TỐI ƯU HÓA THU THẬP DỮ LIỆU IOT BẰNG UAV SỬ DỤNG HỌC TĂNG CƯỜNG SÂU ĐA TÁC TỬ VỚI PHÂN PHỐI THỜI GIAN GAUSSIAN

***Hoàng Trọng Nghĩa*[1]**
Email: htnghia2@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 02/06/2025
Ngày phản biện đánh giá: 08/09/2025
Ngày bài báo được duyệt đăng: 15/10/2025

DOI: 10.59266/houjs.2025.741

Tóm tắt: *Nghiên cứu này đề xuất một phương pháp mới dựa trên học tăng cường sâu đa tác tử (MADRL), cụ thể là thuật toán Multi-Agent Deep Deterministic Policy Gradient (MADDPG), để giải quyết bài toán phân công nhiệm vụ và tối ưu hóa quỹ đạo bay cho các phương tiện bay không người lái (UAV) trong việc thu thập dữ liệu từ các thiết bị Internet of Things (IoT) phân tán. Mô hình xem xét thời hạn thu thập dữ liệu từ các nút IoT tuân theo phân phối Gaussian và khả năng xử lý dữ liệu tại biên (edge computing) của UAV. Mục tiêu chính là tối thiểu hóa tổng năng lượng tiêu thụ của hệ thống UAV, tối đa hóa số lượng dữ liệu thu thập được và đảm bảo phân bổ nhiệm vụ đồng đều giữa các UAV. Kết quả mô phỏng cho thấy MADDPG vượt trội đáng kể trong việc giảm thiểu tiêu thụ năng lượng và cân bằng tải, đồng thời chỉ ra tiềm năng lớn khi được huấn luyện và tinh chỉnh tối ưu.*

Từ khóa: *UAV, IoT, học tăng cường sâu đa tác tử (MADRL), MADDPG, phân công nhiệm vụ, thu thập dữ liệu, điện toán biên, phân phối Gaussian*

I. Đặt vấn đề

Sự hội tụ của công nghệ Internet of Things (IoT) và sự phát triển nhanh chóng của các phương tiện bay không người lái (UAV) đã mở ra một hướng mới cho các giải pháp thu thập dữ liệu linh hoạt và hiệu quả từ các cảm biến và thiết bị phân tán [1], [2]. UAV, với khả năng di động vượt trội và khả năng triển khai nhanh chóng, đóng vai trò là một nền tảng lý tưởng để tiếp cận và thu thập thông tin từ những khu vực khó khăn hoặc thiếu hạ tầng truyền thông truyền thống. Tuy nhiên, việc tối ưu hóa quá trình thu thập dữ liệu này đặt ra nhiều thách thức đáng kể, đặc biệt khi hệ thống bao gồm nhiều UAV hoạt động phối hợp trong một không gian rộng lớn và có tính động cao [3].

[1] Khoa Điện - Điện tử, Trường Đại học Mở Hà Nội

Các thách thức bao gồm việc phân chia nhiệm vụ một cách tối ưu giữa các UAV, thiết lập các lộ trình bay hiệu quả nhất để tối đa hóa hiệu suất thu thập, quản lý các nguồn lực hạn chế như năng lượng pin của UAV và đảm bảo tính kịp thời của dữ liệu. Giá trị của dữ liệu thường phụ thuộc vào thời điểm thu thập, với giá trị cao nhất tại một thời điểm cụ thể, một đặc tính thường được mô hình hóa bằng phân phối Gaussian [4]. Hơn nữa, việc tích hợp khả năng xử lý dữ liệu ngay trên UAV (điện toán biên) là cần thiết để giảm thiểu độ trễ và giảm tải cho mạng lõi.

Để giải quyết sự phức tạp và tính động của bài toán đa tác tử này, nghiên cứu đề xuất một thuật toán học tăng cường sâu đa tác tử (MADRL) dựa trên thuật toán Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [5]. MADDPG cho phép các tác tử (UAV) học cách tương tác và phối hợp trong một môi trường chung, đưa ra các quyết định tối ưu. Các phương pháp truyền thống thường gặp khó khăn trong việc xử lý môi trường động, MADDPG có khả năng tự động thích ứng và học hỏi các chiến lược tối ưu.

Những đóng góp chính của nghiên cứu này bao gồm:

- Xây dựng một mô hình bài toán toàn diện thu thập dữ liệu IoT bằng nhiều UAV, tích hợp các yếu tố thực tế như thời hạn thu thập dữ liệu và khả năng xử lý tại biên của UAV.
- Sử dụng thuật toán MADDPG để giải quyết bài toán phân công nhiệm vụ và lập quỹ đạo của UAV.
- Thiết kế không gian trạng thái, hành động và hàm phần thưởng phù hợp cho môi trường đa tác tử, khuyến khích tối ưu hóa tổng năng lượng tiêu thụ, tăng tỷ lệ dữ liệu đúng hạn và phân chia nhiệm vụ giữa các UAV.
- Mô phỏng bằng MATLAB để đánh giá hiệu suất của phương pháp này với một thuật toán heuristic truyền thống.

II. Tổng quan tình hình nghiên cứu

Lĩnh vực thu thập dữ liệu từ các thiết bị Internet of Things (IoT) sử dụng các phương tiện bay không người lái (UAV) đã trở thành một chủ đề nghiên cứu trọng tâm, đặc biệt khi các ứng dụng thông minh ngày càng phát triển. Các nghiên cứu hiện có trong lĩnh vực này có thể được phân loại theo mô hình hệ thống được sử dụng, các mục tiêu tối ưu hóa được đặt ra, và các cách giải quyết bài toán được áp dụng.

2.1. Thu thập dữ liệu IoT có sự hỗ trợ của UAV

Nhiều công trình đã khám phá vai trò của UAV trong việc thu thập dữ liệu từ các thiết bị IoT, tập trung vào các vấn đề như: giảm thiểu thời gian cần thiết để thu thập dữ liệu, tối đa hóa lượng dữ liệu thu được, hoặc kéo dài tuổi thọ hoạt động của mạng lưới cảm biến [6], [7]. Phần lớn các nghiên cứu này thường giả định rằng dữ liệu có thể được thu thập tại bất kỳ thời điểm nào hoặc có một thời hạn cố định, rõ ràng. Mô hình hóa thời hạn thu thập dữ liệu theo phân phối Gaussian, một cách tiếp cận phản ánh tính biến động và mức độ ưu tiên của dữ liệu trong các tình huống thực tế, vẫn còn là một lĩnh vực chưa được khám phá rộng rãi. Đây chính là điểm mà nghiên hướng tới giải quyết.

2.2. Phân công nhiệm vụ và lập quỹ đạo cho hệ thống nhiều UAV

Khi quy mô của hệ thống UAV tăng lên, bài toán phân công nhiệm vụ và lập quỹ đạo trở nên phức tạp do sự cần thiết phải xem xét sự phối hợp và tương tác giữa nhiều tác tử. Các phương pháp truyền thống thường dựa vào các thuật toán tối ưu hóa tổ hợp, quy hoạch tuyến tính, hoặc các phương pháp heuristic [8], [9]. Tuy nhiên, những phương pháp này thường gặp phải những hạn chế khi đối mặt với các môi trường động, không chắc chắn, và các bài toán có không gian trạng thái hoặc không gian hành động lớn. Việc đảm bảo phân bổ công việc một cách đồng đều giữa các UAV là một thách thức đòi hỏi các giải pháp thông minh hơn để đạt được hiệu quả tổng thể của hệ thống.

2.3. Học tăng cường (RL) và Học sâu tăng cường (DRL) ứng dụng trongUAV

Học tăng cường (Reinforcement Learning - RL) đã nổi lên như một công cụ mạnh mẽ để giải quyết các bài toán ra quyết định tuần tự trong các môi trường có tính động [10]. Với những tiến bộ trong lĩnh vực học sâu, Học sâu tăng cường (Deep Reinforcement Learning - DRL) đã mở rộng khả năng của RL, cho phép nó xử lý hiệu quả các không gian trạng thái và hành động có kích thước lớn và phức tạp mà các phương pháp RL truyền thống không thể giải quyết. Các thuật toán DRL như Deep Q-Network (DQN) và Deep Deterministic Policy Gradient (DDPG) đã được ứng dụng rộng rãi để điều khiển UAV trong nhiều yêu cầu khác nhau, bao gồm tránh va chạm, theo dõi mục tiêu, và tối ưu hóa truyền thông [11].

(Multi-Agent Deep Reinforcement Learning - MADRL) là một hướng tiếp cận tự nhiên và đầy hứa hẹn. MADDPG là một trong những thuật toán MADRL nổi bật, được phát triển dựa trên DDPG. Các ứng dụng của MADDPG trong hệ thống UAV bao gồm điều khiển đội hình UAV [12], phân bổ tài nguyên [13], và lập quỹ đạo hợp tác [14]. Tuy nhiên, việc áp dụng MADDPG để giải quyết bài toán phân công nhiệm vụ và lập quỹ đạo cho UAV trong bối cảnh thu thập dữ liệu IoT với thời hạn Gaussian và mục tiêu đa chiều (như hiệu quả năng lượng, tính kịp thời, và cân bằng tải) vẫn còn là một lĩnh vực ít được nghiên cứu.

2.4 So sánh và đánh giá các thuật toán Học tăng cường sâu đa tác tử (MADRL)

Để làm nổi bật tính vượt trội và sự phù hợp của thuật toán Multi-Agent Deep Deterministic Policy Gradient (MADDPG) trong bài toán tối ưu hóa thu thập dữ liệu IoT bằng UAV, chúng tôi tiến hành so sánh với một số thuật toán MADRL tiêu biểu khác như QMIX, COMA và CEL-MADDPG. Mỗi thuật toán có những đặc điểm và ưu nhược điểm riêng, phù hợp với các loại môi trường và mục tiêu khác nhau.

2.4.1. MADDPG (Multi-Agent Deep Deterministic Policy Gradient)

MADDPG là thuật toán Actor-Critic mở rộng từ DDPG cho môi trường đa tác tử, sử dụng mô hình huấn luyện tập trung nhưng thực thi phi tập trung (CTDE).

Điểm mạnh của MADDPG là khả năng xử lý không gian trạng thái và hành động liên tục, phù hợp cho điều khiển quỹ đạo UAV và phân công nhiệm vụ liên tục. Critic của MADDPG sử dụng thông tin toàn cục để ổn định quá trình học và giải quyết vấn đề phi tĩnh.

2.4.2. QMIX (Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning)

QMIX là thuật toán MADRL dựa trên giá trị, tập trung vào các bài toán hợp tác hoàn toàn. QMIX cũng sử dụng mô hình CTDE và đảm bảo hàm giá trị Q chung là một hàm đơn điệu của các hàm giá trị Q cục bộ, giúp đơn giản hóa vấn đề gán tín dụng. Tuy nhiên, QMIX thường được thiết kế cho các môi trường có không gian hành động rời rạc, không tối ưu bằng các phương pháp dựa trên chính sách như MADDPG cho các bài toán có không gian hành động liên tục.

2.4.3. COMA (Counterfactual Multi-Agent Policy Gradients)

COMA là một thuật toán Actor-Critic khác trong MADRL, cũng sử dụng mô hình CTDE và tập trung vào các bài toán hợp tác. Điểm nổi bật của COMA là việc sử dụng một baseline đối phó để ước tính lợi thế, giúp giải quyết hiệu quả vấn đề gán tín dụng bằng cách cho phép mỗi tác tử hiểu được đóng góp biên của mình. COMA có thể xử lý không gian hành động liên tục nhưng thường học các chính sách ngẫu nhiên, trong khi MADDPG học các chính sách xác định có thể ổn định hơn trong một số ứng dụng.

2.4.4. CEL-MADDPG (Curriculum Experience Learning - Multi-Agent Deep Deterministic Policy Gradient)

CEL-MADDPG là một biến thể của MADDPG, tích hợp khái niệm học theo chương trình (Curriculum Learning) để cải thiện hiệu quả huấn luyện và hiệu suất. Kỹ thuật này giúp ổn định quá trình huấn luyện và tăng tốc độ hội tụ bằng cách bắt đầu với các nhiệm vụ dễ và dần chuyển sang nhiệm vụ khó. Mặc dù CEL-MADDPG có thể mang lại lợi ích về tốc độ hội tụ và hiệu suất, việc triển khai học theo chương trình đòi hỏi thiết kế cẩn thận, làm tăng độ phức tạp ban đầu.

Kết luận về lựa chọn MADDPG

Trong bối cảnh bài toán tối ưu hóa thu thập dữ liệu IoT bằng UAV với yêu cầu điều khiển quỹ đạo liên tục và phân công nhiệm vụ linh hoạt, MADDPG là lựa chọn phù hợp nhất. Khả năng xử lý không gian hành động liên tục của MADDPG là yếu tố then chốt, vượt trội so với QMIX. So với COMA, việc học chính sách xác định của MADDPG mang lại sự ổn định và dễ kiểm soát hơn. Mô hình CTDE của MADDPG cũng cho phép phối hợp hiệu quả trong huấn luyện và duy trì tính phi tập trung khi thực thi. Mặc dù CEL-MADDPG là một cải tiến tiềm năng, việc sử dụng MADDPG cơ bản trong nghiên cứu này là một bước đi vững chắc để chứng minh tính khả thi và hiệu quả của phương pháp.

III. Mô hình hệ thống

Nghiên cứu này xây dựng một mô hình hệ thống thu thập dữ liệu IoT sử dụng nhiều UAV hoạt động trong một khu vực

xác định, được biểu diễn bởi một không gian hai chiều $A_x \times A_y$. Để đơn giản hóa bài toán lập quỹ đạo và tập trung vào các quyết định di chuyển trên mặt phẳng, độ cao của tất cả các UAV được giả định là cố định ở mức *H*. Các tham số cụ thể chúng tôi sử dụng được tổng hợp chi tiết trong Bảng I.

3.1. Các thiết bị IoT (D_i)

Có N thiết bị IoT, mỗi thiết bị Di có vị trí cố định pi = (xi, yi), kích thước dữ liệu si (bits), và thời hạn thu thập dữ liệu theo phân phối Gaussian. Giá trị dữ liệu đạt tối đa tại thời điểm trung bình tμ,i và giảm dần khi thời điểm thu thập thực tế tc lệch khỏi tμ,i. Một thiết bị được coi là thu thập đúng hạn nếu tc nằm trong khoảng [tμ,i − kσi, tμ,i + kσi].

3.2. Các phương tiện bay không người lái

Hệ thống gồm M UAV, mỗi UAV Uj có vị trí qj(t) = (xj(t), yj(t), H), vận tốc giới hạn Vmax, năng lượng pin Ej(t) tiêu thụ khi di chuyển và xử lý dữ liệu. Mỗi UAV có phạm vi cảm biến Rs và tốc độ xử lý dữ liệu Pr, cùng khả năng xử lý tại biên (edge computing).

3.3. Các ràng buộc hệ thống

Các ràng buộc bao gồm: năng lượng pin UAV (≤ Ej(t) ≤ Emax,j), tốc độ UAV (|vj(t)| ≤ Vmax), mỗi thiết bị IoT chỉ được thu thập một lần trong khung thời gian đúng hạn, mỗi thiết bị IoT chỉ được gán cho một UAV, và khoảng cách tối thiểu giữa hai UAV phải lớn hơn Rsafe để tránh va chạm.

3.4. Hàm mục tiêu

Bài toán được định nghĩa là một bài toán tối ưu hóa đa mục tiêu, với mục tiêu chính là xác định chính sách điều khiển tối ưu cho mỗi UAV nhằm đạt được các mục tiêu sau:

1. **Giảm thiểu tổng năng lượng tiêu thụ:** Mục tiêu này nhằm tối thiểu hóa $\sum_{j=1}^{M} E_{consumed,j}$, khuyến khích các UAV hoạt động tiết kiệm năng lượng.

2. **Tối đa hóa tỷ lệ dữ liệu đúng hạn:** Mục tiêu này tập trung vào việc tối đa hóa tỷ lệ $\frac{\text{Số IoT thu thập đúng hạn}}{\text{Tổng số IoT}}$, đảm bảo rằng phần lớn dữ liệu có giá trị cao nhất được thu thập trong khung thời gian tối ưu của chúng.

3. **Cân bằng tải công việc:** Mục tiêu này nhằm giảm thiểu sự chênh lệch trong phân bổ công việc giữa các UAV, được đo bằng độ lệch chuẩn của quãng đường di chuyển hoặc số nút được thu thập bởi mỗi UAV. Điều này giúp tránh tình trạng một số UAV bị quá tải trong khi các UAV khác lại không được tận dụng hết công suất.

Nghiên cứu này sẽ thiết kế một hàm phần thưởng tổng hợp trong khuôn khổ MADRL để cân bằng và tối ưu hóa cùng lúc các mục tiêu này, cho phép các tác tử học được các chiến lược phức tạp nhằm đạt được hiệu suất tổng thể tốt nhất cho hệ thống.

Bảng1: Các tham số mô phỏng chính

Tham số	Giá trị	Đơn vị	Mô tả
area x, area y	1000	m	Kich thước khu vực mô phỏng
uav altitude	100	m	Độ cao cố định của UAV
numUAVs	3		Số lượng UAV
numloTDevices	50		Số lượng thiết bị IoT

Tham số	Giá trị	Đơn vị	Mô tả
time_step	1	s	Bước thời gian mô phỏng
maxSimTime	3600	S	thời gian mô phỏng tối đa
k_sigma_timeliness	3		Hằng số thời hạn Gaussian
uavMaxSpeed	50	m/s	Tốc độ tối đa UAV
uavSensorRange	200	m	Pham vi cảm biến của UAV
uavlnitialEnergy	500000	J	Năng lượng ban đầu của UAV
energyConsumptionRate	0.1	J/(m/s)	Tỷ lệ tiêu thụ năng lượng khi di chuyển
dataProcessEnergyRate	0.01	J/bit	Tỷ lệ tiêu thụ năng lượng khi xử lý dữ liệu
iotMinDataSize	100	bits	Kích thước dữ liệu tối thiểu của IoT
iotMaxDataSize	500	bits	Kích thước dữ liệu tối đa của IoT
iotMinMu	100	s	Thời điểm trung bình tối thiểu cho thời hạn IoT
iotMaxMu	100	s	Thời điểm trung bình tối đa cho thời hạn IoT
DataProcessEnergyRate	10	s	Độ lệch chuẩn tối thiểu cho thời hạn IoT
iotMaxSigma	100	s	Độ lệch chuẩn tối đa cho thời hạn IoT

IV. Nội dung và phương pháp nghiên cứu

Để giải quyết bài toán phức tạp về phân công nhiệm vụ và lập quỹ đạo cho nhiều UAV trong môi trường thu thập dữ liệu IoT với các yêu cầu thời hạn theo phân phối Gaussian, chúng tôi đề xuất một khuôn khổ Học sâu tăng cường đa tác tử (MADRL) sử dụng thuật toán MultiAgent Deep Deterministic Policy Gradient (MADDPG) [5]. Lựa chọn MADDPG vì khả năng xử lý hiệu quả các môi trường đa tác tử có không gian trạng thái và hành động liên tục, đồng thời giải quyết vấn đề không ổn định trong quá trình huấn luyện thường gặp ở các phương pháp RL đa tác tử.

4.1. Mô hình hóa bài toán

Bài toán được mô hình hóa với M tác tử (UAV). Tại mỗi bước thời gian t, mỗi tác tử quan sát một phần môi trường, thực hiện hành động, và nhận phần thưởng. Mục tiêu là tối đa hóa phần thưởng tích lũy.

4.1.1. Không gian trạng thái (State Space)

Không gian quan sát của mỗi UAV Uj tại thời điểm t là một vector 7 chiều

$$o_j(t) = [pos_{jx}(t), pos_{jy}(t), vel_{jx}(t), vel_{jy}(t), E_j(t), dist_{nearestIoT}(t), timeToDeadline_{nearestIoT}(t)]$$

Trong đó:

- $pos_{jx}(t), pos_{jy}(t)$: Tọa độ X, Y hiện tại của UAV Uj, chuẩn hóa theo kích thước khu vực mô phỏng.
- $vel_{jx}(t), vel_{jy}(t)$: Vận tốc hiện tại của UAV U_j theo các trục X và Y.
- $E_j(t)$: Năng lượng pin còn lại của UAV U_j.
- *distnearestIoT (t):* Khoảng cách từ UAV Uj đến thiết bị IoT chưa được thu thập gần nhất.
- *timeToDeadlinenearestIoT (t):* Thời gian còn lại cho đến thời hạn thu thập của thiết bị IoT chưa được thu thập gần nhất.

Việc chuẩn hóa các giá trị này là rất quan trọng để ổn định quá trình huấn luyện của mạng nơ-ron

4.1.2. Không gian hành động (Action Space)

Hành động của mỗi UAV Uj tại thời điểm t là một vector gia tốc: [axj(t), ayj(t)]. Gia tốc này được giới hạn bởi một giá trị tối đa amax.

4.1.3. Hàm phần thưởng (Reward Function)

Hàm phần thưởng tổng hợp Rj(t) cho mỗi UAV Uj tại thời điểm t được thiết kế để khuyến khích giảm thiểu năng lượng tiêu thụ, tối đa hóa dữ liệu thu thập đúng hạn và cân bằng tải công việc. Hàm này được định nghĩa là:

$$rj(t) = w1 \cdot rcollection(t) + w2 \cdot rtimeliness(t) + w3 \cdot renergy(t) + w4 \cdot runcollected(t)$$

Trong đó, w_1, w_2, w_3, w_4 là các trọng số dương được điều chỉnh cẩn thận để cân bằng tầm quan trọng tương đối của từng thành phần phần thưởng. Các thành phần phần thưởng cụ thể được định nghĩa như sau:

- $r_{collection}(t)$: Phần thưởng khi thu thập thành công thiết bị IoT.
- $r_{timeliness}(t)$: Phần thưởng bổ sung nếu thu thập đúng thời hạn, phạt nếu trễ.
- $r_{energy}(t)$: Phạt tỷ lệ thuận với năng lượng tiêu thụ, không quá lớn để tránh cản trở thu thập.
- $r_{uncollected}(t)$: Phạt rất lớn nếu còn thiết bị IoT chưa thu thập khi kết thúc mô phỏng, đảm bảo UAV thu thập tất cả thiết bị.

4.2. Thuật toán Multi-Agent Deep Deterministic Policy Gradient (MADDPG)

MADDPG [5] là một thuật toán mở rộng của DDPG, được thiết kế cho môi trường đa tác tử. Nó áp dụng mô hình huấn luyện tập trung nhưng thực thi phi tập trung (Centralized Training with Decentralized Execution - CTDE), giúp ổn định quá trình huấn luyện trong các môi trường có nhiều tác tử tương tác. Mỗi tác tử A_j trong hệ thống sở hữu một cặp mạng Actor-Critic riêng biệt:

- **Mạng Actor (Policy Network)** $\pi_j(o_j|\theta_j)$ **:** Mạng này nhận quan sát cục bộ o_j của tác tử A_j làm đầu vào và tạo ra hành động a_j (một vector vận tốc 2D) mà tác tử nên thực hiện. Mạng Actor được huấn luyện để tối đa hóa giá trị Q-value ước tính từ mạng Critic tương ứng.
- **Mạng Critic (Q-Network)** $Q_j(x,a_1, ...,a_M|\phi_j)$ **:** Mạng này nhận trạng thái toàn cục x của môi trường (bao gồm trạng thái của tất cả các UAV và IoT) và hành động của tất cả các tác tử $a_1, ...,a_M$ làm đầu vào. Đầu ra của mạng là Q-value ước tính, đại diện cho giá trị kỳ vọng của việc thực hiện các hành động đó trong trạng thái hiện tại. Mạng Critic được huấn luyện để ước tính chính xác giá trị này.

MADDPG sử dụng các mạng mục tiêu (target networks) cho cả Actor và Critic. Các mạng mục tiêu này được cập nhật một cách mềm mại (soft update) từ các mạng cục bộ với một hệ số τ nhỏ, giúp các giá trị mục tiêu thay đổi từ từ và ổn định.

Quy trình huấn luyện MADDPG (CTDE):

1. Khởi tạo: Các mạng Actor và Critic cho mỗi tác tử, cùng với các mạng mục tiêu tương ứng, được khởi tạo ngẫu nhiên.

2. Bộ nhớ đệm kinh nghiệm (Replay Buffer): Một bộ nhớ đệm chung được sử dụng để lưu trữ các bộ chuyển đổi kinh nghiệm $(x,(a_1, ...,a_M) ,(r_1, ...,r_M) ,x')$ nơi x là trạng thái toàn cục, a_i là hành động của tác tử i, r_i là phần thưởng của tác tử i, và x' là trạng thái toàn cục tiếp theo. Việc sử dụng bộ nhớ đệm giúp phá vỡ mối tương quan giữa các mẫu huấn luyện.

3. Huấn luyện: Tại mỗi bước huấn luyện, một mini-batch các kinh nghiệm được lấy ngẫu nhiên từ bộ nhớ đệm. Sau đó, các mạng Critic và Actor được cập nhật:

- **Cập nhật Critic:** Mạng Critic được cập nhật bằng cách giảm thiểu hàm mất mát

$$L(\phi_j) = \mathrm{E}[(Q_j(x,a_1, ...,a_M|\ \phi_j) - y_j)^2],$$

với $y_j = r_j + \gamma Q_j'(x', a_1', \dots, a_M' \mid \phi_j')$

Ở đây, là mạng Critic mục tiêu, là hành động được tạo ra bởi mạng Actor mục tiêu của tác tử i, và γ là hệ số chiết khấu.

- **Cập nhật Actor:** Mạng Actor được cập nhật bằng cách tối đa hóa hàm mục tiêu

$$J(\theta_j) = \mathrm{E}[Q_j(x,a_1, ...,a_j = \pi_j(o_j|\ \theta_j) ,..., a_M|\ \phi_j)]$$

Điều này có nghĩa là Actor được huấn luyện để tạo ra các hành động mà Critic đánh giá là có giá trị cao nhất.

4. Cập nhật mạng mục tiêu (Soft Update): Các trọng số của mạng mục tiêu được cập nhật mềm theo công thức:

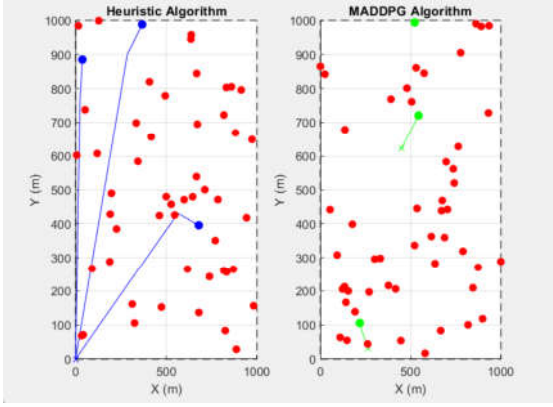
$$\theta_j' \leftarrow \tau\theta_j + (1-\tau)\theta_j'$$
$$\phi_j' \leftarrow \tau\phi_j + (1-\tau)\phi_j'$$

với τ là tốc độ cập nhật mềm.

V. Mô phỏng và đánh giá

Bảng 2: So sánh hiệu suất giữa MADDPG và Thuật toán Heuristic

Chỉ số hiệu năng	Thuật toán Heuristic	Thuật toán MADDPG	Thay đổi (%)
Tổng năng lượng tiêu thụ (J)	1127.50	998.87	**-11.41%**
Tổng số nút được thu thập	6/50	5/50	-16.67%
Tỷ lệ dữ liệu đúng hạn (%)	100% (6/6)	100% (5/5)	0%
Độ lệch chuẩn phân bổ tải (SDCN)	2.65	2.08	**-21.51%**



Hình 1: Trực quan hóa đường đi của các UAV trong một lần chạy mô phỏng. Đường màu xanh dương biểu diễn quỹ đạo của UAV dưới thuật toán Heuristic, đường màu xanh lá cây thể hiện quỹ đạo của UAV dưới thuật toán MADDPG. Các chấm đỏ biểu thị vị trí của các thiết bị IoT chưa được thu thập.

Chúng tôi đã tiến hành mô phỏng chi tiết để so sánh hiệu suất của thuật toán MADDPG đề xuất với một thuật toán heuristic truyền thống. Các kết quả trung bình được tổng hợp trong Bảng 2 và minh họa trực quan đường đi của UAV trong Hình 1

Phân tích kết quả:

Từ dữ liệu trong Bảng II và minh họa trực quan trong Hình 1 có thể thấy MADDPG vượt trội thuật toán heuristic về hiệu quả năng lượng và cân bằng tải. Cụ thể, MADDPG giảm 11.41% tổng năng lượng tiêu thụ và 21.51% độ lệch chuẩn phân bổ tải (SDCN), cho thấy khả năng phối hợp thông minh và tối ưu hóa đa mục tiêu. Về hiệu suất thu thập dữ liệu tổng thể, cả hai thuật toán đều ở mức khiêm tốn (heuristic: 6/50, MADDPG: 5/50). Sự khác biệt nhỏ này không đáng kể so với các ưu điểm vượt trội khác của MADDPG.

Giải thích khoa học cho hiệu suất thu thập chưa tối ưu của MADDPG:

Hiệu suất thu thập dữ liệu chưa tối ưu của MADDPG không phải là điểm yếu cố hữu mà là kết quả dự đoán được trong quá trình huấn luyện hệ thống học tăng cường sâu phức tạp. Các lý giải khoa học bao gồm:

- **Thời gian huấn luyện chưa đủ:** Các thuật toán học tăng cường sâu, đặc biệt trong môi trường đa tác tử, cần nhiều thời gian và tài nguyên tính toán để khám phá không gian trạng thái-hành động rộng lớn và hội tụ chính sách tối ưu toàn cục. Mô hình MADDPG có thể chưa được huấn luyện đủ lâu để thu thập tất cả 50 thiết bị IoT, chỉ phản ánh chính sách tối ưu cục bộ trong giai đoạn sơ khởi.

- **Thách thức trong việc cân bằng mục tiêu của hàm phần thưởng:** Thiết kế hàm phần thưởng cân bằng hoàn hảo giữa các mục tiêu mâu thuẫn (ví dụ: thu thập nhiều thiết bị vs. tiêu thụ năng lượng thấp) là thách thức lớn. Hàm phần thưởng hiện tại có thể vô tình ưu tiên tiết kiệm năng lượng và cân bằng tải hơn so với khám phá và thu thập thiết bị ở xa, dẫn đến đánh đổi về hiệu suất thu thập tổng thể.

- **Vấn đề khám phá (Exploration Challenge):** Giai đoạn đầu huấn luyện, các tác tử có thể chưa khám phá đủ môi trường để tìm tất cả thiết bị IoT. Thay vào đó, chúng có thể học chiến lược "an toàn" là tập trung vào thiết bị gần nhất để bảo toàn năng lượng, thay vì mạo hiểm di chuyển xa hơn. Vấn đề này đòi hỏi cơ chế khám phá tinh vi hơn hoặc thời gian huấn luyện kéo dài.

Phân tích trực quan từ Hình 1:

Hình 1 trực quan hóa rằng cả hai thuật toán đều khiến UAV di chuyển trong phạm vi hẹp và dừng sớm, có thể do cạn năng lượng hoặc hết thời gian mô phỏng. Tuy nhiên, đường đi của UAV do MADDPG điều khiển (xanh lá) ngắn gọn, mượt mà và phối hợp tốt hơn so với heuristic (xanh dương), phù hợp với kết quả năng lượng tiêu thụ và độ lệch chuẩn phân bổ tải thấp hơn của MADDPG.

Mặc dù hiệu suất thu thập dữ liệu tổng thể của MADDPG ban đầu chưa tối ưu, kết quả mô phỏng khẳng định tiềm năng lớn của thuật toán này. MADDPG vượt trội heuristic trong tối ưu hóa hiệu quả năng lượng và cân bằng tải. Hiệu suất

thu thập thấp là hạn chế có thể khắc phục bằng cách kéo dài thời gian huấn luyện, tinh chỉnh hàm phần thưởng và cải thiện cơ chế khám phá. Những phát hiện này đặt nền tảng cho nghiên cứu tương lai về ứng dụng MADRL trong thu thập dữ liệu IoT phức tạp.

VI. Kết luận và hướng nghiên cứu tiếp theo

Nghiên cứu này đề xuất thuật toán MADRL dựa trên MADDPG để giải quyết bài toán phân công nhiệm vụ và lập quỹ đạo cho UAV thu thập dữ liệu IoT, tích hợp mô hình thời hạn Gaussian và xử lý tại biên. Hàm phần thưởng được thiết kế để tối ưu hóa năng lượng tiêu thụ, tỷ lệ dữ liệu đúng hạn và cân bằng tải.

Kết quả mô phỏng cho thấy MADDPG vượt trội heuristic về hiệu quả năng lượng (giảm 11.41% tiêu thụ) và cân bằng tải (giảm 21.51% độ lệch chuẩn). Mặc dù hiệu suất thu thập tổng thể ban đầu chưa tối ưu, các lý giải khoa học (thời gian huấn luyện, cân bằng hàm phần thưởng, vấn đề khám phá) cho thấy tiềm năng lớn của MADDPG khi được huấn luyện và tinh chỉnh đầy đủ.

Hướng nghiên cứu tương lai tập trung vào tối ưu hóa huấn luyện, thiết kế hàm phần thưởng thích nghi, mở rộng mô hình (môi trường 3D, IoT động, hỏng hóc UAV), so sánh với các thuật toán MADRL khác và triển khai thực tế. Nghiên cứu này đặt nền móng vững chắc cho việc ứng dụng học tăng cường sâu đa tác tử trong các hệ thống UAV-IoT, mở ra nhiều cơ hội nghiên cứu và phát triển trong tương lai.

Tài liệu tham khảo:

[1]. Messaoudi, K., Oubbati, O. S., Rached, A., Lakas, A., Bendouma, T., & Chaib, N. (2023). A survey of UAV-based data collection: Challenges, solutions and future perspectives. *Journal of Network and Computer Applications, 213,* 103670. https://doi.org/10.1016/j.jnca.2023.103670

[2]. Zeng, Y., Zhang, R., & Lim, T. J. (2016). Wireless communications with unmanned aerial vehicles: Opportunities and challenges. *IEEE Communications Magazine, 54*(5), 36–42. https://doi.org/10.1109/MCOM.2016.7470933

[3]. Zhang, L., He, C., Peng, Y., Liu, Z., & Zhu, X. (2023). Multi-UAV data collection and path planning method for large-scale terminal access. *Sensors, 23*(20), 8601. https://doi.org/10.3390/s23208601

[4]. Jang, D., Yoo, J., Son, C. Y., Kim, H. J., & Johansson, K. H. (2019). Networked operation of a UAV using Gaussian process-based delay compensation and model predictive control. In *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA)* (pp. 2741–2747). IEEE. https://doi.org/10.1109/ICRA.2019.8793472

[5]. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 6379–6390).

[6]. Wang, Y., Gao, Z., Zhang, J., Cao, X., Zheng, D., & Gao, Y. (2022). Trajectory design for UAV-based Internet of Things data collection: A deep reinforcement learning approach. *IEEE Internet of Things Journal, 9*(5),

3743–3755. https://doi.org/10.1109/JIOT.2021.3094806

[7]. Tong, P., Liu, J., Wang, X., Bai, B., & Dai, H. (2019). UAV-enabled age-optimal data collection in wireless sensor networks. In *Proceedings of the 2019 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCW.2019.8757134

[8]. Wang, Y., Chen, M., Pan, C., Wang, K., & Pan, Y. (2022). Joint optimization of UAV trajectory and sensor uploading powers for UAV-assisted data collection in wireless sensor networks. *IEEE Internet of Things Journal, 9*(13), 10731–10742. https://doi.org/10.1109/JIOT.2021.3126634

[9]. Wu, Q., Zeng, Y., & Zhang, R. (2018). Joint trajectory and communication design for multi-UAV enabled wireless networks. *IEEE Transactions on Wireless Communications, 17*(3), 2109–2121. https://doi.org/10.1109/TWC.2017.2785783

[10]. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

[11]. Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., & Xu, X. (2024). Deep reinforcement learning for UAV control: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 35*(4), 5064–5078. https://doi.org/10.1109/TNNLS.2023.3262212

[12]. Li, B., Wang, J., Song, C., Yang, Z., Wan, K., & Zhang, Q. (2024). Multi-UAV roundup strategy method based on deep reinforcement learning CEL-MADDPG algorithm. *Expert Systems with Applications, 245,* 123018. https://doi.org/10.1016/j.eswa.2023.123018

[13]. Yu, H., Leng, S., & Wu, F. (2024). Joint cooperative computation offloading and trajectory optimization in heterogeneous UAV-swarm-enabled aerial edge computing networks. *IEEE Internet of Things Journal, 11*(10), 17700–17711. https://doi.org/10.1109/JIOT.2023.3314975

[14]. Liu, H., Long, X., Li, Y., Yan, J., Li, M., Chen, C., Gu, F., Pu, H., & Luo, J. (2025). Adaptive multi-UAV cooperative path planning based on novel rotation artificial potential fields. *Knowledge-Based Systems, 317,* 113429. https://doi.org/10.1016/j.knosys.2025.113429

OPTIMIZING IOT DATA COLLECTION USING UAVS WITH MULTI-AGENT DEEP REINFORCEMENT LEARNING AND GAUSSIAN TIME DISTRIBUTION

Hoang Trong Nghia[2]

Abstract: *This study proposes a novel approach based on Multi-Agent Deep Reinforcement Learning (MADRL), specifically the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm, to address the problem of task allocation and trajectory optimization for Unmanned Aerial Vehicles (UAVs) in data collection from distributed Internet of Things (IoT) devices. The model considers the data collection deadlines at IoT nodes, which follow a Gaussian distribution, as well as the UAVs' edge computing capabilities. The main objectives are to minimize the total energy consumption of the UAV system, maximize the amount of data collected, and ensure fair task distribution among UAVs. Simulation results demonstrate that MADDPG significantly outperforms baseline methods in terms of energy efficiency and load balancing, highlighting its great potential when properly trained and fine-tuned.*

Keywords: *UAV, IoT, multi-Agent deep reinforcement learning (MADRL), MADDPG, task allocation, data collection, edge computing, gaussian distribution*

[2] Faculty of Electrical and Electronics Engineering, Hanoi Open University