Tạp chí Khoa học và Công nghệ Trường Đại học Công nghệ Đông Á

Ứng dụng học sâu tăng cường trong tối ưu công suất phát trong mạng di động

Lưu Bách Hưng[1], Lâm Sinh Công[1*]

[1] Khoa Điện tử - Viễn thông, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội

*Email: congls@vnu.edu.vn

Tóm tắt

Sự phát triển nhanh chóng của Internet vạn vật (IoT) đã thúc đẩy xu hướng dày đặc hóa cơ sở hạ tầng mạng, trong đó mạng siêu dày đặc (UDN) trở thành công nghệ then chốt cho 5G. Bài báo này nghiên cứu bài toán tối ưu hóa tài nguyên động nhằm cân bằng hiệu quả năng lượng (EE) và hiệu quả phổ (SE) trong môi trường 5G UDN. Do tính chất động của lưu lượng người dùng, chúng tôi đề xuất mô hình trọng số thích nghi để chuyển đổi bài toán tối ưu đa mục tiêu thành quá trình quyết định Markov (MDP). Để giải quyết vấn đề bùng nổ không gian trạng thái-hành động trong UDN, thuật toán Dueling Deep Q-Network (Dueling DQN) được áp dụng kết hợp với cơ chế truyền thông hợp tác sử dụng kỹ thuật Maximum Ratio Combining (MRC). Cơ chế này xác định 50% người dùng có chất lượng kênh thấp nhất và gán hai SgNB tốt nhất để phục vụ đồng thời, giúp tăng tỷ số tín hiệu trên nhiễu (SNR) cho người dùng yếu. Kết quả mô phỏng với 10 SGNBs và 6 người dùng cho thấy phương pháp đề xuất tiết kiệm công suất 32.5% tại mức tải cao so với phương pháp không hợp tác, đồng thời đảm bảo yêu cầu chất lượng dịch vụ (QoS) cho toàn bộ người dùng.

Từ khóa: Deep Q Network; Mạng di động 5G siêu dày đặc; hiệu quả năng lượng; hiệu suất phổ

Abstract

The rapid development of the Internet of Things (IoT) has driven the trend of densifying network infrastructure, in which ultra-dense networks (UDN) become a key technology for 5G. This paper studies the dynamic resource optimization problem to balance energy efficiency (EE) and spectral efficiency (SE) in the 5G UDN environment. Due to the dynamic nature of user traffic, we propose an adaptive weighting model to transform the multi-objective optimization problem into a Markov decision process (MDP). To address the state-action space explosion issue in UDN, the Dueling Deep Q-

Network (Dueling DQN) algorithm is applied in combination with a cooperative communication mechanism using Maximum Ratio Combining (MRC) technique. This mechanism identifies the 50% of users with the lowest channel quality and assigns the two best SgNBs to serve them simultaneously, helping to increase the signal-to-noise ratio (SNR) for weak users. Simulation results with 10 SgNBs and 6 users show that the proposed method saves 32.5% power at high load compared to the non-cooperative method, while ensuring quality of service (QoS) requirements for all users.

Keywords: Deep Q Network; 5G Ultra Dense Networks; Energy Efficiency; Spectrum Efficiency

1. MỞ ĐẦU

Trong những năm gần đây, với sự phát triển nhanh chóng của Internet di động và công nghệ Internet vạn vật (IoT), nhu cầu về dịch vụ truyền thông di động cũng tăng trưởng nhanh chóng. Các dịch vụ mới nổi như video siêu nét 4K, thực tế ảo, thực tế tăng cường và lái xe tự động đặt ra yêu cầu cao hơn về hiệu suất mạng 5G . Nhu cầu truyền thông của các thiết bị thông minh và môi trường IoT quy mô lớn thúc đẩy sự dày đặc hóa cơ sở hạ tầng mạng. Mạng siêu dày đặc (UDN) đã xuất hiện đúng thời điểm và sẽ trở thành công nghệ then chốt trong mạng 5G [1]. Trong kiến trúc UDN, bán kính cell sẽ được thu nhỏ hơn nữa, dẫn đến tăng nhiễu giữa các cell [2-3]. Hơn nữa, do đặc điểm không đồng nhất của thiết bị người dùng trong không gian và thời gian, việc quản lý tài nguyên của UDN trở nên khó khăn hơn. Do đó, cách phân bổ tài nguyên thích ứng trong UDN là vấn đề đáng được nghiên cứu sâu.

Các chính sách phân bổ tài nguyên trong UDN ảnh hưởng đến hiệu suất mạng và trải nghiệm người dùng. Nhiều công trình đã nghiên cứu có hệ thống về phân bổ tài nguyên trong các kịch bản mạng không đồng nhất (HetNet) và UDN từ các góc độ khác nhau. Trong [4], việc tối đa hóa tổng tốc độ của người dùng tế bào con trong hệ thống NOMA được nghiên cứu. Việc tối ưu hóa hiệu quả năng lượng trong mạng tế bào con dày đặc được nghiên cứu trong [5]. Sự cân bằng giữa hiệu quả năng lượng và hiệu quả phổ được xem xét và đánh giá . Tuy nhiên, các nghiên cứu này chỉ xem xét hiệu suất mạng và yêu cầu người dùng ở trạng thái cố định. Khi môi trường mạng thay đổi, yêu cầu hiệu suất của hệ thống cũng có thể thay đổi. Do đó, phân bổ tài nguyên mạng cần xem xét sự tương tác với môi trường.

Học tăng cường (RL) không cần mô hình có thể được áp dụng để giải quyết bài toán tối ưu hóa ngẫu nhiên trong mạng không dây. Trong môi trường không xác định, RL sẽ thu được

chính sách tối ưu thông qua tương tác với môi trường [6], [7]. Q-learning là một trong những thuật toán RL phổ biến nhất. Tuy nhiên, do sự bùng nổ không gian trạng thái hành động trong các bài toán thực tế, Q-learning hội tụ chậm và khó tìm được hành động tối ưu để giải quyết bài toán. Deep Q Network (DQN) là thuật toán học tăng cường sâu (DRL) mới, có thể kết hợp quá trình RL với một loại mạng nơ-ron gọi là mạng nơ-ron sâu để xấp xỉ hàm giá trị trạng thái-hành động. Do đó, các hạn chế của Q-learning được giải quyết trong DQN. Trong [8], [9], DQN được áp dụng để tối ưu hóa chiến lược BẬT/TẮT của các trạm gốc nhỏ nhằm nâng cao EE trong khi đáp ứng yêu cầu QoS. Trong hệ thống thông tin di động thực tế, có nhiều người dùng có kênh truyền rất thấp như bị che khuất hoặc nằm ranh giới giữa các trạm. Khi đó, việc tăng công suất nên có thể tăng nhiễu giữa các trạm, kéo theo hiệu năng của những người dung này không tăng, hoặc thậm chí giảm. Do đó, việc triển khai kỹ thuật truyền thông hợp tác trong mạng di động được coi là giải pháp phù hợp để nâng cao chất lượng dịch vụ của người dùng [10].

Bài báo đề xuất cơ chế truyền thông hợp tác sử dụng kỹ thuật Maximum Ratio Combining (MRC) để hỗ trợ người dùng có kênh yếu. Cụ thể, hệ thống xác định 50% người dùng có độ lợi kênh trung bình thấp nhất làm Người dung yếu thông qua việc sắp xếp độ lợi kênh trung bình từ tất cả SGNB. Với mỗi người dung yếu, hai SGNB có độ lợi kênh tốt nhất được chọn để đồng thời phục vụ người dung yếu. Kỹ thuật MRC kết hợp tín hiệu từ các SGNB hợp tác, cho phép tăng tín hiệu SNR tổng hợp của người dung yếu. Hàm thưởng được bổ sung điểm thưởng cho kỹ thuật hợp tác dựa trên tỷ lệ phủ sóng người dung yếu và hệ số giảm công suất lý thuyết với bonus tối đa khoảng 15%. Kết quả mô phỏng cho thấy phương pháp Dueling DQN kết hợp truyền thông hợp tác đạt hiệu suất tiêu thụ công suất thấp hơn đáng kể so với phương pháp không hợp tác.

2. MÔ HÌNH HỆ THỐNG

Nghiên cứu xây dựng kịch bản mạng di động không dây trong đó, N trạm con SgNB được phân bố đều trong không gian. Trong đó, vị trí của mỗi trạm là tâm của hình lục giác. Bên cạnh các trạm con, 1 trạm lớn MgNB được triển khai để quản lý tất cả các trạm SgNB. Trong kịch bản này, trạm MgNB được tích hợp 1 tác tử học máy để trách nhiệm thu thập thông tin từ các trạm SgNB, xử lý và đưa ra các quyết định về công suất cũng như phân bổ tài nguyên cho các người dung. Các quyết định này được chuyển về các trạm SgNB để thực thi. Mỗi trạm phục vụ M người dùng, trong đó người dùng phân bố ngẫu nhiên trong vùng phục vụ của trạm đó. Trong mô hình này, người dùng sẽ kết nối trực tiếp với các trạm SgNB mà không kết nối trực tiếp tới MgNB.

Trong mạng truyền thông di động, người dung thường truyền tin một cách rời rạc theo khe thời gian. Do đó, bài báo mô hình hóa quá trình đến và đi của người dùng trong mỗi mạng như hai quá trình ngẫu nhiên độc lập [4]. Trong mỗi khe thời gian, UE đến mỗi tế bào con theo

phân phối Poisson với tham số λ_t. Do đó, xác suất có x SUE mới đến tế bào con trong khoảng thời gian τ là:

$$P(x) = \frac{(\lambda_t \tau)^x}{x!} e^{(-\lambda_t \tau)} \quad (1)$$

Tương tự, SUE rời khỏi tế bào con cũng theo phân phối Poisson với tham số μ_t. Do đó, xác suất có y SUE rời khỏi tế bào con trong khoảng thời gian τ là:

$$P(y) = \frac{(\mu_t \tau)^y}{y!} e^{(-\mu_t \tau)} \quad (2)$$

Trong công trình này, chúng tôi giả định rằng việc kết nối SUE với tế bào con được hoàn thành trước khi phân bổ tài nguyên. Sau đó, chúng tôi có thể định nghĩa tập các thiết bị người dùng được kết nối với SgNB n trong khe thời gian t là $U_n(t) = 1,2,\dots,s_n(t),\dots,S_n(t)$.

Để nâng cao hiệu năng sử dụng phổ, các trạm SgNB được phép tái sử dụng tần số với hệ số tỉ lệ 1. Khi số lượng người dùng đủ lớn và số lượng tài nguyên vô tuyến hữu hạn thì SgNB phải sử dụng tất cả băng tần con để cấp cho người dùng của đó. Bên cạnh đó, kỹ thuật truyền tin không trực giao (NOMA) được áp dụng, cho phép 1 trạm SgNB sử dụng 1 sóng mang con để đồng thời cấp cho nhiều người dùng. Người dùng lúc này được phân biệt với nhau thông qua công suất phát. Do việc sử dụng chung băng tần con, các người dùng có thể gây nhiễu cho nhau. Như vậy, trong kịch bản xấu nhất, tín hiệu đường lên của mỗi người dùng sẽ chịu nhiễu từ tất cả người dùng sử dụng chung băng tần con, trong đó có bao gồm cả người dùng ở trạm lân cận và người dùng trong cùng trạm đó. Giả sử trạm SgNB cấp băng tần con cho người dùng của nó và người dùng này được phép truyền tín hiệu với công suất p_i. Khi đó, tỷ số tín hiệu trên nhiễu cộng tạp âm (SINR) của người dùng n thuộc trạm con SgNB m trên song mang con này được cho bởi:

$$\gamma_{s_n}(t) = \frac{p_n g_n^m}{\Sigma_{i=1,i\neq n}^{M} x_i^m(t) p_i g_i^m(t) + \Sigma_{i=2}^{N} \Sigma_{i=1}^{M} x_n^i(t) p_i g_i^i(t) + \sigma^2} \quad (3)$$

trong đó g_n^m là độ lợi kênh từ SgNB n đến SUE s_n trong trạm SgNB m, p_n là công suất được gán bởi SgNB n cho mỗi SUE, σ là phương sai của nhiễu Gauss trắng cộng (AWGN).

Bên cạnh đó, để đảm bảo hiệu năng của người dùng có kênh truyền yếu kỹ thuật truyền thông hợp tác được sử dụng. Với việc sử dụng kỹ thuật này, người dùng có kênh truyền yếu có thể được kết nối đồng thời với 2 trạm có tín hiệu SgNB mạnh nhất. Tại trạm MgNB, kỹ thuật Maximum Ratio Combining (MRC) được sử dụng phép tổng hợp tín hiệu của người dùng yếu. Khi đó, tín hiệu đường lên của người dùng được tính như sau:

$$h_{combined} = \sqrt{\Sigma_i^K \quad |h_i|^2}$$

trong đó h_i là độ lợi kênh từ SgNB thứ i; K là số trạm hợp tác phục vụ người dùng thứ i.

Theo biểu thức SINR của SUE s_n, tổng thông lượng đường lên trong tế bào con thứ n có thể được tính theo công thức Shannon:

$$R_n(t) = \Sigma_{s_n=1}^{S_n} R_{s_n}(t) = \Sigma_{s_n=1}^{S_n} B_m log_2[1 + \gamma_{s_n}(t)] \quad (4)$$

và tổng thông lượng của toàn bộ hệ thống là:

$$R(t) = \Sigma_{n=1}^{N} R_n(t) = \Sigma_{n=1}^{N} \Sigma_{s_n=1}^{S_n} B_m log_2[1 + \gamma_{s_n}(t)] \quad (5)$$

Trong bài báo này, hiệu quả năng lượng (EE) của hệ thống mạng 5G siêu dày đặc được định nghĩa là tỷ số giữa tổng thông lượng và công suất tiêu thụ của toàn bộ hệ thống. Trong khe thời gian t, EE có thể được biểu diễn như:

$$\eta_E^{(t)} E = R(t)/(p_M + \Sigma_{n=1}^{N} \Sigma_{s=1}^{S} s_n(t) p_n + p_C) \quad (6)$$

trong đó p_M và p_C lần lượt là công suất của MgNB và công suất tiêu thụ được tạo ra bởi tất cả các truyền dẫn mạch.

Hiệu quả phổ (SE) của hệ thống mạng 5G siêu dày đặc được định nghĩa là tỷ số giữa tổng thông lượng và băng thông của toàn bộ hệ thống, trong khe thời gian t được cho bởi:

$$\eta_S^{(t)} E = R(t)/\Sigma_{m=1}^{M} B_m \quad (7)$$

Tối đa hóa SE của hệ thống tương đương với tối đa hóa các tài nguyên khả dụng cho hệ thống. Việc tối đa hóa tài nguyên khả dụng sẽ làm tăng công suất tiêu thụ của hệ thống, có thể dẫn đến giảm EE của hệ thống. Ngược lại, tối đa hóa EE có thể dẫn đến giảm SE. Do đó, không thể đáp ứng yêu cầu hiệu suất hệ thống chỉ bằng cách xem xét tối đa hóa SE hoặc EE. Cần phải xem xét sự đánh đổi giữa SE và EE. Giải quyết sự đánh đổi giữa SE và EE là một bài toán tối ưu đa mục tiêu.

Tuy nhiên, trong các giai đoạn cao điểm và thấp điểm, người dùng có yêu cầu khác nhau về SE và EE trong hệ thống. Do đó, chúng tôi sử dụng phương pháp tổng trọng số động để chuyển đổi MOOP của sự đánh đổi giữa SE và EE thành SOOP. Trọng số động được đưa ra theo yêu cầu của người dùng về SE và EE. Càng nhiều người dùng, tỷ lệ SE càng lớn. Số lượng người dùng càng nhỏ, EE càng trở nên quan trọng. Do đó, trọng số giữa SE và EE trong khe thời gian t có thể được biểu diễn như:

$$\xi^{(t)} = \frac{\Sigma_{n=1}^{N} U_n(t)}{M \times N} \quad (8)$$

Khi đó sự đánh đổi giữa SE và EE có thể được định nghĩa như:

$$max_x \Sigma_{t=1}^{T} \eta(t) = \Sigma_{t=1}^{T} \left[(1 - \xi^{(t)}) \eta_E^{(t)} E + \xi^{(t)} \eta_S^{(t)} E \right] \quad (9)$$

với các ràng buộc: (a) $R_{s_n}(t) \geq R_0$ đảm bảo QoS của SUE;

(b) $\Sigma_{m=1}^{M} x_n^m(t) = S_n(t)$;

(c) $\Sigma_{s=1}^{S} s_n(t) \leq S_{max}$

Hàm mục tiêu: Hàm mục tiêu $ma \Sigma_{t=1}^{T} \eta(t) = \Sigma_{t=1}^{T} \left[(1 - \xi^{(t)}) \eta_{EE}^{(t)} + \xi^{(t)} \eta_{SE}^{(t)} \right]_x$ nhằm tối đa hóa tổng hiệu suất của hệ thống qua tất cả các khe thời gian từ t=1 đến T. Hiệu suất tổng hợp η(t) tại mỗi thời điểm là tổ hợp tuyến tính có trọng số của hiệu quả năng lượng (EE) và hiệu quả phổ (SE). Hệ số trọng số ξ(t) thay đổi động theo số lượng người dùng trong hệ thống: khi có nhiều người dùng, ξ(t) tăng lên ưu tiên SE để đáp ứng nhu cầu băng thông; khi ít người dùng, ξ(t) giảm xuống ưu tiên EE để tiết kiệm năng lượng. Cách tiếp cận trọng số động này chuyển đổi bài toán tối ưu đa mục tiêu (MOOP) thành bài toán tối ưu đơn

mục tiêu (SOOP), giúp đơn giản hóa việc giải quyết trong khi vẫn cân bằng được hai mục tiêu quan trọng của hệ thống mạng.

Ràng buộc (a): $R_{s_n}(t) \geq R_0$Ràng buộc này đảm bảo chất lượng dịch vụ (QoS) cho mỗi người dùng SUE trong hệ thống. Cụ thể, tốc độ dữ liệu $R_{s_n}(t)$của người dùng s_ntại thời điểm t phải lớn hơn hoặc bằng ngưỡng tối thiểu R_0(ví dụ 1 Mbps).

Ràng buộc (b): $\sum_{m=1}^{M} \quad x_n^m(t) = S_n(t)$ ràng buộc này quy định rằng tổng số khối tài nguyên (RB) được phân bổ cho tế bào con n phải bằng chính xác số lượng người dùng $S_n(t)$đang hoạt động trong cell đó tại thời điểm t.

Ràng buộc (c): $\sum_{s=1}^{S} \quad s_n(t) \leq S_{max}$ ràng buộc này giới hạn tổng số người dùng mà mỗi tế bào con có thể phục vụ đồng thời không vượt quá giá trị tối đa S_{max}. Khi số người dùng đến vượt quá S_{max}, hệ thống phải từ chối kết nối mới hoặc chuyển giao (handover) sang cell lân cận. Ràng buộc này đảm bảo hệ thống hoạt động trong vùng ổn định, tránh quá tải gây suy giảm chất lượng dịch vụ cho tất cả người dùng.

3. MÔ TẢ THUẬT TOÁN

Trong bài báo này, ngoài EE, SE cũng được xem xét toàn diện, vì vậy bài toán phân bổ tài nguyên trở thành bài toán NP-hard, và khó thu được nghiệm tối ưu. Khi đưa ra quyết định cho mỗi khe MgNB, nó chỉ phân bổ tài nguyên cho SUE mới vào khe đó. Sẽ không có phân bổ lại cho SUE đã được phân bổ. Để cân bằng EE và SE của hệ thống, bài toán phân bổ tài nguyên trong kịch bản 5G UDN có thể được biểu diễn như MDP. Bài báo này áp dụng phương pháp học tăng cường sâu để giải quyết bài toán này.

3.1. Mô hình cơ bản của học tăng cường

Học tăng cường coi việc học như một quá trình đánh giá heuristic. Agent chọn một hành động cho môi trường. Trạng thái thay đổi sau khi môi trường chấp nhận hành động. Đồng thời, một phản hồi thưởng được tạo ra cho Agent. Agent chọn hành động tiếp theo theo thưởng và trạng thái hiện tại của môi trường. Nguyên tắc lựa chọn là tăng xác suất thưởng, để thu được chiến lược tối ưu. Do đó, môi trường, trạng thái và thưởng là ba yếu tố then chốt trong học tăng cường. Đối với hệ thống được xem xét trong bài báo này, chúng tôi định nghĩa không gian trạng thái, không gian hành động và thưởng trong khe thời gian t dựa trên khung học tăng cường:

• Không gian trạng thái: Các quyết định được đưa ra bởi agent MgNB. Agent cần biết trạng thái của mỗi tế bào con để xác định hành động. Do đó, trạng thái của agent trong khe thời gian t là: $s_t = s_1(t), R_1(t), \dots, s_n(t), R_n(t), x(t)$. Điều này có nghĩa là agent sẽ biết số lượng và thông lượng của tất cả các tế bào con và việc phân bổ tất cả RB trong hệ thống.

• Không gian hành động: Agent quyết định những RB nào được tái sử dụng bởi tế bào con . Vì vậy hành động là: $a_t = x_1(t), x_2(t), \dots, x_n(t)$. Không gian hành động tăng theo cấp số nhân với sự tăng của SgNB. Sự bùng nổ không gian hành động sẽ là một vấn đề quan trọng và khó khăn cần xử lý.

• Hàm thưởng: Khi agent MgNB thực hiện hành động a bằng cách quan sát trạng thái s_t, nó sẽ nhận được thưởng tức thì r_t. Mục tiêu của chúng tôi là tối đa hóa $\Sigma_{t=1}^{T} \eta(t)$, vì vậy hàm thưởng là η(t) tại thời điểm t:

$$r_t = \eta(t) = \left(1 - \xi^{(t)}\right)\eta_E^{(t)}E + \xi^{(t)}\eta_S^{(t)}E \quad (11)$$

3.2. Chiến lược Deep Q Network

Tại mỗi thời điểm t, agent MgNB xác định hành động $a_t = \pi(s_t)$ thông qua chính sách π theo trạng thái hiện tại s_t. Tức là, MgNB được thưởng bằng cách phân bổ các khối tài nguyên khả dụng cho SgNB. Trong học tăng cường, lợi nhuận kỳ vọng được định nghĩa bởi hàm giá trị trạng thái-hành động $Q^{\pi}(s_t, a_t)$, có thể được biểu diễn như:

$$Q^{\pi}(s_t, a_t) = E_{\pi}[\Sigma_{t=1}^{T}\gamma^t\eta(s = s_t, a = a_t)] \quad (12)$$

trong đó γ là hệ số chiết khấu. Mục tiêu là làm cho lợi nhuận tương lai ít liên quan hơn đến hiện tại. Mục tiêu của MDP là tìm một chiến lược tối ưu, tức là chiến lược nhận được nhiều thưởng nhất. Điều này có nghĩa là đối với tất cả các hành động được chọn bởi chiến lược tối ưu $\pi*(s_t) = argmaxQ(s_{t+1}, a_{t+1}), a_t = \pi*(s_t)$ sẽ tối đa hóa hàm giá trị trạng thái-hành động.

Khi không gian rất lớn, rất khó tìm chiến lược tối ưu bằng cách tra cứu bảng Q-value trong Q-learning. Trong DQN, mạng nơ-ron sâu (DNN) có thể được sử dụng để xấp xỉ chiến lược tối ưu và hàm giá trị tối ưu: $Q*(s_t, a_t; \theta) \approx Q(s_t, a_t)$, trong đó θ là tham số của mạng nơ-ron.

Để đảm bảo tính ổn định của $Q*(s_t, a_t; \theta)$, mỗi bước cần huấn luyện mạng nơ-ron được đánh giá để tối thiểu hóa hàm mất mát L(θ):

$$L(\theta) = E\left[\left(\eta(t) + \gamma maxQ(s_{t+1}, a_{t+1}; \theta^{-}) - Q*(s_t, a_t; \theta)\right)^2\right] \quad (13)$$

trong đó θ^{-} là tham số của target network, và θ là tham số của behavior network.

3.3. Chiến lược Dueling Deep Q Network

Trong nhiều trạng thái, kích thước hàm giá trị của RB được gán cho các người dùng khác nhau là khác nhau. Tuy nhiên, trong một số trạng thái, các chính sách phân bổ khác nhau có thể dẫn đến các hàm giá trị giống hệt nhau. Theo [1], Dueling DQN là thuật toán cải tiến dựa trên DQN, sử dụng cấu trúc mô hình để biểu diễn hàm giá trị ở dạng chi tiết hơn. Để mô hình có thể có hiệu suất tốt hơn. Đặc biệt, hàm giá trị trạng thái-hành động được phân tách thành hàm giá trị dựa trên trạng thái và hàm lợi thế:

$$Q*(s_t, a_t; \theta, \mu, \omega) = V(s_t; \theta, \mu) + A(s_t, a_t; \theta, \omega) \quad (14)$$

trong đó μ, ω và θ lần lượt đại diện cho các tham số của luồng giá trị trạng thái, luồng lợi thế hành động và các phần còn lại của mô hình. Tuy nhiên, trong thực tế, luồng lợi thế của hành động thường được đặt là giá trị của hàm lợi thế của từng hành động trừ đi giá trị trung bình của tất cả các hàm lợi thế của hành động trong một trạng thái nhất định. Do đó, hàm lợi thế thực sự được biểu diễn như:

$$A(s_t, a_t; \theta, \omega) = A(s_t, a_t; \theta, \omega) - (1/|A|)\Sigma_{a'}A(s_t, a_t'; \theta, \omega) \quad (15)$$

Phép toán của (15) không chỉ đảm bảo rằng hàm trội của mỗi hành động trong trạng thái này không thay đổi, mà còn giảm phạm vi của Q-value và loại bỏ bậc tự do dư thừa, cải thiện tính ổn định.

Thuật toán 1: Phân bổ tài nguyên động dựa trên Dueling DQN với truyền thông hợp tác

Đầu vào: (state, action), QoS, tập RB khả dụng M, ngưỡng weak user

Đầu ra: Chuỗi hành động tối ưu $x_n^m(t)$, tổng trọng số Ση(t) của EE và SE

1: Khởi tạo replay buffer D với dung lượng N
2: Khởi tạo mạng giá trị trạng thái-hành động $Q*(s_t, a_t; \theta)$ với trọng số θ
3: Khởi tạo target network $Q*(s_t', a_t'; \theta^-)$ với trọng số θ⁻
4: for episode = 1 : K do
5: Khởi tạo môi trường hệ thống 5G UDN, MgNB nhận trạng thái ban đầu s_1
6: Xác định weak users dựa trên chất lượng kênh (50% thấp nhất)
7: for t = 1 : T do
8: MgNB chọn hành động a_t tại trạng thái s_t sử dụng ε-greedy từ $a_t = maxQ*(s_t, a_t, \theta)$
9: if coop_mode then
10: Gán 2 SGNB tốt nhất cho mỗi weak user
11: Áp dụng MRC: $h_{combined} = \sqrt{(\Sigma|h_i|^2)}$
12: end if
13: MgNB thực hiện hành động a_t để phân bổ các RB đã chọn cho SUE
14: Tính thưởng tức thì r_t dựa trên (11) (bao gồm cooperative bonus)
15: MgNB nhận trạng thái hệ thống tại thời điểm tiếp theo s_{t+1}
16: MgNB lưu kinh nghiệm s_t, a_t, r_t, s_{t+1} vào replay buffer D
17: if dung lượng của D đã đạt N then
18: MgNB chọn ngẫu nhiên một batch mẫu s_j, a_j, r_j, s_{j+1} từ D
19: MgNB tính hai luồng mạng: $V(s_t, \theta, \mu)$ và $A(s_t, a_t, \theta, \omega)$
20: Kết hợp thành $Q*(s_t, a_t; \theta, \mu, \omega)$ dựa trên (14)
21: if s_{j+1} là s_T then $y_j = r_j$
22: else $y_j = r_j + \gamma \max Q*\left(s_{j+1}, a_{j+1}'; \theta^-, \mu^-, \omega^-\right)$
23: MgNB tối thiểu hóa hàm mất mát $L(\theta, \mu, \omega) = E\left[\left(y_j - Q*(s_t, a_t; \theta)\right)^2\right]$
24: thông qua gradient descent dựa trên (13)
25: MgNB hoàn thành cập nhật tham số target network $\theta^- = \theta$ mỗi C bước
26: end if
27: end for
28: end for

4. ĐÁNH GIÁ HIỆU SUẤT

Trong phần này, chúng tôi đánh giá Dueling DQN để giải quyết bài toán phân bổ tài nguyên động. Kết quả mô phỏng cho thấy so với thuật toán không sử dụng truyền thông hợp tác, phương pháp Dueling DQN với truyền thông hợp tác được cải thiện.

4.1. Môi trường mô phỏng

Chúng tôi xem xét một MgNB, một mạng siêu dày đặc của các SgNB. Trong mô phỏng, mỗi người dùng chỉ có thể chiếm một RB. Trong Dueling DQN, để tránh hội tụ của mục tiêu tối ưu về cực tiểu cục bộ, chiến lược ε-greedy thích ứng được sử dụng. Các tham số mô phỏng được sử dụng được hiển thị trong Bảng 1.

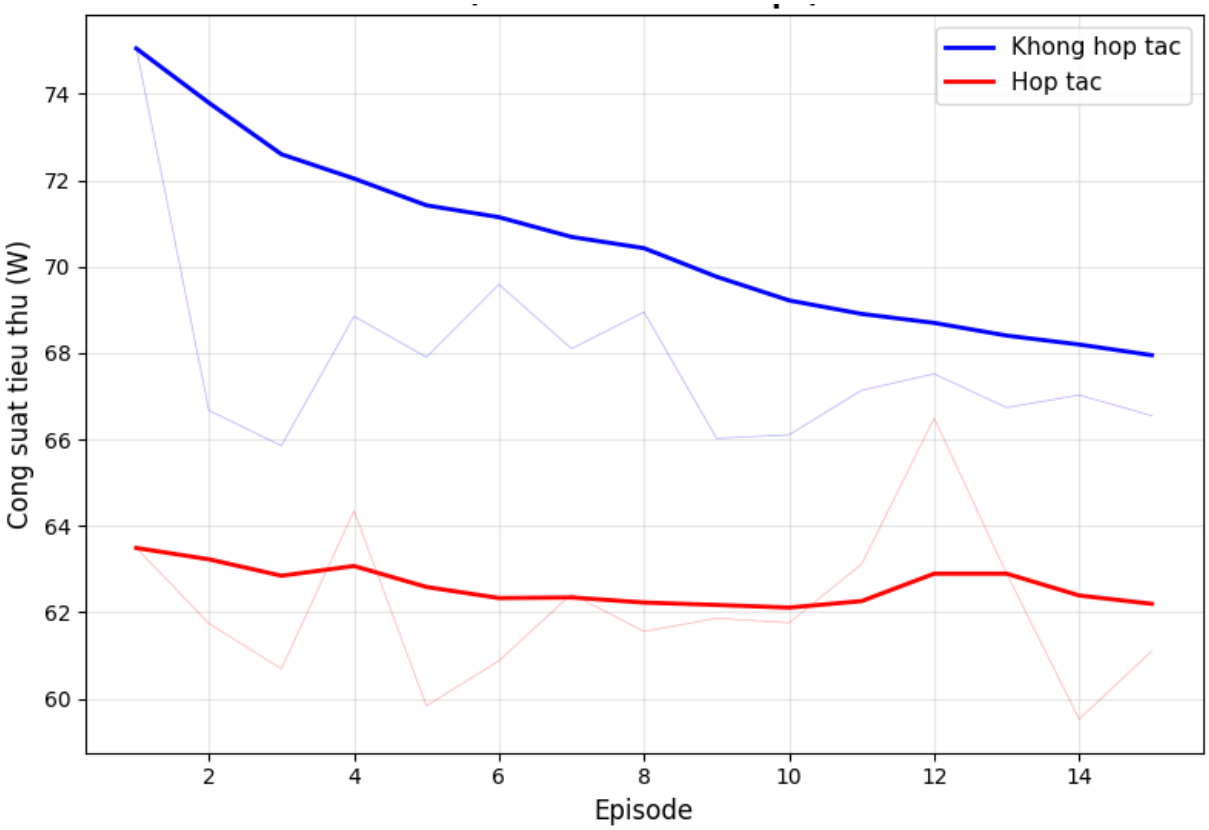
Bảng 1. Tham số mo phỏng

Tham số mô phỏng	Giá trị
Tổng băng thông hệ thống	10 MHz
Số lượng SGNB (N)	10
Số lượng người dùng (K)	5
Công suất SGNB hoạt động (p_{on})	6.8 W
Công suất SGNB ngủ (p_{slp})	4.3 W
Công suất phát tối đa (p_{tsm})	1.0 W
Hiệu suất bộ khuếch đại công suất (η)	0.25
Tốc độ học (learning rate)	0.001
Hệ số chiết khấu (γ)	0.99
Kích thước replay memory D	100000
Ngưỡng phân loại người dùng	50%

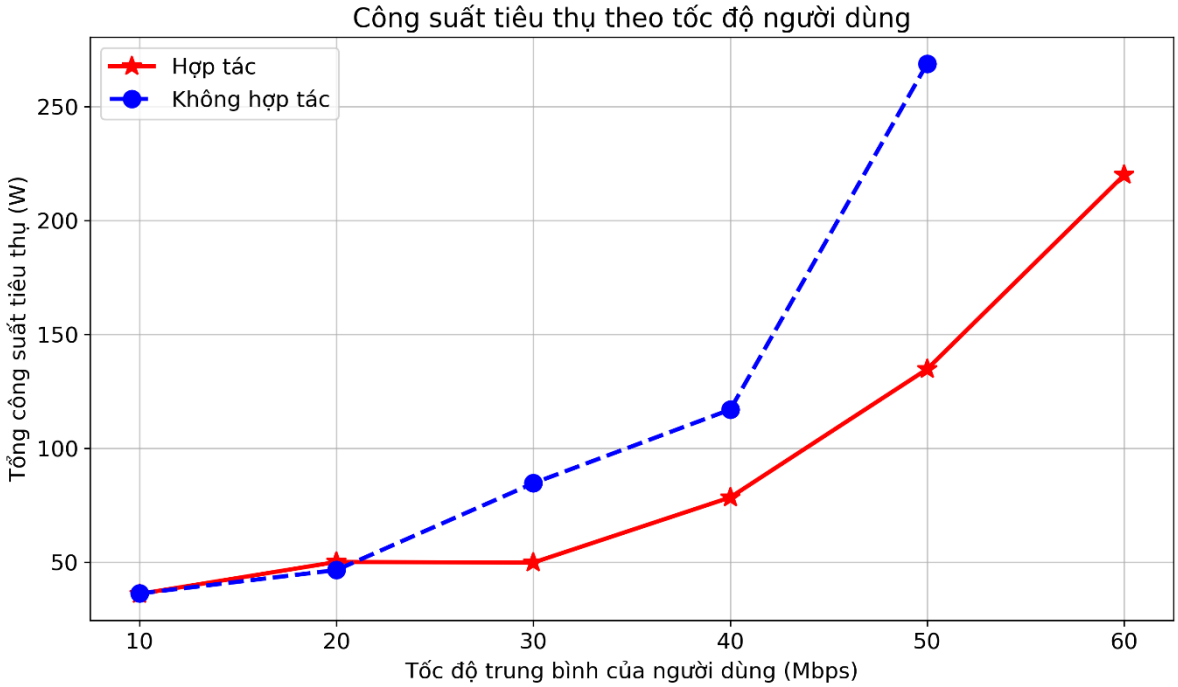
4.2. Kết quả mô phỏng và thảo luận

Hình 1 minh họa quá trình học của thuật toán Dueling DQN trong kịch bản nhu cầu trung bình (20–30 Mbps) với 10 SGNBs và 6 người dùng. Trục hoành biểu diễn số episode, trục tung là công suất tiêu thụ trung bình (W). Đường xanh đậm (không hợp tác) cho thấy công suất ban đầu cao (~75 W) và giảm dần theo episode, sau đó hội tụ tương đối nhanh từ khoảng episode 8–10, ổn định quanh mức 68–69 W. Các đường xanh nhạt phía sau thể hiện giá trị thô trước khi smoothing, phản ánh dao động do quá trình exploration của agent trong giai đoạn đầu học. Ngược lại, đường đỏ đậm (hợp tác) luôn duy trì ở mức thấp hơn, dao động quanh 62–63 W, với biên độ biến thiên nhỏ hơn so với trường hợp không hợp tác. Các đường đỏ nhạt cho thấy agent thử nghiệm các chiến lược hợp tác khác nhau (bật/tắt SGNB phối hợp), nhưng nhìn chung quá trình học ổn định hơn và hội tụ sớm. Khoảng chênh lệch công suất giữa hai cơ chế vào cuối quá trình học vào khoảng 5–6 W, tương đương 8–9%, cho thấy trong điều kiện tải trung bình, cơ chế hợp tác MRC kết hợp Dueling DQN vẫn mang lại lợi ích tiết kiệm năng

lượng rõ rệt so với phương pháp không hợp tác.



Hình 1: Công suất tiêu thụ trong quá trình huấn luyện



Hình 2: Công suất tiêu thụ theo tốc độ người dung

Hình 2 minh họa mối quan hệ giữa công suất tiêu thụ và nhu cầu người dùng trung bình cho hai phương pháp: Dueling DQN với truyền thông hợp tác (đường đỏ) và Dueling DQN không hợp tác (đường xanh). Ở mức nhu cầu thấp (10-40 Mbps), cả hai phương pháp có công suất tiêu thụ tương đương (~67-69W) do agent tối ưu hóa bằng cách tắt các SGNB không cần thiết. Khi nhu cầu tăng lên 50 Mbps, sự khác biệt bắt đầu rõ rệt: phương pháp hợp tác chỉ tiêu thụ 73.8W so với 101.2W của phương pháp không hợp tác, tiết kiệm 27.1% công suất. Ưu điểm vượt trội của cơ chế hợp tác thể hiện ở hai khía cạnh: (1) Tiết kiệm năng lượng đáng kể tại mức tải cao nhờ kỹ thuật MRC kết hợp tín hiệu từ 2 SGNB tốt nhất, giúp tăng SNR tổng hợp mà không cần tăng công suất phát; (2) Mở rộng khả năng phục vụ - phương pháp hợp tác có thể đáp ứng nhu cầu lên đến 60 Mbps với 79.5W, trong khi phương pháp không hợp tác chỉ đạt tối đa 50 Mbps và không thể tìm được giải pháp khả thi cho nhu cầu cao hơn do thiếu tài nguyên hỗ trợ người dùng yếu.

5. KẾT LUẬN

Bài báo đã đề xuất phương pháp tối ưu hóa tài nguyên động cho mạng 5G UDN dựa trên thuật toán Dueling Deep Q-Network kết hợp cơ chế truyền thông hợp tác sử dụng kỹ thuật Maximum Ratio Combining (MRC). Phương pháp mô hình hóa bài toán cân bằng EE-SE như quá trình quyết định Markov với hàm mục tiêu trọng số thích nghi. Cơ chế MRC tự động nhận diện 50% người dùng yếu và phân bổ hai SGNB tốt nhất để phục vụ đồng thời, cải thiện SNR tổng hợp mà không tăng công suất phát. Kết quả mô phỏng với 10 SGNBs và 6 người dùng cho thấy hai ưu điểm vượt trội: (1) Tiết kiệm năng lượng - tại mức tải 50 Mbps, phương pháp hợp tác tiêu thụ 73.8W so với 101.2W của phương pháp không hợp tác, giảm 27.1% công suất; (2) Mở rộng khả năng phục vụ - phương pháp hợp tác có thể đáp ứng nhu cầu lên đến 60 Mbps trong khi phương pháp không hợp tác chỉ đạt tối đa 50 Mbps do không đủ tài nguyên cho người dùng yếu.

TÀI LIỆU THAM KHẢO

[1] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-Dense Networks: A Survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 4, pp. 2522-2545, 2016.

[2] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," IEEE Access, vol. 3, pp. 1206-1232, 2015.

[3] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource Allocation for Ultra-Dense Networks: A Survey," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 2134-2168, 2019.

[4] Y. Cai, P. Cheng, Z. Chen, M. Ding, B. Vucetic, and Y. Li, "Deep Reinforcement Learning for Online Resource Allocation in Network Slicing," IEEE Trans. Mobile Computing, vol. 23, no. 6, pp. 7099-7116, 2024.

[5] M. Khani, S. Jamali, and M. K. Sohrabi, "Resource Allocation in 5G Cloud-RAN Using Deep Reinforcement Learning: A Review," Trans. Emerging Telecom. Tech., vol. 35, no. 1, e4929, 2024.

[6] V. Mnih et al., "Human-Level Control Through Deep Reinforcement Learning," Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[7] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling Network Architectures for Deep Reinforcement Learning," in Proc. ICML, pp. 1995-2003, 2016.

[8] Z. Liu, X. Chen, Y. Chen, and Z. Li, "Deep Reinforcement Learning Based Dynamic Resource Allocation in 5G Ultra-Dense Networks," in Proc. IEEE SmartIoT, pp. 168-174, 2019.

[9] H. Li, H. Gao, T. Lv, and Y. Lu, "Deep Q-Learning Based Dynamic Resource Allocation for Self-Powered Ultra-Dense Networks," in Proc. IEEE ICC Workshops, pp. 1-6, 2018.

[10] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative Communication in Wireless Networks," IEEE Communications Magazine, vol. 42, no. 10, pp. 74-80, 2004.