본 연구에서는 양질의 경험만 리플레이 버퍼에 저장되도록 액션 마스킹을 도입한다[18-20]. 기존의 액션 마스킹에서 환경은 각 에이전트의 불가한 액션을 감지하고 타임 스텝마다 이를 각 에이전트에게 동시에 알려준다. 그러나 이 방식으로는 AGV 창고에서 여전히 불가한 액션이 선택될 수 있다.

모든 AGV가 유효한 액션을 선택하더라도 다른 AGV의 액션으로 인해 각 AGV의 유효한 액션이 불가한 액션으로 변경될 수 있다. 이러한 딜레마를 해결하기 위해 본 논문에서는 액션 마스킹을 동시에 적용하지 않고 순차적인 액션 마스킹 기법을 제안한다. Fig. 2는 순차적 액션 마스킹 프로세스를 보여준다. IQL 및 QMIX와 같은 모든 MARL 알고리즘에 적용할 수 있는 이 기법은 AGV의 가능한 모든 충돌 경우를 방지할 뿐만 아니라 리플레이 버퍼에 효율적인 경험만 저장하여 알고리즘의 성능을 향상시킨다.

4.2 추가 로컬 오차

본 논문에서는 분산 실행 및 중앙집중 훈련 프레임워크를 유지하면서 QMIX [17]에 추가 기술을 적용하는 구조를 제안한다. QMIX는 미니 배치에 대해 합동 액션 가치 Q와 타겟 사이의 제곱 오차를 다음과 같이 사용한다:

$$\mathcal{L}_{global} = \sum_{i=1}^{b}\left[\left(y^{tot} - Q^{tot}(\boldsymbol{\tau}, \boldsymbol{u}, s;\, \theta^{tot})\right)^2\right], \qquad (2)$$

여기서, $y^{tot} = \mathrm{r} + \gamma \max_{\boldsymbol{u}'} Q^{tot}(\boldsymbol{\tau}', \boldsymbol{u}', s'; \bar{\theta}^{tot})$이다. 합동 액션 가치 Q는 각 에이전트의 액션 가치 Q^a를 혼합 네트워크에 입력하여 생성된다. QMIX는 합동 액션 가치 Q와 액션 가치 Q^a 간의 관계를 단조성 제약으로 제한한다. 이러한 제약을 적용하기 위해 혼합 네트워크의 가중치는 음이 아닌 수로 제약된다.

혼합 네트워크의 오차인 글로벌 오차 L_{global}은 Equation (2)와 같이 계산된 후, 혼합 네트워크와 에이전트 네트워크로 구성된 피드포워드 네트워크로 역전파된다. 이 과정에서 그래디언트는 에이전트 네트워크에 도달하지만, 그 효과는 상당히 미미할 수 있다. 구체적으로, 단조성 제약이 보장될지라도 혼합 네트워크가 단독으로 개별 에이전트들의 기여도를 구별하는 것은 어렵다. 따라서, 에이전트 네트워크를 위해 다음과 같이 공식화되는 추가 로컬 손실을 제안한다:

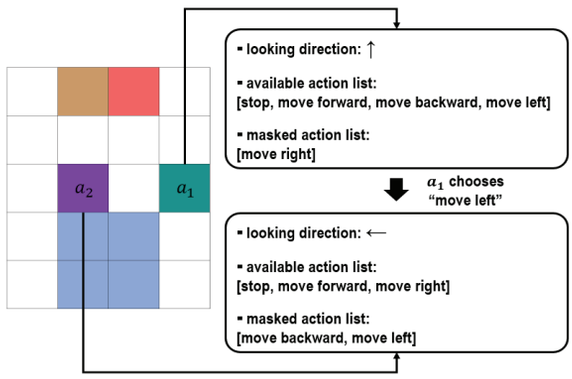


Fig. 2. Example of Sequential Action Masking

$$\mathcal{L}^a_{local} = \sum_{i=1}^{b}\left[\left(y^a - Q^a(\tau^a, u^a; \theta^a)\right)^2\right], \qquad (3)$$

여기서, $y^a = r^a + \gamma \max_{u'_a} Q^a(\tau'_a, u'_a; \bar{\theta}^a)$이다. 글로벌 오차 L_{global}은 혼합 네트워크의 Q를 사용하여 계산되는 반면, 로컬 오차 L^a_{local}는 각 에이전트에 대해 에이전트 네트워크의 Q^a를 사용하여 계산된다.

본 연구에서는 에이전트 네트워크가 공유되지만, 필수적 요소는 아니다. 각 에이전트의 관찰-액션 히스토리 τ^a는 공유되지 않으며 로컬 오차 L^a_{local}는 각각 별도로 계산된다. 로컬 오차 L^a_{local}에 의해 계산된 그래디언트는 혼합 네트워크와 관계없이 오직 에이전트 네트워크를 통해 역전파된다. 글로벌 오차 L_{global}을 통한 역전파도 에이전트 네트워크에 어느 정도 영향을 주지만 로컬 오차 L^a_{local}를 통한 역전파는 개별 에이전트의 액션에 대한 피드백을 명확하게 할 수 있다. 최종 오차는 다음과 같이 정의된다:

$$\mathcal{L} = \mathcal{L}_{global} + \sum_{a=1}^{n} \mathcal{L}^a_{local}. \qquad (4)$$

Fig. 3은 제안하는 구조가 적용된 QMIX의 전체 네트워크 아키텍처를 간략하게 보여준다. 에이전트 네트워크는 GRU [21]와 MLP로 구성되는 반면, 혼합 네트워크는 MLP로만 구성되지만, 가중치와 편향은 단조성 제약을 보장하기 위해 하이퍼 네트워크[22]의 출력으로부터 얻어진다. 상태와 관찰은 먼저 CNN을 통해 1차원 특징으로 변환된 다음 MLP에 입력된다.

각 에이전트는 자신의 관찰만을 활용하여 분산 실행을 수행하는 환경과의 상호작용을 통해 경험을 리플레이 버퍼에

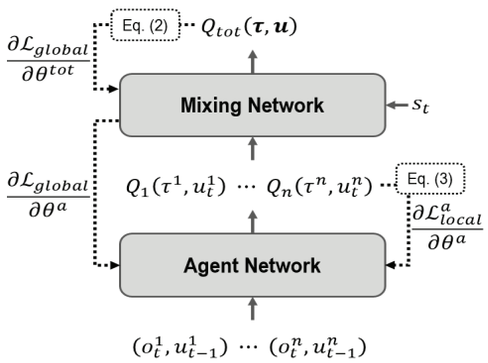


Fig. 3. Overall Network Architecture

저장한다. 혼합 네트워크는 모든 에이전트의 액션 가치 Q^a를 합동 액션 가치 Q로 추정하여 에이전트 네트워크를 업데이트하는 데 사용한다. 즉, 혼합 네트워크는 환경과 직접 상호작용하지 않고 리플레이 버퍼에서 샘플링된 각 에이전트의 상태 및 액션 가치 Q^a만을 필요로 한다. 중앙집중 훈련의 주요 기반을 형성하는 혼합 네트워크는 훈련이 완료된 후 테스트에는 사용되지 않는다.

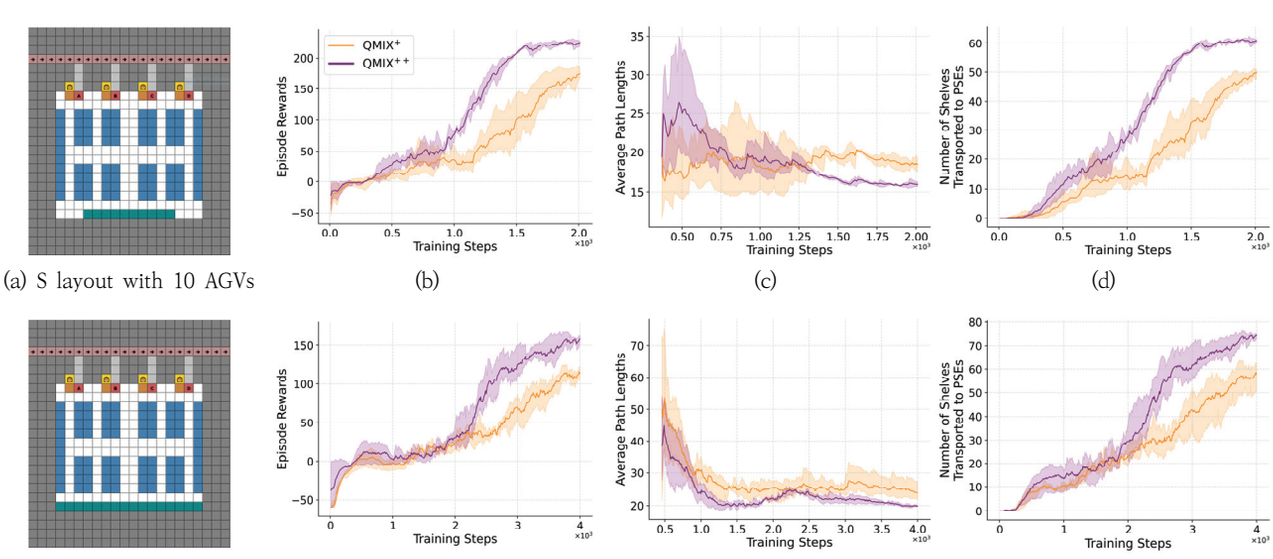
5. 실 험

본 장에서는 제안하는 기법의 우수성을 입증하는 실험을 수행하고 그 성능을 분석한다. 제안하는 기법이 QMIX를 기반으로 하므로 기존 QMIX의 성능을 베이스라인으로써 함께 제시한다. 하지만, 순차적 액션 마스킹 기법이 적용되지 않았을 때 AGV 간의 충돌로 인해 학습이 정상적으로 진행되지 않아, 기존 QMIX에 순차적 액션 마스킹 기법을 적용하였다. 순차적 액션 마스킹 기법은 AGV 간의 가능한 모든 충돌 경우를 해결하며 추가 로컬 오차 기법은 충돌이 발생하지 않는다는 보장 하에 개별 에이전트에 대한 명확한 피드백을 통해 전체 시스템의 성능을 크게 향상시킬 수 있다. QMIX$^+$는 기존 QMIX에 순차적 액션 마스킹 기법이 적용된 것을 의미하고, QMIX^{++}는 순차적 액션 마스킹 기법과 추가 로컬 오차 기법이 모두 적용된 것을 나타낸다.

자세한 평가를 위해 [7-12]를 참고하여 에피소드마다 측정되는 세 가지 메트릭을 채택하였다. 첫째, *episode rewards*는 받은 글로벌 보상의 총 합계이다. 둘째, *average path lengths*는 각 운반을 완료하는 데 걸린 타임 스텝의 평균이다. 셋째, *number of shelves arrived at PSEs*는 PSE에 도착한 선반의 수로 AGV가 선반 보관소에서 피킹 스테이션으로 선반을 운반한 횟수이다. 300 타임 스텝에 도달하면 에피소드가 종료되고 새로운 에피소드가 시작된다. 훈련 중 열 번의 모델 업데이트마다 각 메트릭은 모든 에이전트가 탐욕스러운 액션을 선택하는 다섯 번의 개별 테스트 에피소드에서의 평균값으로 측정된다. 또한, 각 메트릭은 5회 실행에 대한 95% 신뢰 구간으로 표시하였다.

Fig. 4와 5는 각각 S 레이아웃과 L 레이아웃에서 QMIX$^+$와 QMIX^{++}를 사용하여 훈련한 결과를 보여준다. 동일한 주문처리 센터 레이아웃에서 AGV의 수가 성능에 미치는 영향을 알아보기 위해 각 주문처리 센터 레이아웃에 대해 두 종류의 AGV 수로 실험하였다. 다른 모든 조건이 동일하더라도 AGV의 수를 조정하면 필요한 협력 역량이 변경된다. 즉, 한정된 공간에서 AGV의 수가 증가할수록 더 치밀한 운반 경로 제어가 요구된다.

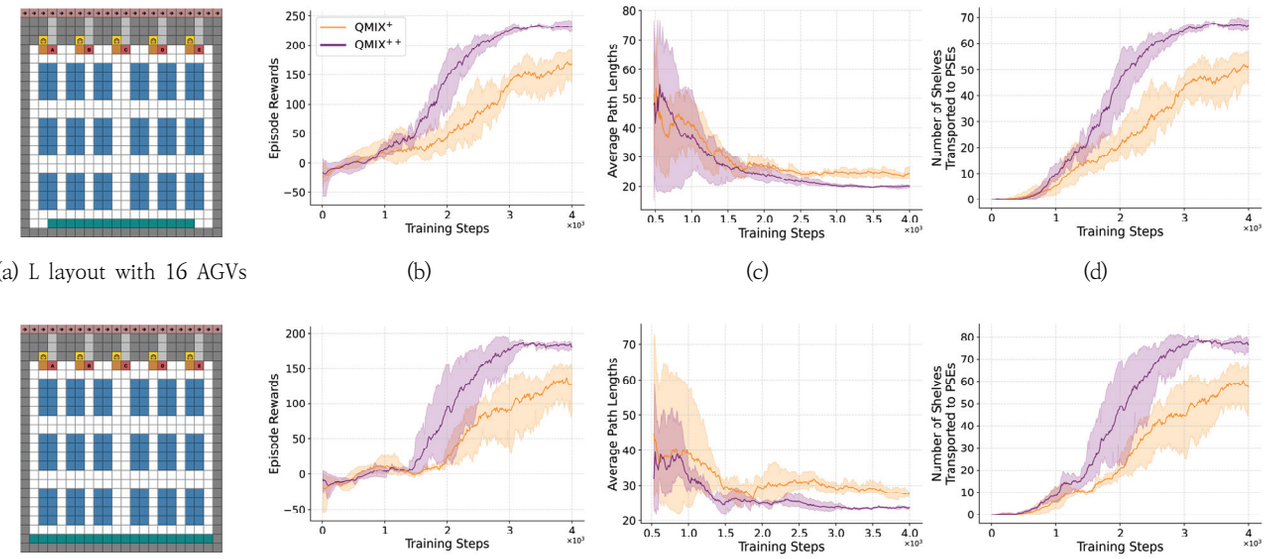
본 실험에서는 순차적 액션 마스킹을 적용하여 모든 충돌 경우를 제거했지만, 여전히 AGV는 서로를 통과할 수 없다. 또한, 주문처리 센터 레이아웃의 공간은 제한되어 있으며, 이 제약을 Fig. 4(a) 및 5(a)와 같이 선반과 선반 사이 또는 선반과 접근 금지 구역 사이의 공간을 의도적으로 일방통행로로 만듦으로써 부각하였다. 궁극적으로 작업을 효율적으로 수행



(a) S layout with 10 AGVs (b) (c) (d)

(e) S layout with 16 AGVs (f) (g) (h)

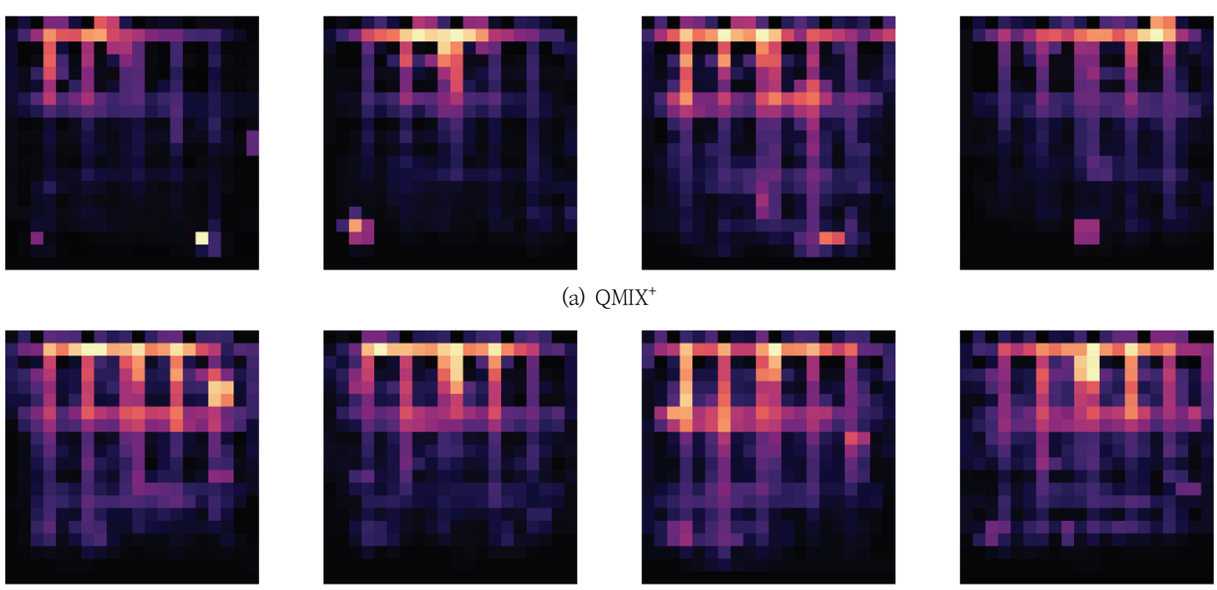
Fig. 4. Training results on S layout.



(a) L layout with 16 AGVs (b) (c) (d)

(e) L layout with 20 AGVs (f) (g) (h)

Fig. 5. Training Results on L Layout



(a) QMIX+

(b) QMIX++

Fig. 6. Heat Maps on L Layout

하기 위해서는 특정 시간 동안 어떤 AGV가 일방통행로를 사용할 것인지 통신해야 한다. 모든 AGV가 자신의 이익만을 추구한다면 이 핵심 과제를 효율적으로 해결할 수 없다. 일부 AGV는 양보하거나 차선책을 사용함으로써 손해를 볼 수 있지만, 이러한 희생을 통해 전체 시스템의 최종 성능을 높일 수 있다. 이 협업은 AGV가 극복해야 하는 숨겨진 과제이며 실시간 운반 경로 제어를 적용할 당위성을 설명한다.

QMIX+는 에이전트 간의 통신을 기반으로 서로를 인식하고 협업을 시도하지만, 개별 에이전트에 대한 정확한 운반 경로 분배가 이루어지지 않아 훈련이 불안정하고 상대적으로 좋지 못한 성능을 보인다. 대조적으로, QMIX++는 추가 로컬 오차 기법이 적용되어 각 에이전트에 대한 명확한 개별 피드백으로 인해 훈련이 안정적이고 상대적으로 우월한 성능을 보인다. 일반적으로, 주문처리 센터의 규모가 커질수록 AGV의 수도 늘어난다. S 레이아웃과 L 레이아웃은 규모만 다를 뿐 전체적인 구조는 비슷하지만, AGV의 수가 다르다. 즉, 이 두 주문처리 센터 레이아웃에서의 운반 경로 제어 복잡성은 AGV의 수로 인해 크게 차이 난다. 예상대로 QMIX++는 레이아웃의 규모에 따른 적응성 측면에서 QMIX+보다 더 높은 안정성과 성능을 보여준다.

Fig. 6은 L 레이아웃에서 AGV들의 운반 경로가 얼마나 고르게 분포되어 있는지를 히트 맵으로 보여준다. 피킹 스테

이션은 레이아웃의 북쪽에 위치하기 때문에 AGV들 대부분의 운반 경로는 레이아웃의 북쪽에 있을 것으로 예상된다. QMIX^{+}는 에이전트 간의 통신을 지원하므로 대체로 레이아웃의 북쪽이 밝지만, 유휴 상태인 여러 AGV가 구석에서 발견된다. 대조적으로, QMIX^{++}의 히트 맵은 AGV의 보다 유연하고 고르게 분포된 운반 경로 사용량을 보여준다. 특히 QMIX^{+}로 훈련했을 때 자주 발생하는 유휴 AGV가 추가 로컬 오차 기법을 활용한 각 AGV에 대한 명확한 개별 피드백으로 인해 발생하지 않는다.

6. 논 의

3장에서 언급한 것처럼, 각 AGV는 자신의 바라보는 방향을 기준으로 액션을 수행하며, 그것에 맞게 자신의 관찰도 회전된다. 또한, 관찰은 주문처리 센터 레이아웃의 전체 영역이 아닌 자신의 주변 9×9 영역의 정보로 이루어져 있다. 이를 통해 추론할 수 있는 것은 주문처리 센터의 레이아웃이 변경되더라도 AGV들의 목적인 선반 운반 경로 제어는 동일하기 때문에 AGV들의 행동 양식은 주문처리 센터의 레이아웃에 구애받지 않아야 한다. 이론적으로, 특정 주문처리 센터 레이아웃에 AGV가 경험할 수 있는 모든 종류의 관찰이 존재한다면 해당 주문처리 센터 레이아웃에서 학습된 모델은 다른 주문처리 센터 레이아웃에서 추가적인 학습 없이 동일한 성능을 유지할 수 있다. 하지만, 한 주문처리 센터 레이아웃에 모든 종류의 관찰이 존재하는 것은 불가능하므로 모델의 높은 일반화 성능을 위해서는 서로 다른 구조를 지닌 여러 종류의 주문처리 센터 레이아웃에서의 학습이 필요하다. 가능한 관찰의 경우는 매우 많지만 비슷한 구조에서 요구되는 액션은 유사할 확률이 높을 것으로 예상된다. 즉, AGV가 모든 경우의 관찰을 경험하지 않더라도 다양한 구조의 관찰을 어느 정도 경험한다면 새로운 주문처리 센터 레이아웃에 강건한 성능을 보일 수 있다.

7. 결 론

본 논문에서는 공동의 목표를 달성하기 위해 제한된 공간에서 함께 작업하는 여러 개체의 실시간 제어가 필요한 AGV 창고 문제를 다뤘다. 이 작업을 Dec-POMDP로 형식적으로 정의하고 다중 에이전트 강화학습을 사용하여 해결하였다. 제안된 방식은 개별적으로 조직적인 경로 제어를 가능하게 하여 전체 시스템의 성능을 향상시킨다. 결과는 *episode rewards*, *average path lengths*, *number of shelves arrived at PSEs*의 세 가지 메트릭으로 측정하였다. 또한 운반 경로 사용량 분포를 검증하기 위해 히트 맵을 제시하였다. 제안된 추가 로컬 오차 기법은 각 에이전트에 대한 구체적인 개별 피드백을 통해 알고리즘의 성능을 크게 향상시킨다. 특히 제안된 순차적 액션 마스킹 기법은 가능한 모든 충돌 경우를 제거한다. 향후 연구로 일반화 성능을 향상시키기 위해 여러 에이전트에 걸쳐 여러 레이아웃을 동시에 훈련하는 연합 강화학습을 적용하는 것을 목표로 한다.

References

[1] L. Buşoniu, R. Babuška, and B. Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in Multi-agent Systems and Applications-1*, pp.183-221, 2010.

[2] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Transactions on Wireless Communications*, Vol.19, No.2, pp.729-743, 2019.

[3] X. Li, J. Zhang, J. Bian, Y. Tong, and T. Liu, "A cooperative multi-agent reinforcement learning framework for resource balancing in complex logistics network," In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

[4] X. Li, X. Hu, W. Li, and H. Hu, "A multi-agent reinforcement learning routing protocol for underwater optical sensor networks," In *Proceedings of IEEE International Conference on Communications*, 2019.

[5] F. A. Oliehoek, M. T. J. Spaan, and N. Vlassis, "Optimal and approximate Q-value functions for decentralized POMDPs," *Journal of Artificial Intelligence Research*, Vol.32, pp.289-353, 2008.

[6] F. A. Oliehoek and C. Amato, "A concise introduction to decentralized POMDPs," *SpringerBriefs in Intelligent Systems*, Springer, 2016.

[7] J. J. Enright and P. R. Wurman, "Optimization and coordinated autonomy in mobile fulfillment systems," In *Proceedings of the AAAI Workshop on Automated Action Planning for Autonomous Mobile Robots*, pp.33-38, 2011.

[8] J. Bae and W. Chung, "A heuristic for a heterogeneous automated guided vehicle routing problem," *International Journal of Precision Engineering and Manufacturing*, Vol.18, No.6, pp.795-801, 2017.

[9] Z. Han, D. Wang, F. Liu, and Z. Zhao, "Multi-AGV path planning with double-path constraints by using an improved genetic algorithm," *PloS one*, Vol.12, No.7, 2017.

[10] Y. Lian and W. Xie, "Improved A* multi-AGV path planning algorithm based on grid-shaped network," In *2019 Chinese Control Conference*, 2019.

[11] R. Kamoshida and Y. Kazama, "Acquisition of automated guided vehicle route planning policy using deep reinforcement learning," *IEEE International Conference on Advanced Logistics and Transport (ICALT)*, 2017.

[12] Y. Yang, J. Li, and L. Peng, "Multi-robot path planning based on a deep reinforcement learning DQN algorithm," *CAAI Transactions on Intelligence Technology*, Vol.5, No.3, pp.177-183, 2020.

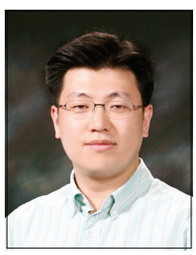
[13] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, Vol.8, pp.279-292, 1992.
[14] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, Vol.518, No.7540, pp.529-533, 2015.
[15] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," In *Proceedings of the Tenth International Conference on Machine Learning*, pp.330-337, 1993.
[16] P. Sunehag et al., "Value-decomposition networks for cooperative multi-agent learning based on team reward," In *Proceedings of 17th International Conference on Autonomous Agents and Multiagent Systems*, Stockholm, Sweden, 2018.
[17] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018.
[18] O. Vinyals et al., "Starcraft II: A new challenge for reinforcement learning," *arXiv preprint arXiv:1708.04782*, 2017.
[19] D. Ye et al., "Mastering complex control in moba games with deep reinforcement learning," In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp.6672-6679, 2020.
[20] S. Huang and S. Ontañón, "A closer look at invalid action masking in policy gradient algorithms," *arXiv preprint arXiv:2006.14171*, 2020.
[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," In *NIPS 2014 Workshop on Deep Learning*, 2014.
[22] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

최 호 빈

https://orcid.org/0000-0002-2071-652X
e-mail : chb3350@koreatech.ac.kr
2019년 한국기술교육대학교 컴퓨터공학부 (학사)
2021년 한국기술교육대학교 컴퓨터공학과 (석사)
2021년 ~ 현 재 한국기술교육대학교 컴퓨터공학과 미래융합공학전공 박사과정
관심분야 : 강화학습, 다중 에이전트 강화학습, 게임 개발

김 주 봉

https://orcid.org/0000-0001-6406-3092
e-mail : rlawnqhd@koreatech.ac.kr
2017년 한국기술교육대학교 컴퓨터공학부 (학사)
2019년 한국기술교육대학교 컴퓨터공학과 (석사)
2019년 ~ 현 재 한국기술교육대학교 컴퓨터공학과 미래융합공학전공 박사과정
관심분야 : 강화학습, 다중 에이전트 강화학습, 딥러닝



한 연 희

https://orcid.org/0000-0002-5835-7972
e-mail : yhhan@koreatech.ac.kr
1998년 고려대학교 컴퓨터학과(석사)
2002년 고려대학교 컴퓨터학과(박사)
2002년 삼성종합기술원 전문연구원
2013년 ~ 2014년 SUNY at Albany, Department of Computer Science 방문교수
2006년 ~ 현 재 한국기술교육대학교 컴퓨터공학과 미래융합공학전공 교수
관심분야 : 사물인터넷, 기계 학습, 강화학습

오 세 원

https://orcid.org/0000-0002-3078-7866
e-mail : sewonoh@etri.re.kr
1999년 포항공과대학교 산업공학과(학사)
2001년 포항공과대학교 산업공학과(석사)
2018년 충남대학교 컴퓨터공학과(박사)
2001년 ~ 현 재 한국전자통신연구원 책임연구원
2022년 ~ 현 재 (주)인투와이즈 책임연구원
관심분야 : SW 구조 설계, IoT/지능화 응용 솔루션 개발

김 귀 훈

https://orcid.org/0000-0002-0798-1687
e-mail : kimkh@knue.ac.kr
1998년 KAIST 전기 및 전자공학과(학사)
2000년 KAIST 전기 및 전자공학과(석사)
2019년 KAIST 전기 및 전자공학과(박사)
2000년 ~ 2005년 LG데이콤 주임연구원
2005년 ~ 2020년 한국전자통신연구원 책임연구원
2006년 ~ 현 재 ITU-T SG11 Rapporteur, Editor
2020년 ~ 현 재 한국교원대학교 인공지능융합교육전공 교수
관심분야 : 인공지능융합교육, 지능형에지컴퓨팅, 강화학습