심층 강화학습 기반 다중 에이전트 협력을 통한 통신 기밀성 에너지 효율 최적화

장현성, 김정현

세종대학교

23012724@sju.ac.kr, j.kim@sejong.ac.kr

Optimizing Communication Secrecy Energy Efficiency via Multi-Agent Cooperation Based on Deep Reinforcement Learning

Hyeonseong Chang, Junghyun Kim

Sejong Univ.

요 약

본 논문에서는 무인항공기 (Unmanned Aerial Vehicle, UAV)와 지능형 재구성 표면 (Reconfigurable Intelligent Surface, RIS)을 사용하는 무선 통신 시스템의 통신 기밀성 에너지 효율 (Secrecy Energy Efficiency, SEE)을 향상하기 위해 심층 강화학습 기반 근위 정책 최적화 (Proximal Policy Optimization, PPO) 알고리즘을 적용하였다. 더 나아가, 신경망의 깊이를 늘려 표현력을 향상할 수 있는 SimBa 구조와 PPO를 결합한 새로운 모델을 제안한다. 실험 결과, 제안 모델은 SEE가 기존 모델 대비 약 4.3% 향상되었다.

Ⅰ. 서론

최근 5G 및 6G와 같은 차세대 통신 기술 분야에서는 무인항공기 (Unmanned Aerial Vehicle, UAV) 기반 통신 방식이 주목받고 있다 [1]. UAV는 사용자 통신 범위를 확장하고, 인구 밀집 지역에서 기지국의 트래픽을 분산 처리하는 데 효과적이다 [2]. 특히, 고고도에서 운용되는 특성 덕분에 가시거리 (Line-of-Sight, LoS) 경로를 보다 많이 확보할 수 있어 무선 전송 품질을 향상시킬 수 있다 [3]. 그러나 이러한 UAV 기반 무선 통신은 감청 등 악의적인 공격에 취약하다는 보안 한계를 지닌다.

감청과 같은 물리적 공격에 대한 해결책 중 하나는 UAV의 이동 경로를 변경하는 것이다. 예를 들어 감청자 (eavesdropper)가 한 지점에 고정되어 있다고 가정했을 때, UAV의 이동 경로를 조정하여 감청 범위에서 벗어나 기밀성을 유지할 수 있다. 한편, UAV의 이동 경로를 조정하는 것은 통신 경로 제한으로 인해 커버리지가 감소된다는 단점이 있다.

UAV의 통신 경로를 다양하게 만들어 통신 커버리지를 늘리는 해결책 중 하나로 지능형 재구성 표면 (Reconfigurable Intelligent Surface, RIS)이 있다. RIS는 통신 신호를 반사하여 사용자에게 우회적으로 도달할 수 있도록 하는 장치이다.

본 논문에서는 높은 통신 기밀성 에너지 효율 (Secrecy Energy Efficiency, SEE)을 달성하기 위해 RIS를 사용하는 UAV 통신 시나리오를 가정한다 [3]. RIS와 UAV 각각을 심층 강화학습 에이전트로 만들어 협동 학습을 수행한다.

기존 모델 [3]은 정책 수렴이 불안정하고 정책 신경망과 가치 신경망의 깊이가 3층에 불과하여 표현력이 제한된다. 본 논문에서는 보편적으로 우수한 성능을 보이는 SimBa 구조 [5]를 사용하여 신경망을 심층화하고 복잡한 환경에서도 잘 학습하도록 모델을 설계했다.

Ⅱ. 본론

본 논문에서는 그림 1의 시나리오를 가정한다. 기지국에서 송신한 신호가 건물 등 장애물에 의해 가로막히므로 UAV와 RIS를 통해 기지국과 사용자 간 신호를 우회·반사하여 통신한다. 이때, 감청자는 UAV와 RIS의 송신·반사되는 신호를 감청한다고 가정한다. 초록색 선들은 정상적인 다운 링크 (down-link) 신호들을 의미하고, 빨간색 선들은 감청자가 감청한 신호에 해당한다. 점선은 RIS에 반사된 신호이다.

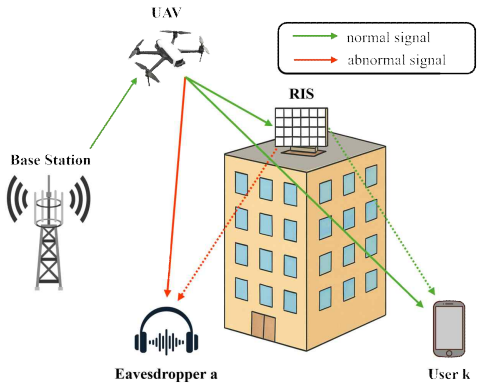

그림 1. UAV와 RIS 기반 통신 시스템 감청 시나리오

그림 1에 표현된 시나리오는 마르코프 의사결정 과정 (Markov Decision Process, MDP)으로 다음과 같이 표현될 수 있다:

$$e = \{s_t, a_t, r_t, s_{t+1}\}, \tag{1}$$

e는 에피소드를 의미하고, 전체 시간을 0.1초 간격으로 나눈 타임스텝 (time step) t에서 상태 s_t, 행동 a_t, 보상 r_t, 다음 상태 s_{t+1}로 표현된다. UAV의 상태는 UAV의 3차원 위치 좌표를 표현하고, RIS의 상태는 UAV에서 송신한 신호의 채널 정보를 의미한다. UAV의 행동은 이전 위치에서 d만큼 더한 움직임으로, UAV의 변위를 의미한다. RIS의 행동은 RIS 각 반사 소자의 수동 빔포밍 각도 조절 행위이다. 다음 상태 s_{t+1}은 에이전트가 상태 s_t에서 행동을 취하여 변화된 상태를 의미한다. 보상은 다음과 같다:

$$r_t = \tanh\left(\sum_{k=1}^{K} \mathrm{R}_{\mathrm{k}}^{\mathrm{sec}}[t] - c_1 p_m - c_2 p_r - c_3 p_g - c_4 p_e\right), \tag{2}$$

$\mathrm{R}_k^{\mathrm{sec}}[t]$는 k번째 사용자가 타임 스텝 t에서 달성한 기밀성 전송률이고, 하이퍼볼릭 탄젠트를 사용해 -1부터 1까지의 범위로 조절했다. p는 페널티 (penalty) 항들을 의미하고, c는 각 페널티의 계수들을 의미한다. 페널티들은 특정 조건을 만족할 때 활성화된다. p_m은 UAV의 이동 범위를 제

한한다. p_r은 전송률 $\mathrm{R_t^{sec}[t]}$의 전송률을 최소 보장치 이상으로 유지하도록 제한하며, p_g는 UAV의 능동 빔포밍 전력을 제한한다. 그리고 p_e는 전송률이 0일 때, 한 타임 스텝 t당 UAV의 에너지 소비량을 제한한다. 마지막으로 $\mathrm{R_k^{sec}}[t]$는 다음과 같다:

$$\mathrm{R_k^{sec}}[t] = [\mathrm{R_k^U[t]} - \max_{\forall a} \mathrm{R_{a,k}^E[t]}]^{\dagger}, \quad (3)$$

$\mathrm{R_k^U[t]}$는 타임 스텝 t에서 k번째 사용자의 보안 수신 신호 비율이며, $\mathrm{R_{a,k}^E}[t]$는 모든 감청자 a에 대해 a가 감청한 k번째 사용자의 수신 신호이다. †는 양수 부분만 반환한다는 의미이다. U와 E는 각각 사용자와 감청자를 의미한다.

제안 모델의 기본 강화학습 알고리즘으로 기존 모델에 비해 학습 곡선이 완만하고 수렴이 안정적인 근위 정책 최적화 (Proximal Policy Optimization, PPO)를 사용하였다 [4]. PPO는 다음과 같은 클립된 대체 목적 함수 (clipped surrogate objective)를 사용한다:

$$L(\theta) = \hat{\mathrm{E}}[\min(\mathrm{r_t}(\theta)\hat{\mathrm{A}}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t], \quad (4)$$

$r_t(\theta)$는 타임 스텝 t에서 현재 정책과 이전 정책의 비이고, $\hat{A}_t$는 타임 스텝 t에서 현재 정책의 이점 (advantage)을 나타낸다. 클립 (clip)은 $r_t(\theta)$를 1에서 허용범위 ϵ만큼 조정한다.

기존 모델의 얕은 신경망으로 인한 표현력 한계를 해결하기 위해 레이어를 심층화할 수 있는 SimBa 구조를 적용했다. SimBa는 레이어 정규화 (Layer Normalization, LayerNorm)와 잔차 연결 (Residual Connection), 이동 통계 정규화 (Running Statistics Normalization, RSNorm)를 조합한 구조이다. 제안한 모델의 구조는 그림 2에서 확인할 수 있다.

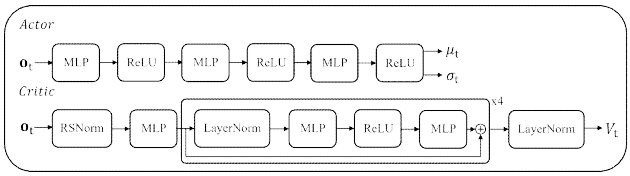

그림 2. 제안 모델 구조

먼저 레이어 정규화는 다음과 같다:

$$\mathbf{z}_t = \mathrm{LayerNorm}(\mathbf{x}_t^{\mathrm{L}}), \quad (5)$$

$\mathbf{z}_t$는 크리틱 (Critic)에서 마지막에 수행되는 정규화된 출력을 의미하고, $\mathbf{x}_t^{\mathrm{L}}$는 레이어 정규화를 할 계층의 입력이다. 레이어 정규화는 PPO에서 경험 재사용 시 발생할 수 있는 분산 변동을 줄여 안정적인 학습을 가능하게 한다. 다음으로 잔차 연결은 다음과 같다:

$$\mathbf{x}_t^{l+1} = \mathbf{x}_t^{l} + \mathrm{MLP}(\mathrm{LayerNorm}(\mathbf{x}_t^{l+1})), \quad (6)$$

$\mathbf{x}_t^{l}, \mathbf{x}_t^{l+1}$는 각각 l번째 레이어의 입력, $l+1$번째 레이어의 입력이다. 잔차 연결은 층이 깊어질수록 소실될 수 있는 입력 정보를 보존하여 안정적인 학습을 가능하게 한다. 마지막으로 이동 통계 정규화는 다음과 같다:

$$\overline{\mathbf{o}}_t = \mathrm{RSNorm}(\mathbf{o}_t) = \frac{\mathbf{o}_t - \mu_t}{\sqrt{\sigma_t^2 + \epsilon}}, \quad (7)$$

RSNorm은 관찰값 (observation) $\mathbf{o}_t$을 평균 0, 표준 편차 1로 정규화한다. ϵ은 분모가 0이 되는 것을 방지하는 작은 상수이다. 평균 μ_t와 표준 편차 σ_t는 다음과 같다:

$$\mu_t = \mu_{t-1} + \frac{1}{t}\delta_t, \ \ \sigma_t^2 = \frac{t-1}{t}(\sigma_{t-1}^2 + \frac{1}{t}\delta_t^2), \quad (8)$$

δ_t는 $\delta_t = \mathbf{o}_t - \mu_{t-1}$로 현재 관찰값과 이전 관찰값들 평균의 차이다.

실험 결과는 그림 3과 표 1에서 확인할 수 있다. 기존 모델 대비 학습 안정성이 개선되었으며 평균 SEE가 약 4.3% 향상되었다. 평균 SSR은 6.7% 향상되었다. 비록 총 에너지 소모량은 1.8% 증가했으나, SEE 향상으로 에너지 효율이 개선된 긍정적인 결과로 해석할 수 있다.

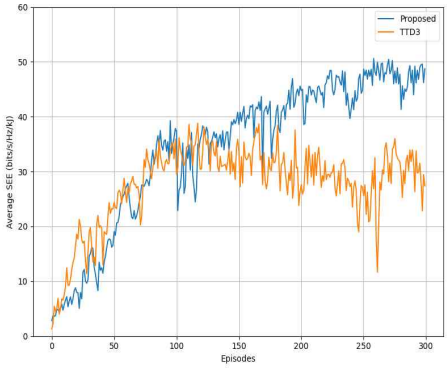

그림 3. 기존 모델 (TTD3)와 제안 모델의 학습 곡선 비교 (SEE 기준)

표 1. 기존 모델과 제안 모델의 성능 비교

알고리즘	평균 SSR	총 에너지 소모량 (kJ)	평균 SEE
기존 모델[3]	5.39	11.2	48.4
제안 모델	5.75	11.4	50.5

Ⅲ. 결론

본 논문에서는 PPO를 적용하고 SimBa 구조와 결합하여 새로운 모델을 제안하였다. 실험 결과 기존 모델 대비 평균 SEE가 약 4.3%, 평균 SSR이 약 6.7% 향상되었다. UAV와 RIS의 소비 전력으로 인한 총 에너지 소모량은 1.8% 증가했지만, 에너지 효율이 개선되었으므로 긍정적인 성과로 평가할 수 있다. 향후 연구에서는 모델 구조 개선을 통해 총 에너지 소모량을 줄이고, 경험 버퍼를 효율적으로 재설계하는 방안을 탐구할 예정이다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (RS-2023-00271991).

참 고 문 헌

[1] Zeng, Y., Zhang, R. and Lim, T. J., "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36-42, May 2016.

[2] Lee, J. and Friderikos, V., "Interference-aware path planning optimization for multiple UAVs in beyond 5G networks," *Journal of Communications and Networks*, vol. 24, no. 2, pp. 125-138, Apr. 2022.

[3] Tham, M. L., Wong, Y. J., Iqbal, A., Ramli, N. B., Zhu, Y. and Dagiuklas, T., "Deep reinforcement learning for secrecy energy-efficient UAV communication with reconfigurable intelligent surface," in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC)*, Glasgow, U.K., 2023.

[4] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., "Proximal policy optimization algorithms," *arXiv*:1707.06347, Jul. 2017.

[5] Lee, H., Hwang, D., Kim, D., Kim, H., Tai, J. J., Subramanian, K., Wurman, P. R., Choo, J., Stone, P. and Seno, T., "SimBa: Simplicity bias for scaling up parameters in deep reinforcement learning," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2025.