강화학습 기반 다중 에이전트 간 시맨틱 통신 작업 환경

김진혁, 서세진, 박지훈, 김성륜*
*연세대학교 전기전자공학부

{jh.kim, sjseo, jhpark}@ramo.yonsei.ac.kr, *slkim@yonsei.ac.kr

Semantic Communication Task Environment based on Multi-Agent Reinforcement Learning

Jinhyuk Kim, Sejin Seo, Jihoon Park, Seong-Lyun Kim*
Dept. Electrical & Electronic engineering, Yonsei Univ.

요 약

본 논문은 고성능 저지연 차세대 통신을 위한 시맨틱 통신 구조의 작업 환경을 제안한다. 다양한 무선 네트워크 시나리오에서 저지연, 고성능을 보장하며 공통 작업을 수행하기 위해서는 네트워크의 여러 주체가 협력해야 한다. 최근 많은 연구에서 이용하는 강화학습 기법을 통한 통신 시나리오의 지능 에이전트가 생성하는 시맨틱 통신을 분석하고 분석방법을 소개한다.

Ⅰ. 서 론

앞으로 나타나는 6G 네트워크에는 AI 와 센싱 기능을 갖춘 다양한 종류의 통신 기기들이 등장할 것으로 예상된다[1]. 이러한 통신 기기들이 대량으로 배치됨에 따라, 에너지 제약(energy constraints) 및 저지연(low latency)의 문제가 나타날 것으로 보인다. 시맨틱 통신(Semantic communication)은 기존의 통신과 달리, 시맨틱 수준에서 메시지를 주고받는 통신으로, 같은 작업대비 더 적은 양의 통신양이 필요하다.

본 논문에서는 지능을 가진 여러 에이전트간 의미 전달을 통해 작업(task)의 성능을 보장하며 통신 효율적인 시맨틱 통신 구조를 제안한다. 최근 시맨틱 통신의 분석과 평가지표 등의 발달에 따라, 기존 강화학습 모델기반의 통신이 기계 지각적 도메인에서 이루어지는 시맨틱 통신으로 해석과 분석이 용이해졌다. 이에 따라 본 논문에서 여러 에이전트들이 강화학습을 통해 시맨틱 통신을 학습하고, 분석한다.

Ⅱ. 본론

본 논문에서는 다중 에이전트 파티클 환경(Multi-Agent Particle Environment)의 협력적 커뮤니케이션(Cooperative Communication)[2] 상황에서 다중 에이전트 근접 정책최적화(Multi-Agent Proximal Policy Optimization)[3]기반의 시맨틱 통신 학습 기법을 제안한다.

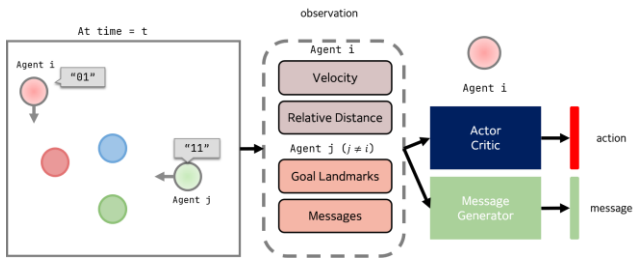

그림 1. 시스템 모델

가. 시스템 모델

부분적으로 관찰 가능한 마르코프 결정 과정(POMDP)을 가정하고, 다중 에이전트간 협력적 커뮤니케이션 상황에서 각 에이전트는 각기 다른 3 개의 랜드마크들(Landmarks) 중 한 곳으로 할당되어 이동하는 것을 목표로 한다. 기존 연구와 다르게, 더욱 부분적으로 관찰 가능하게 바꾸기 위해서, 각 에이전트는 오로지 자신을 제외한 다른 에이전트들의 목표 랜드마크와 메시지(message)를 입력 받고, 자기 자신으로부터 속도와 랜드마크까지의 상대적인 거리를 관찰한다. 관측한 값을 입력 받아 각 에이전트는 어느 방향으로 움직일지 혹은 정지할지 행동(Action)을 취한다. 그 후, 각 에이전트가 목표 랜드마크에 가까워질수록 보상을 많이 받도록 설정하여 학습시킨다.

나. 다중 에이전트 근접 정책최적화 학습 방법

다중 에이전트 행위자-비평가 학습방법에 더해 근접 정책최적화 방법을 이용한다.[3] 근접 정책최적화 방법은 최근 많은 연구에서 사용되고 있는 보편적인 강화학습 방법이며 신뢰 구역 정책최적화(TRPO)만큼의 성능을 가지며 신뢰 구역 정책최적화 기법의 근사를 통해 더욱 구현이 쉽고 복잡도가 낮다. 특히 잡음에 민감했던 신뢰 구역 정책최적화에 비해 더욱 다중 에이전트 간 간섭이 있고 잡음이 심한 통신상황에 적합하다.

다. 시뮬레이션 결과

다중 에이전트 간 통신상황을 학습 후, 각 에이전트는 자신의 목표로 성공적으로 나아가는 것을 확인했다. 본 논문에서는 다중 에이전트 근접 정책최적화 학습 방법을 통해 학습된 에이전트가 주고받는 방식을 시맨틱 통신의 관점에서 분석한다.

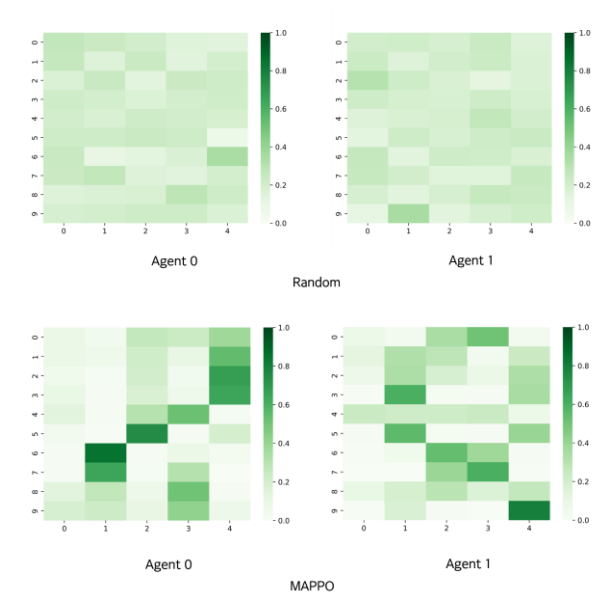

그림 2. 수신한 메시지에 대한 행동의 조건부확률

각 에이전트가 만드는 의미 없던 이머전트 통신 메시지들이 다중 에이전트 행위자-비평가 및 근접 정책최적화 방법으로 학습 후, 그림 2 와 같이 수신된 메시지에 따라 높은 확률로 특정 행동을 하도록 제어하는 의미가 형성되었다. 직접적인 학습 없이 형성된 의미를 분석하기위해서 본 논문에서는 3 가지의 지표를 사용했다. 첫 번째는 즉각적인 조정(Instantaneous Coordination, IC)로 긍정적인 경청(Positive Listening)을 나타내는 지표이다. 두 번째는 화자의 일관성(Speaker Consistency, SC)로, 에이전트의 메시지와 행동의 일치 정도를 나타내는 정량화한 지표이다.[4] 세 번째는 정보 엔트로피(Entropy)로, 에이전트가 생성하는 메시지의 정보 엔트로피를 이용하여 메시지를 분석하는 지표이다.

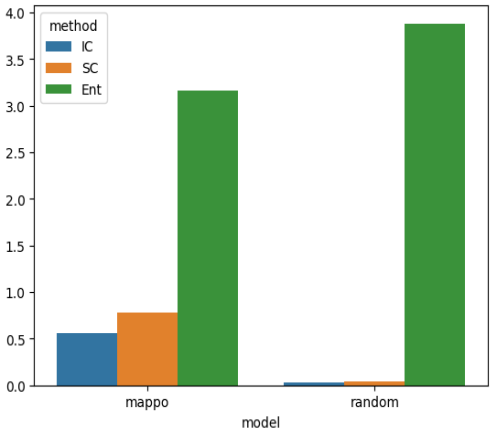

그림 3. 두 모델간 IC, SC, Entropy 값 비교

그림 3 에서 나타나듯, 다중 에이전트 근접 정책최적화 학습 후, 즉각적인 동조의 값과 화자의 일관성을 크게 증가하였고, 메시지 엔트로피의 양은 줄어든 것을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 최근 연구에서 자주 사용되는 기법인 다중 에이전트 근접 정책최적화 기법을 이용하여, 다중 에이전트 간 제어 통신 상황에서 생성되는 의미를 분석해 보았다. 추후 대형 언어모델 혹은 자연어처리 모델들을 이용하여 에이전트가 만드는 메시지의 의미를 인간의 관점에서 분석할 계획이다. 이는 시맨틱 통신 연구에 큰 기여를 할 것으로 기대된다.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2022-0-00420, Development of Core Technologies enabling 6G End-to-End On-Time Networking & No. 2023-11-1836)

참 고 문 헌

[1] M. Chafii, S. Naoumi, R. Alami, E. Almazrouei, M. Bennis, and M. Debbah, “Emergent Communication in Multi-Agent Reinforcement Learning for Future Wireless Networks,” arXiv preprint arXiv:2309.06021, 2023.

[2] C. Yu, A. Velu, E. Vinitsky, Y. Wang, A. Bayen, and Y. Wu, “The surprising effectiveness of PPO in cooperative, multi-agent games,” in Proc. NIPS, 2017

[3] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. Neural Information Processing Systems (NIPS), 2017.

[4] Lowe, R., Foerster, J., Boureau, Y. L., Pineau, J., & Dauphin, Y. (2019). On the Pitfalls of Measuring Emergent Communication. arXiv e-prints, arXiv-1903.