Ứng dụng phương pháp học tăng cường đa tác nhân giải bài toán lựa chọn phương tiện hỏa lực trong hệ thống tự động hóa chỉ huy-điều khiển

Nguyễn Xuân Trường[1*], Vũ Hỏa Tiễn[2], Hoàng Văn Phúc[1],
Nguyễn Quang Thi[1], Vũ Chí Thanh[3]

[1]Viện Tích hợp hệ thống, Học viện Kỹ thuật Quân sự, số 236 Hoàng Quốc Việt, Bắc Từ Liêm, Hà Nội, Việt Nam;
[2]Viện Tên lửa và Kỹ thuật điều khiển, Học viện Kỹ thuật Quân sự, Số 236 Hoàng Quốc Việt, Bắc Từ Liêm, Hà Nội, Việt Nam;
[3]Viện Ra đa, Viện Khoa học và Công nghệ quân sự, Số 17 Hoàng Sâm, Cầu Giấy, Hà Nội, Việt Nam.
[*]Email: truongnx@mta.edu.vn
Nhận bài: 19/01/2024; Hoàn thiện: 11/3/2024; Chấp nhận đăng: 08/4/2024; Xuất bản: 22/04/2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.94.2024.11-21

TÓM TẮT

Bài báo trình bày phương pháp học tăng cường sâu đa tác nhân giải bài toán lựa chọn phương tiện hỏa lực (PTHL) động trong hệ thống TĐH CH-ĐK phòng không. Mô hình hoạt động của PTHL được xây dựng dựa trên dự đoán quỹ đạo tối ưu của các mô hình mục tiêu trên không đã được huấn luyện trước đó [1] và trạng thái các đối tượng trên mặt đất, cũng như phương án tối ưu phối hợp hoạt động của các PTHL trong hệ thống. Mô hình PTHL được xây dựng trên bộ thư viện OpenAI Gym sử dụng thuật toán học tăng cường sâu (DQL) để tối ưu hóa hàm giá trị Q. Sau khi được huấn luyện qua 200 nghìn vòng, mô hình PTHL đã có khả năng tự động phân tích, nhận thức tình huống, phối hợp các PTHL trong hệ thống, xây dựng phương án tương tác đối kháng động và chọn ra phương án tối ưu có tính tới các ràng buộc thực tế, để thu được giá trị cực tiểu của hàm tổn thất tổng thể cho toàn bộ quá trình chiến đấu. So với mô hình PTHL sử dụng thuật toán PPO được huấn luyện trong cùng một điều kiện môi trường, sau 1000 chu trình tác chiến tương tác với mô hình mục tiêu trên không, mô hình PTHL đề xuất đạt tỉ lệ chiến thắng 89,1% lớn hơn nhiều so với 77,2% của mô hình sử dụng thuật toán PPO.

Từ khoá: Học tăng cường; Tự động hóa chỉ huy; C4I; DWTA; DQL; OpenAI Gym.

1. MỞ ĐẦU

Bài toán lựa chọn PTHL (*hay bài toán chỉ định vũ khí*) phòng không là bài toán cốt lõi trong hệ thống tự động hóa (TĐH) chỉ huy – điều khiển, hỗ trợ người chỉ huy ra quyết định lựa chọn PTHL để tiêu diệt mục tiêu, nhằm phát huy tối đa khả năng của các PTHL khác nhau về chủng loại trong cụm phòng không hỗn hợp [2, 3]. Bài toán lựa chọn PTHL là bài toán tối ưu tổ hợp, có không gian nghiệm mở rộng theo cấp số nhân với sự gia tăng của số lượng PTHL và mục tiêu hàng không. Ngoài ra, một quyết định hợp lý về lựa chọn PTHL cần quan tâm tới tính chất đa yếu tố tác động và các đặc tính kỹ - chiến thuật của mỗi loại PTHL cụ thể trong hệ thống. Mô hình bài toán lựa chọn PTHL được mô tả trong hình 1 [4].

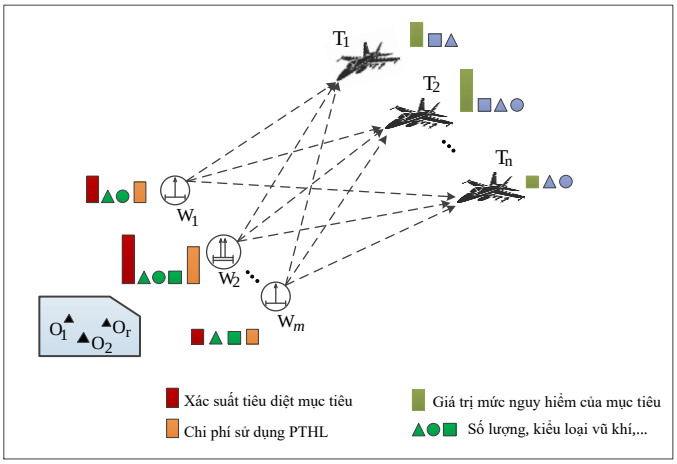
Các phương pháp giải bài toán lựa chọn PTHL đã được đề xuất [2, 5, 6], gồm: phương pháp quy hoạch tuyến tính, quy hoạch đồ thị ngẫu nhiên và quy hoạch hỗn hợp số nguyên; sử dụng thuật toán tối ưu siêu mô phỏng. Ngoài ra, trong [7, 8] trình bày các phương pháp giải bài toán ở dạng ma trận được phát triển từ thuật toán Hungary. Ứng dụng phương pháp học tăng cường (*Reinforcement Learning*) để giải bài toán lựa chọn PTHL được Mouton và cộng sự trình bày trong [9] (2011), và so sánh hai thuật toán học tăng cường Monte Carlo ES và TD Q-Learning giải bài toán lựa chọn PTHL rút gọn. Tong Wang và cộng sự [10] (2023), ứng dụng mạng học sâu Q (*Deep Q-Learning Network*) với thuật toán đàn ong nhân tạo đa mục tiêu cải tiến để lựa chọn PTHL không người lái trên mặt đất trong môi trường tác chiến đô thị phức tạp. Trong [11] (2023), B. Gaudet và cộng sự ứng dụng phương pháp học sâu tăng cường (*Deep Reinforcement Learning*) để tối ưu hóa chính sách lựa chọn PTHL tiêu diệt đa mục tiêu siêu thanh.

Bài báo trình bày một phương pháp giải bài toán lựa chọn PTHL động (*Dynamic weapon target assignment -DWTA*) sử dụng kỹ thuật học tăng cường sâu đa tác nhân (*Multi-Agent Deep Reinforcement Learning*), trong đó, các PTHL phòng không đóng vai trò là các tác nhân (agent), được xây dựng và huấn luyện trên bộ thư viện OpenAI Gym [12], đưa ra quyết định tiêu diệt mục tiêu trên không trong môi trường tác chiến phức tạp, không chắc chắn. Khác với các phương pháp trước đây, trong phương pháp đề xuất mô hình PTHL được huấn luyện bằng cách tương tác trực tiếp với mô hình mục tiêu trên không (một mô hình AI) đã được huấn luyện trước đó (xem trong [1]), trong môi trường tác chiến động. Điều này làm tăng tính đa dạng và sát với thực tế của bộ dữ liệu đầu vào huấn luyện. Thuật toán học sâu Q (*Deep Q-Learning – DQL*) được sử dụng để tối ưu hóa hàm giá trị Q (*Q-value function*) của mô hình PTHL đề xuất, bằng kết quả thử nghiệm chứng minh, với số lượng lớn PTHL thì thuật toán DQL giúp mô hình PTHL có khả năng học nhanh hơn so với thuật toán tối ưu hóa luật tiệm cận (Proximal Policy Optimization - PPO) [11, 13]. Sau khi được huấn luyện, các mô hình PTHL được thử nghiệm qua 1000 chu trình, kết quả thu được mô hình PTHL với thuật toán DQL đạt tỉ lệ chiến thắng 89,1%, so với tỉ lệ 77,2% của mô hình PPO. Điều này chứng minh mô hình PTHL đề xuất có khả năng ứng dụng trong các hệ thống TĐH CH-ĐK phòng không thời gian thực.

2. XÂY DỰNG MÔ HÌNH TOÁN HỌC BÀI TOÁN LỰA CHỌN PTHL ĐỘNG TRONG HỌC TĂNG CƯỜNG SÂU

2.1. Xây dựng mô hình toán học của bài toán lựa chọn PTHL phòng không động

Trong bài toán lựa chọn PTHL động, toàn bộ khoảng thời gian tấn công của kẻ địch (mục tiêu trên không) tính từ khi được phát hiện đến khi bị tiêu diệt hoàn toàn hoặc hoàn thành nhiệm vụ chiến đấu là K, được chia thành các đoạn thời gian tương ứng với chu kỳ bắn của các PTHL. Chu kỳ bắn của các PTHL có độ dài đủ để cho các PTHL trong cụm phòng không bắn một tập hợp con vũ khí (tên lửa) của nó và quan sát (một cách hoàn hảo) kết quả của tất cả các lần bắn. Với phản hồi của thông tin này, hệ thống TĐH CH-ĐK chọn một nhóm PTHL và chỉ định chúng tiêu diệt các mục tiêu. Trong mỗi chu kỳ, các PTHL được chọn và chỉ định tiêu diệt mục tiêu nhằm tối đa hóa tổng giá trị dự kiến thiệt hại của các mục tiêu trên không và tối thiểu hóa chi phí sử dụng PTHL ở giai đoạn cuối cùng của cuộc chiến.



Hình 1. *Mô hình bài toán lựa chọn PTHL trong hệ thống TĐH CH-ĐK phòng không.*

Bài toán lựa chọn PTHL động trong hình 1, bao gồm các tham số đầu vào như sau:

$\mathbf{T}=\{T_1,T_2,...,T_n\}$ là tập hợp gồm *n* mục tiêu được phát hiện bởi các cảm biến (các đài ra đa) trong khu vực tác chiến của hệ thống TĐH CH-ĐK;

$\mathbf{W}=\{W_1,W_2,...,W_m\}$ là tập hợp gồm *m* PTHL phòng không kết nối vào hệ thống;

$\mathbf{O}=\{O_1,O_2,...,O_r\}$ là tập hợp gồm *r* đối tượng cần bảo vệ trên mặt đất;

V_{ir} là giá trị mức nguy hiểm của một mục tiêu T_i đối với một đối tượng cần bảo vệ O_r.

Trên cơ sở tối đa hóa khả năng gây thiệt hại cho mục tiêu hàng không, giải bài toán lựa chọn PTHL động là tìm phương án chỉ định vũ khí nhằm tối đa giá trị của hàm mục đích tại một chu kỳ t trong biểu thức (1).

$$F_t(X^t)=\sum_{i=1}^{n(t)}V_i\left(1-\prod_{k=t}^{K}\prod_{j=1}^{m(t)}\left[\left(1-p_{ij}(k)\right)^{x_{ij}(t)}\right]\right) \quad (1)$$

trong đó: K là toàn bộ khoảng thời gian tấn công của kẻ địch (mục tiêu trên không); $n(t)$ và $m(t)$ tương ứng là số lượng mục tiêu trên không và số lượng PTHL còn lại trong chu kỳ t; V_i là giá trị mức nguy hiểm tổng thể của mục tiêu T_i với tất cả các đối tượng cần bảo vệ; $X^t=\left[X_t,X_{t+1},...,X_k\right]$ với $X_t=\left[x_{ij}(t)\right]_{W\times T}$ là ma trận quyết định tại chu kỳ t, và $x_{ij}(t)$ là biến điều khiển có giá trị: $x_{ij}(t)=1$ khi PTHL W_j được lựa chọn tiêu diệt mục tiêu T_i; ngược lại $x_{ij}(t)=0$; $p_{ij}(k)$ - Xác suất PTHL W_j tiêu diệt mục tiêu T_i trong một chu kỳ bắn k;

Trên cơ sở tối thiểu hóa chi phí sử dụng PTHL trong mọi chu kỳ bắn, cần tìm giá trị tối thiểu của hàm mục đích chi phí sử dụng PTHL trong biểu thức (2).

$$F_c=\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j u_{ij}(k)x_{ij}(k) \quad (2)$$

trong đó: u_{ij} là số lượng đạn tiêu thụ của PTHL W_j để tiêu diệt mục tiêu T_i trong chu kỳ k; β_j là chi phí một đơn vị đạn của PTHL W_j; $x_{ij}(k)$ là tham số điều khiển tại chu kỳ *k*.

Như vậy, giải bài toán lựa chọn PTHL động theo tiêu chí "Hiệu quả - Chi phí", là tìm giá trị tối thiểu của các hàm như trong biểu thức (3).

$$\begin{aligned}&\min f_1=\frac{1}{F_t(X^t)}\\&\min f_2=F_c\end{aligned} \quad (3)$$

Với các điều kiện ràng buộc của bài toán lựa chọn PTHL động được mô tả bằng các biểu thức (4), (5), (6), như sau:

$$\sum_{j=1}^{m}x_{ij}(t)\le m_i,\quad \forall i\in\{1,2,...,n\},\quad \forall t\in\{1,2,...,K\} \quad (4)$$

$$\sum_{i=1}^{n}x_{ij}(t)\le n_j,\quad \forall j\in\{1,2,...,m\},\quad \forall t\in\{1,2,...,K\} \quad (5)$$

$$\sum_{i=1}^{n}\sum_{t=1}^{K}x_{ij}(t)\le N_j,\quad \forall j\in\{1,2,...,m\} \quad (6)$$

Biểu thức (4), biểu thị m_i là số lượng tối đa các PTHL được sử dụng để tiêu diệt một mục tiêu T_i ở giai đoạn *t*. Biểu thức (5) chỉ ra rằng, n_j là số lượng mục tiêu tối đa, mà một PTHL W_j có thể xạ kích đồng thời trong một chu kỳ *t* hay là số kênh mục tiêu của PTHL W_j. Biểu thức (6), giới hạn số lượng đạn có sẵn của mỗi PTHL W_j trong toàn bộ thời gian chiến đấu *K*.

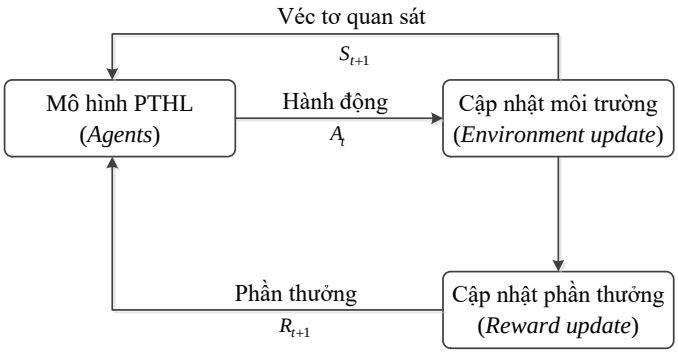
Ngoài ra, có thể tính đến điều kiện ràng buộc về cự ly của mục tiêu so với kích thước vùng hỏa lực của PTHL, được mô tả trong biểu thức (7).

$$x_{ij}(t) \le d_{ij}(t),\ \forall i \in \{1,2,...,n\},\ \forall j \in \{1,2,..,m\},\ \forall t \in \{1,2,..,K\} \quad (7)$$

trong đó, $d_{ij}(t)$ có giá trị là 0 hoặc 1. Nếu trong chu kỳ *t* mục tiêu T_i nằm trong vùng hỏa lực của PTHL W_j thì $d_{ij}(t)=1$ và ngược lại $d_{ij}(t)=0$.

2.2. Xây dựng mô hình học tăng cường sâu giải bài toán lựa chọn PTHL động

Mô hình học tăng cường đề xuất, được xây dựng theo quy trình quyết định Markov (Markov Decision Process – MDP), trong đó, PTHL là một mô hình AI, đóng vai trò là tác nhân (Agent), đang ở một trạng thái (State), đưa ra quyết định tiêu diệt mục tiêu trong môi trường (Environment) tác chiến phức tạp, bằng cách thực hiện các hành động (Action) dựa trên chính sách được thiết kế sẵn, mỗi hành động được thực hiện sẽ trả về một giá trị kết quả được gọi là phần thưởng (Reward). Kỹ thuật học tăng cường tập trung vào việc làm thế nào để tác nhân (PTHL) trong một môi trường có thể hành động sao cho đạt được phần thưởng tối đa qua các vòng lặp huấn luyện. Sơ đồ cấu trúc mô hình học tăng cường PTHL được trình bày trong hình 2.



Hình 2. *Cấu trúc mô hình học tăng cường PTHL theo quy trình quyết định Markov.*

Mô hình học tăng cường sâu được xây dựng và huấn luyện trên bộ thư viện OpenAI Gym [14] sử dụng giải thuật học sâu Q (DQL) [11], gồm các thành phần chính sau:

• ***Môi trường huấn luyện*** *(Enviroment)* được giả lập để đào tạo mô hình PTHL, cung cấp không gian để thực hiện các nhiệm vụ hoặc giải quyết các bài toán cụ thể bằng cách tương tác với môi trường. Mô hình huấn luyện tương tác được xây dựng, gồm các tham số đầu vào:

- Số lượng mục tiêu trên không, tọa độ vị trí, vận tốc, hướng bay, độ cao của từng mục tiêu;
- Số lượng PTHL, tọa độ vị trí của mỗi PTHL và kích thước vùng quan sát, vùng tiêu diệt; chu kì xạ kích,…
- Số lượng đối tượng phòng thủ cần bảo vệ, tọa độ vị trí và phạm vi bảo vệ của mỗi đối tượng phòng thủ;
- Kích thước bản đồ khu vực huấn luyện tương tác 1000 x 1000 (pixel);

• ***Không gian quan sát*** *(Observation space)* là các trạng thái có thể xảy ra mà tác nhân nhận được từ môi trường sau mỗi vòng (round), cung cấp thông tin cho tác nhân để đưa ra các quyết định. Trong mô hình huấn luyện, không gian quan sát được thiết lập dưới dạng hàm rời rạc **Dict**{}, bên trong đó có các trường thông tin được trích xuất từ môi trường huấn luyện, bao gồm:

- Trạng thái của các đối tượng cần bảo vệ O_{active}^{state} nhận các giá trị 0 và 1, $O_{active}^{state}=0$ - đối tượng cần bảo vệ bị tiêu diệt và ngược lại;
- Trạng thái của các mục tiêu trên không T_{active}^{state} nhận các giá trị 0 và 1, $T_{active}^{state}=0$ - mục tiêu bị PTHL tiêu diệt và ngược lại;
- Các tham số quỹ đạo của mục tiêu và tham số trạng thái của PTHL trong hệ thống.

• ***Không gian hoạt động*** *(Action space)* là không gian véctơ chứa tất cả các phương án quyết định của tác nhân tại một thời điểm xác định. Đối với mô hình DQL, mỗi hành động sẽ có giá trị

Q tương ứng, đầu ra của mô hình DQL là mảng một chiều tương ứng với bảng Q. Trong bài toán lựa chọn PTHL đa tác nhân, sau khi lựa chọn hành động có giá trị Q cao nhất, cần phải chuyển đổi từ kết quả được lựa chọn thành hành động tương ứng cho từng tác nhân. Do đó, không gian hoạt động của mô hình DQL sẽ có giá trị A xác định trong khoảng: $0 \le A \le n^m - 1$. Trong đó, n – Số lượng mục tiêu trên không, m – số lượng PTHL.

• ***Hàm phần thưởng*** *(Reward function)* có mục tiêu là đưa ra chiến lược thực hiện hành động của PTHL. Trong bài toán lựa chọn PTHL động, mục tiêu của các PTHL (tác nhân) là tối đa hóa tổng phần thưởng mà nó nhận được trong toàn bộ các giai đoạn của cuộc chiến. Phần thưởng nhận được sau mỗi lần thực hiện hành động (xạ kích) giúp PTHL xác định được mức độ tốt – xấu của hành động, cũng là cơ sở để thay đổi chiến lược. Chiến lược tính phần thưởng và cập nhật môi trường được đặt ra để PTHL có thể học tập sau mỗi lần đưa ra quyết định với mục tiêu là tiêu diệt các mục tiêu trên không để bảo vệ các đối tượng phòng thủ trên mặt đất. Phần thưởng một PTHL nhận được trong một chu kỳ bắn được tính bởi biểu thức (8).

$$R_{W_j} = \omega_A R_A + \omega_{xk} R_{xk} - \omega_{hl} \sum_{i=0}^{n} T_{hl}^{i} - \omega_s (1 - S_{T_i}) \quad (8)$$

trong đó: R_A - Phần thưởng khi PTHL tiêu diệt được mục tiêu trên không;

ω_A - Hệ số phần thưởng tiêu diệt mục tiêu;

R_{xk} - Phần thưởng sau mỗi lần xạ kích, được tính theo hàm $R_{xk} = \lg(n_{xk})$;

n_{xk} - Số lần xạ kích, có thể nhận các giá trị 0, 1, 2, 3;

ω_{xk} - Hệ số phần thưởng xạ kích;

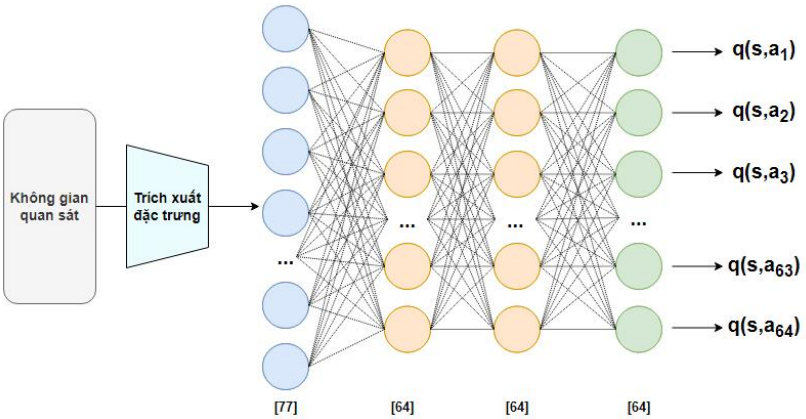
T_{hl}^{i} - Thời gian mục tiêu T_i nằm trong vùng hỏa lực của PTHL W_j;

ω_{hl} - Hệ số phần thưởng khi mục tiêu nằm trong vùng hỏa lực;

S_{T_i} - Trạng thái hoạt động của mục tiêu được lựa chọn, nhận giá trị $S_{T_i} = 0$ nếu mục tiêu bị tiêu diệt, và $S_{T_i} = 1$ nếu mục tiêu đang hoạt động;

ω_s - Hệ số phần thưởng trạng thái của mục tiêu.

- ***Giải thuật học sâu Q (Deep Q-Learning)***



Hình 3. *Kiến trúc mô hình học sâu Q (Deep Q-Learning) sử dụng để huấn luyện mô hình học tăng cường PTHL.*

Trong bài toán lựa chọn PTHL động (DWTA) với số lượng lớn các PTHL (tác nhân), mục tiêu trên không và đối tượng cần bảo vệ, tạo thành một không gian trạng thái lớn, nhiều chiều, do đó, áp dụng bảng Q trở lên không thực tế. Để giải quyết vấn đề này, một giải pháp được sử dụng để ước tính giá trị $Q(s,a)$ là thay thế bảng Q bằng một mạng nơron để học được cách ước lượng

giá trị Q cho các hành động của PTHL một cách chính xác (xem hình 3). Hàm thiệt hại (Loss Function) phải tính được sai số giữa giá trị Q thực tế và dự đoán, như trong biểu thức (9).

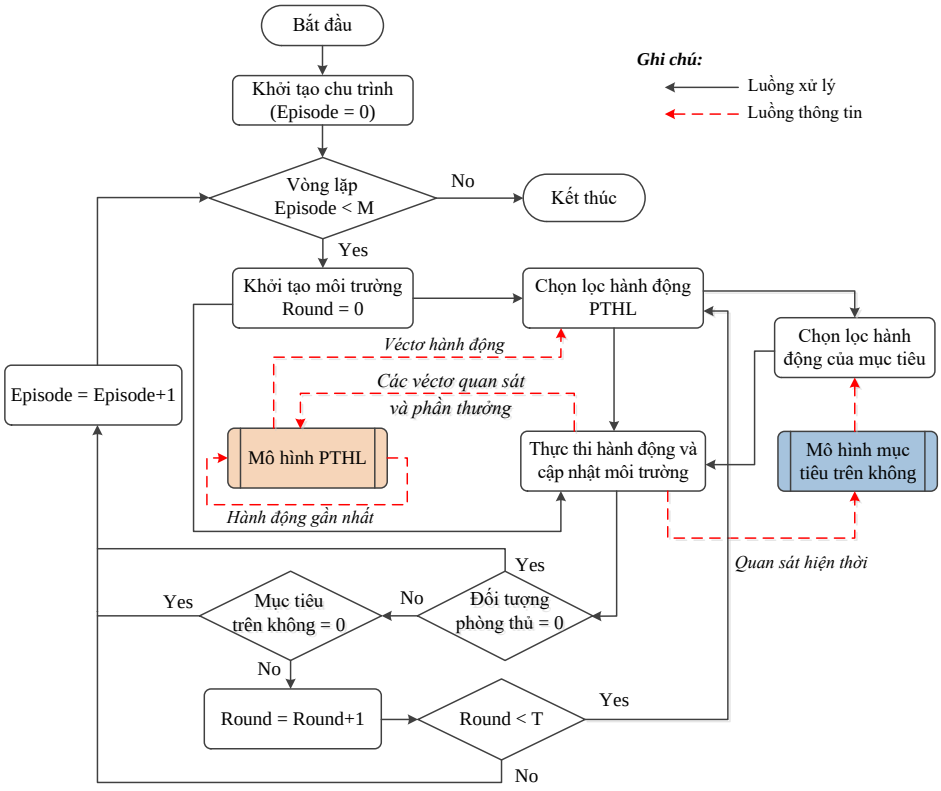
$$f_{Loss} = \left(R(s,a) + \gamma \max_{a'} Q(s',a',\theta') - Q(s,a,\theta) \right)^2 \quad (9)$$

Mô hình DQL được xây dựng ở dạng chiến lược đa đầu vào (Multi Input Policy), đầu vào mô hình là một không gian quan sát rời rạc chứa các thông tin mà mô hình PTHL thu thập được bao gồm: trạng thái của mục tiêu trên không; trạng thái của đối tượng cần bảo vệ; góc tiếp cận và cự li của từng mục tiêu đối với đối tượng cần bảo vệ; cự li của từng mục tiêu với PTHL,... Dữ liệu đầu vào sau khi được trích xuất đặc trưng, được đưa vào mạng nơ ron 3 lớp tuyến tính với hàm kích hoạt ReLU, đầu ra của lớp cuối cùng chính là bảng Q (Q table) của mô hình, là cơ sở dùng để mô hình PTHL đưa ra quyết định cho các tác nhân.

3. HUẤN LUYỆN MÔ HÌNH HỌC TĂNG CƯỜNG SÂU, MÔ PHỎNG TƯƠNG TÁC THEO KỊCH BẢN PHÒNG KHÔNG, ĐÁNH GIÁ KẾT QUẢ

3.1. Lưu đồ thuật toán huấn luyện mô hình học tăng cường sâu

Mô hình học tăng cường cho PTHL được huấn luyện trong môi trường có tích hợp mô hình mục tiêu trên không đã được xây dựng và huấn luyện trước đó (trong [1]), làm dữ liệu đầu vào để huấn luyện tương tác. Mô hình của PTHL và mục tiêu trên không sẽ luân phiên tương tác và làm thay đổi môi trường, giá trị phần thưởng cho PTHL sẽ được tính toán sau mỗi vòng (round) trong một chu trình (episode), làm cơ sở để cập nhật các thông số tối ưu cho mô hình. Thuật toán huấn luyện mô hình học tăng cường PTHL được trình bày trong lưu đồ thuật toán hình 4.



Hình 4. *Lưu đồ thuật toán huấn luyện tương tác mô hình học tăng cường PTHL với mô hình mục tiêu trên không đã được huấn luyện làm dữ liệu đầu vào.*

Mô hình học tăng cường được huấn luyện với số chu trình (episode) nhất định ($n_{Episode} < M$),

được khởi tạo với giá trị ban đầu bằng 0. Khi bắt đầu mỗi episode, môi trường sẽ được khởi tạo, mô hình của PTHL được huấn luyện sẽ tương tác với mô hình mục tiêu trên không làm thay đổi môi trường. Trong mỗi vòng (round) của episode, các mô hình AI thực hiện các hành động:

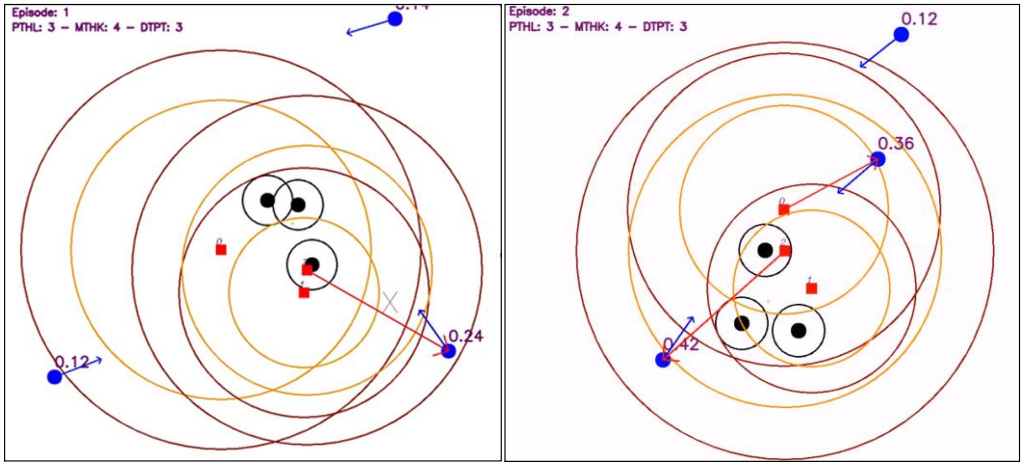
- Mô hình mục tiêu dựa trên quan sát môi trường sẽ đưa ra các hành động (action) thay đổi: vận tốc, hướng bay, độ cao,… làm thay đổi môi trường và mức độ nguy hiểm của mục tiêu.

- Mô hình PTHL sẽ dựa trên quan sát môi trường sau hành động của mục tiêu trên không, hành động của PTHL ở vòng trước và phần thưởng ở vòng trước để tối ưu các tham số trong mô hình, đồng thời đưa ra hành động mới từ mô hình vừa được tối ưu.

- Sau mỗi vòng, chương trình sẽ kiểm tra điều kiện kết thúc của chu trình như sau: Số lượng đối tượng phòng thủ được bảo vệ bằng 0 – toàn bộ đối tượng phòng thủ bị tiêu diệt, khi đó, bên tấn công đường không dành chiến thắng; Số lượng mục tiêu trên không bằng 0 – toàn bộ mục tiêu trên không bị tiêu diệt, khi đó, bên phòng thủ chiến thắng; Khi số vòng $n_{round} \geq T$ với T là số vòng tối đa trong một chu trình (episode) – kết quả trận chiến là hòa.

3.2. Phương pháp, công cụ mô phỏng

Môi trường học tăng cường đa tác nhân được xây dựng theo kịch bản phòng không (hình 5).



Hình 5. *Kịch bản phòng không được xây dựng để huấn luyện tương tác mô hình học tăng cường PTHL và mục tiêu trên không.*

Kịch bản phòng không được xây dựng sử dụng bộ công cụ mở rộng thư viện OpenAI Gym và ngôn ngữ lập trình Python [14], trong đó, các mô hình PTHL và mô hình mục tiêu trên không sẽ tương tác trực tiếp với nhau để hoàn thành nhiệm vụ chiến đấu của mình. Trong hình 5, các đối tượng được thể hiện gồm: 4 mục tiêu trên không – hình tròn màu xanh; 3 đối tượng cần bảo vệ - hình tròn màu đen; có 3 PTHL – hình vuông màu đỏ và các đường tròn màu cam là phạm vi vùng hỏa lực, đường tròn màu đỏ là phạm vi vùng quan sát của PTHL. Mục tiêu của mô hình mục tiêu trên không là tiêu diệt tất cả các đối tượng phòng thủ, mục tiêu của mô hình PTHL là tiêu diệt toàn bộ mục tiêu trên không trong vùng hỏa lực để bảo vệ đối tượng phòng thủ. Vị trí bố trí các PTHL được khởi tạo ngẫu nhiên và phân bố xung quanh các đối tượng phòng thủ, đảm bảo phù hợp với điều kiện thực tế.

3.3. Kết quả mô phỏng và đánh giá

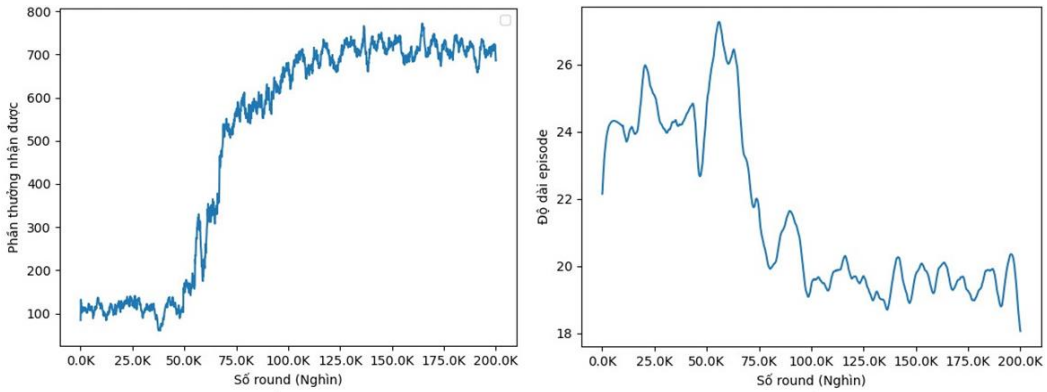
a) Kết quả huấn luyện mô hình học tăng cường

Sau khi xây dựng môi trường huấn luyện, mô hình học tăng cường đa tác nhân của PTHL với thuật toán DQL được huấn luyện trong 200 nghìn vòng (timestep) đã đạt được sự hội tụ giá trị phần thưởng, được trình bày trong hình 6.a. Dựa trên kết quả thu được có thể thấy mô hình PTHL với thuật toán DQL đã học được khả năng nhận thức tình huống, lựa chọn và tiêu diệt mục

tiêu để nhận về phần thưởng tối ưu. Đồng thời, giá trị trung bình của độ dài chu trình huấn luyện giảm dần theo thời gian, được trình bày trong hình 6.b.

Đánh giá kết quả huấn luyện

Trong hình 6.a, ta thấy giá trị phần thưởng trung bình của mô hình PTHL với thuật toán DQL có xu hướng tăng dần chứng tỏ quá trình học tăng cường có hiệu quả, mô hình PTHL được tối ưu và các chiến lược huấn luyện để tiêu diệt mục tiêu và bảo vệ đối tượng phòng thủ đưa ra là đúng đắn. Khi mô hình huấn luyện đến khoảng 200 nghìn vòng, giá trị phần thưởng trung bình của mô hình PTHL có dấu hiệu dao động theo phương ngang, điều này cho thấy mô hình huấn luyện đã đạt đến điểm tối ưu, PTHL đã đạt được khả năng xử lý và nhận về giá trị phần thưởng cao nhất có thể dựa trên môi trường hiện tại và hành động của mục tiêu trên không. Đồng thời, trong hình 6.b giá trị trung bình của độ dài chu trình huấn luyện giảm theo thời gian, điều đó có nghĩa là tác nhân học tăng cường sẽ mất ít thời gian hơn để hoàn thành nhiệm vụ tiêu diệt mục tiêu.

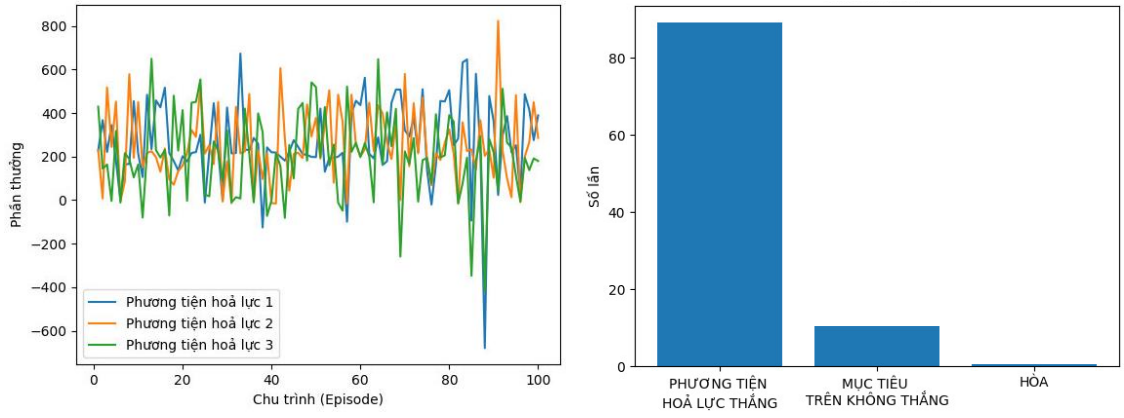


a) *b)*

Hình 6. *a) Sự hội tụ giá trị phần thưởng của mô hình PTHL với thuật toán DQL;*
b) Giá trị trung bình độ dài chu trình huấn luyện theo thời gian.

b) Kết quả mô phỏng tương tác các mô hình sau khi được huấn luyện

Mô hình PTHL và mục tiêu trên không sau khi được huấn luyện, tương tác với nhau qua 100 chu trình (episode) kết quả phần thưởng của mỗi PTHL qua mỗi vòng trong 100 chu trình tương tác, được thể hiện trong hình 7.a, với kết quả tương tác là: mô hình PTHL chiến thắng 89 chu trình, mục tiêu trên không chiến thắng 10 chu trình và 1 chu trình có kết quả hòa, hình 7.b.



a) ***b)***

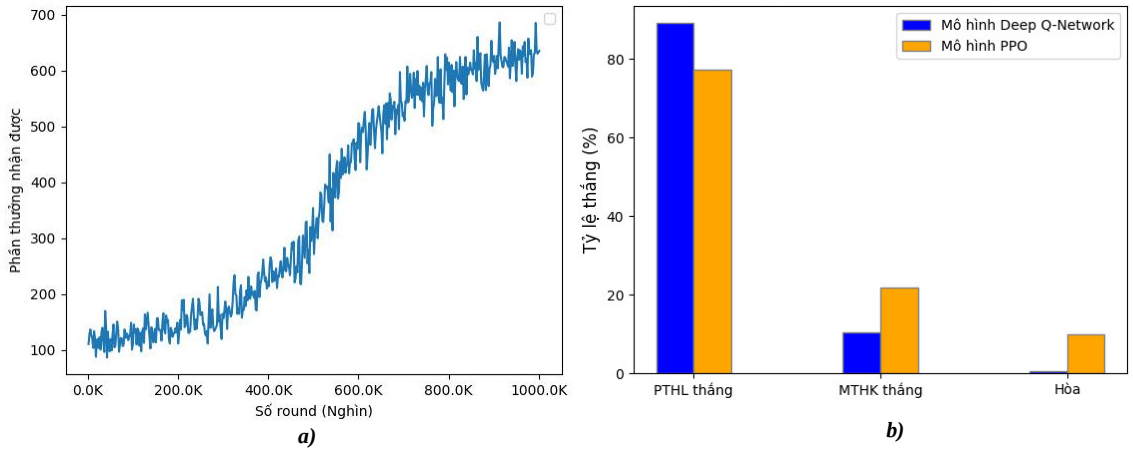
Hình 7. *a) Phần thưởng của mỗi PTHL nhận được qua mỗi vòng trong 100 episode;*
b) Kết quả tác chiến tương tác giữa 2 mô hình học tăng cường sau khi huấn luyện.

Đánh giá kết quả mô phỏng tương tác

Trong hình 7.a, các giá trị phần thưởng trung bình của 3 PTHL được tính lần lượt là: $R_{W_1}=266.07; R_{W_2}=249.33; R_{W_3}=198.96$. Ta thấy rằng, phần thưởng của mỗi PTHL ở mỗi chu trình (episode) đều có khả năng đạt được những giá trị cao hoặc thấp như nhau, phụ thuộc vào yếu tố môi trường được khởi tạo và tương tác giữa hai mô hình. Khi thống kê với số lượng chu trình đủ lớn, mặc dù biên độ giá trị của phần thưởng khá lớn (xấp xỉ từ -600 đến 800) nhưng giá trị trung bình phần thưởng PTHL qua 100 chu trình không có sự chênh lệch lớn, điều này chứng tỏ vai trò của các đối tượng trong mô hình là tương đương nhau và mỗi đối tượng đều đưa ra được hành động hợp lý dựa trên môi trường, mô hình huấn luyện không có hiện tượng "*học lệch*", tối ưu tập trung duy nhất vào tối ưu 1 hoặc 2 đối tượng mà tối ưu hành động của toàn bộ các đối tượng có trong mô hình để đạt được kết quả tác chiến tối ưu nhất.

c) So sánh mô hình PTHL đề xuất với mô hình PTHL sử dụng thuật toán PPO

Để đánh giá so sánh mô hình PTHL với thuật toán DQL đề xuất với mô hình khác, chúng tôi tiến hành huấn luyện mô hình PTHL với thuật toán PPO, trong cùng các điều kiện (môi trường, không gian quan sát, không gian hành động và hàm phần thưởng). Khác với DQL, thuật toán PPO tối ưu hóa chiến lược lựa chọn hành thay vì ước tính giá trị phần thưởng nhận được. Do điều kiện ràng buộc mỗi PTHL chỉ lựa chọn một mục tiêu riêng để tiêu diệt trong mỗi lần thực hiện hành động, không gian hoạt động của PTHL sẽ có giá trị lớn, yêu cầu nhiều mẫu, nên cần thời gian huấn luyện lớn hơn để cập nhật và tối ưu chính sách. Kết quả mô phỏng chứng minh, mô hình PTHL với thuật toán PPO đã tối ưu hóa được chính sách và đạt được hội tụ giá trị phần thưởng qua 900 nghìn vòng huấn luyện, xem hình 8.a. Sau khi được huấn luyện, chúng tôi tiến hành mô phỏng tương tác mô hình PTHL với thuật toán DQL và mô hình PTHL với thuật toán PPO qua 1000 chu trình (episode) để đánh giá kết quả khi tác chiến với mô hình mục tiêu trên không, kết quả tương tác được thể hiện trong hình 8.b. Có thể thấy, tỉ lệ chiến thắng của mô hình PTHL với DQL đạt 89,1% lớn hơn nhiều so với mức 77,2% của PPO, điều này chứng minh mô hình DQL đề xuất có khả năng tác chiến tốt hơn so với mô hình PPO sau khi được huấn luyện.



Hình 8. *a) Giá trị trung bình phần thưởng của mô hình PTHL sử dụng thuật toán PPO; b) Kết quả mô phỏng tương tác của mô hình PTHL sử dụng thuật toán DQL và PPO.*

4. KẾT LUẬN

Bài báo trình bày phương pháp học tăng cường sâu đa tác nhân giải bài toán lựa chọn PTHL động, trong đó, các PTHL phòng không được xây dựng và huấn luyện trên bộ thư viện OpenAI Gym, tương tác trực tiếp với mô hình mục tiêu trên không đã được huấn luyện trước đó, trong môi trường tác chiến động. Thuật toán học sâu Q (DQL) được sử dụng để tối ưu hóa giá trị hàm

phần thưởng, bằng kết quả huấn luyện chứng minh, sau 200 nghìn vòng huấn luyện, giá trị phần thưởng trung bình của mô hình PTHL đạt được sự hội tụ (tối ưu), PTHL đã đạt được khả năng xử lý và nhận về giá trị phần thưởng cao nhất có thể dựa trên môi trường hiện tại và hành động của mục tiêu trên không. Đồng thời, giá trị trung bình của độ dài chu trình huấn luyện giảm theo thời gian, điều đó có nghĩa là tác nhân học tăng cường sẽ mất ít thời gian hơn để hoàn thành nhiệm vụ tiêu diệt mục tiêu. Mô hình PTHL với thuật toán DQL đề xuất được so sánh với mô hình PTHL với thuật toán PPO trong cùng điều kiện huấn luyện chứng minh mô hình DQL có khả năng học nhanh hơn. Sau khi được huấn luyện, tiến hành thử nghiệm tương tác hai mô hình DQL và PPO với mô hình mục tiêu trên không, sau 1000 chu trình, mô hình PTHL với DQL đề xuất có tỉ lệ chiến thắng đạt 89,1% lớn hơn nhiều so với mô hình sử dụng thuật toán PPO đạt 77,2%. Như vậy, mô hình PTHL sử dụng thuật toán DQL sau khi huấn luyện có khả năng tự động nhận thức tình huống, để xây dựng phương án tương tác đối kháng động cùng với các PTHL khác trong hệ thống và chọn ra phương án tối ưu có tính tới các ràng buộc thực tế, chứng minh khả năng ứng dụng mô hình trong phát triển các mô đun phần mềm hỗ trợ tác chiến trong các hệ thống TĐH CH-ĐK phòng không thời gian thực.

TÀI LIỆU THAM KHẢO

[1]. Truong, N.X., Phuong, P.K., Phuc, H.V., Tien, V.H., *"Q-Learning Based Multiple Agent Reinforcement Learning Model for Air Target Threat Assessment,"* in The International Conference on Intelligent Systems & Networks, (2023), https://doi.org/10.1007/978-981-99-4725-6_16.

[2]. Lloyd Hammond, *"Application of a Dynamic Programming Algorithm for Weapon Target Assignment"*, Edinburgh South Australia: Defence Science and Technology Group, (2016).

[3]. Mohammad Babul Hasan and Yaindrila Barua, *"Weapon Target Assignment"*, DOI: 10.5772/intechopen.93665, (2020).

[4]. Fredrik Johansson, Göran Falkman, *"SWARD: System for weapon allocation research & development,"* in Information Fusion (FUSION), DOI:10.1109/ICIF.2010.5712067.

[5]. Yiping Lu, Danny Z. Chen, *"A new exact algorithm for the Weapon-Target Assignment problem,"* Elsevier Ltd, vol. Omega 98,102138, (2021), https://doi.org/10.1016/j.omega.2019.102138, 2019.

[6]. Yang Zhao, Yifei Chen, Ziyang Zhen and Ju Jiang, *"Multi-weapon multi-target assignment based on hybrid genetic algorithm in uncertain environment,"* International Journal of Advanced Robotic Systems, no. https://doi.org/10.1177/1729881420905922, (2020).

[7]. Elias Munapo, *"Development of an accelerating hungarian method for assignment problems,"* Eastern-European Journal of Enterprise Technologies, pp. 6-13, (2020).

[8]. Yuan Zeng Cheng,.. *"Weapon Target Assignment Problem Solving Based on Hungarian Algorithm,"* Applied Mechanics and Materials, doi:10.4028/www.scientific.net/AMM.713-715.2041, (2015).

[9]. Hildegarde Mouton, Jan Roodt, Herman Le Roux, *"Applying Reinforcement Learning to the Weapon Assignment Problem in Air Defence,"* Journal of Military Studies, vol. 39 No. 2, (2011), DOI: https://doi.org/10.5787/39-2-115.

[10]. Tong Wang, Liyue Fu, Zhengxian Wei, *"Unmanned ground weapon target assignment based on deep Q-learning network with an improved multi-objective artificial bee colony algorithm,"* Engineering Applications of Artificial Intelligence. https://doi.org/10.1016/j.engappai.2022.105612, (2023).

[11]. Brian Gaudet, Kristofer Drozd, *"Deep Reinforcement Learning for Weapons to Targets Assignment in a Hypersonic strike,"* University of Arizona. doi:10.13140/RG.2.2.19047.62881, (2023).

[12]. Yuxi Li, *"Deep Reinforcement Learning: An Overview,"* https://arxiv.org/abs/1701.07274, (2018).

[13]. Greg Brockman, Vicki Cheung, Ludwig Pettersson, *"OpenAI Gym,"* https://arxiv.org/pdf/1606.01540.pdf, (2016).

[14]. John Schulman, Filip Wolski, Prafulla Dhariwal, *"Proximal Policy Optimization Algorithms,"* OpenAI, no. https://arxiv.org/pdf/1707.06347.pdf, pp. 1-12, (2017).

ABSTRACT

Application of multi-agent deep reinforcement learning method to solve the dynamic weapon target assignment problem

This paper presents the Multi-Agent Deep Reinforcement Learning method to solve the dynamic weapon target assignment (DTWA) in the air defense command and control system. The weapon model is built based on predicting the optimal trajectory of air targets and the status of objects on the ground, as well as the optimal plan to coordinate the activities of weapons in the system. The weapon model is built on the OpenAI Gym library, describes the rules of the dynamic air defense combat environment and uses deep reinforcement learning algorithms (Deep Q-Leanring) to optimize the policy. Experimental simulation results with different air defense scenarios demonstrate that, after being trained, the deep reinforcement learning model of the air defense weapon has the ability to automatically analyze, perceive situations, and coordinate with other air defense weapons in the system, build a dynamic resistance interaction plan and select the optimal plan taking into account practical constraints so that the overall loss function has a minimum value for the entire combat process. Therefore, the reinforcement learning model has the ability to be applied to develop software modules to support decision-making in the air defense command and control system.

Keywords: Reinforcement Learning; Command and control system; C4I; DWTA; DQL; OpenAI Gym.