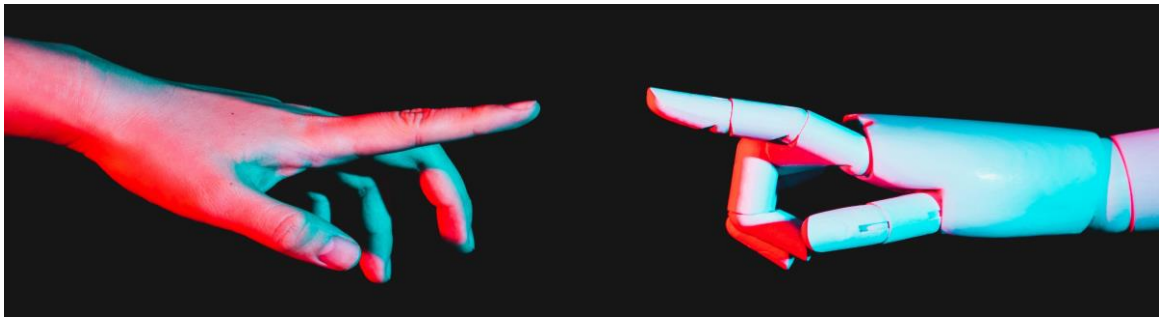


ĐỒ ÁN CS106: TÌM HIỂU VÀ NGHIÊN CỨU PRESSLIGHT TRONG BÀI TOÁN ĐIỀU HƯỚNG GIAO THÔNG

Members :
Le Truong Ngoc Hai – 20520481
Le Thi Phuong Vy – 20520355
Nguyen Nhat Truong – 20522087
Lai Chi Thien – 20520309
University of Information Technology, Ho Chi Minh, Viet Nam

Keywords

Deep reinforcement learning, traffic signal control, multi-agent system

Note

Vì việc sử dụng Tiếng Việt sẽ không mô tả đầy đủ được định nghĩa và tư tưởng của các phương pháp, kỹ thuật, phép toán,.... Do đó, chúng em xin phép được giữ nguyên các thuật ngữ Tiếng Anh, cũng như các đại lượng

ABSTRACT

Thế giới ngày càng phát triển, công nghệ ngày càng tân tiến, số lượng phương tiện giao thông cũng vì thế mà tăng lên đáng kể, gây ra không ít tác động tiêu cực đến tình hình giao thông vận tải nói chung. Do đó, việc thiết kế một phương pháp để giải quyết vấn đề này là thực sự cần thiết, đó là bài toán Traffic Signal Control (Điều khiển tín hiệu giao thông). Traffic Signal Control là điều cần thiết cho một hệ thống giao thông vận tải hiệu quả trong mạng lưới đường bộ. Đây là một vấn đề mang đầy thách thức bởi tính linh động và phức tạp của giao thông. Các nghiên cứu về giao thông thông thường gặp phải tình trạng không đủ năng lực để thích ứng với các tình huống giao thông phức tạp. Hầu hết các mô hình dựa trên reinforcement learning (RL) tập trung thiết kế các nhân tố chính - reward (điểm thưởng) và state (trạng thái) - theo một cách heuristic. Điều này dẫn đến một kết quả có độ nhạy quá cao và tiêu tốn rất nhiều thời gian để học.

Để loại bỏ tính heuristic trong các nhân tố của RL, một phương pháp được đề xuất bằng cách liên kết RL với các nghiên cứu lân cận về giao thông vận tải. Phương pháp PressLight được lấy cảm hứng từ một phương pháp state-of-the-art (SOTA) max pressure trong lĩnh vực giao thông vận tải. Các lý thuyết của phương pháp này đều được hỗ trợ nhờ MP - một phương pháp có thể tối đa hóa thông lượng của một mạng giao thông. Phương pháp này sẽ được chạy thực nghiệm trên cả dữ liệu tổng hợp (Synthetic data) và dữ liệu thế giới thực (Real-world data).

1. INTRODUCTION

Bài toán điều hướng giao thông luôn luôn là một đề tài nghiên cứu đã và đang nhận được nhiều sự chú ý bởi tính chất dao động và phức tạp của bài toán. Như chúng ta thường thấy, các tình huống giao thông luôn luôn dao động ở mức độ cao và từ đó, việc thiết kế một chiến lược tín hiệu giao thông để điều chỉnh các tình trạng này là thực sự cần thiết, ít nhất là tránh được ùn tắc giao thông.
Các kỹ thuật học tăng cường (RL) đã và đang được ứng dụng nhiều hơn cho bài toán điều hướng giao thông. Rất nhiều các nghiên cứu đã cho thấy sự vượt trội của việc ứng dụng các kỹ thuật RL [1, 2, 3, 4, 5, 6]. Lợi ích to lớn nhất mà RL mang lại là nó sẽ học cách đưa ra hành động tiếp theo sau khi quan sát và phản hồi từ những thử nghiệm trước đó. Một bất lợi chính của các phương pháp điều khiển tín hiệu giao thông mà tiếp cận dựa trên RL là quá trình thiết lập sẽ thường mang tính heuristic và thiếu lý thuyết cơ sở từ các tài liệu giao thông vận tải. Vấn đề này làm cho kết quả đạt được có độ nhạy quá cao, nói cách khác là không ổn định, dẫn đến một quá trình học tiêu tốn rất nhiều thời gian. Bằng việc phân tích hai yếu tố cơ bản - reward và state - của các phương pháp RL-based, vấn đề sẽ được giải thích rõ ràng hơn.
Về việc thiết kế reward, rất nhiều thiết kế dành riêng cho reward được đề xuất dựa trên cơ sở lý thuyết. Trong bài toán mà chúng ta đang xét, nhiệm vụ tối cao cần giải quyết là giảm thời gian di chuyển của các phương tiện (travel time) nhưng đây lại là một . Travel time là một reward lâu dài và bị chi phối bởi một chuỗi các hành động, do đó mức độ ảnh hưởng của một hành động khó có thể được phản ánh qua travel time bởi nó chiếm một phần rất nhỏ so với tổng bộ. Vì vậy, người ta sẽ có thiên hướng thiết kết reward ngắn hạn như độ dài hàng chờ hay độ trễ để tính toán xấp xỉ travel time. Những phương pháp trước đây chủ yếu tập trung vào việc điều hướng giao thông cho một ngã tư, nhưng trong PressLight, nhóm tác giả hướng tới bài toán điều hướng cho nhiều ngã tư (multi-intersection).

2. RELATED WORKS

Individual Traffic Signal Contral: Rất nhiều nghiên cứu đã được thực hiện về Individual Traffic Signal Control trong lĩnh vực giao thông. Lời giải mà những nhà nghiên cứu là làm sao để tối ưu hóa thời gian di chuyển hoặc độ trễ của các phương tiện [7, 8, 9, 10, 11] giả sử các phương tiện đang di chuyển theo một lộ trình thiết lập từ trước. Bên cạnh đó, các phương pháp dựa trên RL tiếp cận bằng cách học trực tiếp từ dữ liệu cho trước, từ đó đưa ra lời giải tối ưu. Hơn nữa, deep RL [12, 13, 14, 4, 5, 15] lại càng được khai thác bởi tính chất phức tạp của các trạng thái trong bài toán. Rõ ràng ta thấy, một trạng thái càng phức tạp thì sẽ càng dễ ảnh hưởng đến hiệu suất mô hình. Khác với những phương pháp trước khi họ thiết kế reward để điều khiển tín hiệu cho một giao lộ, PressLight tập trung vào nhiều giao lộ một lúc (multi-intersection).
RL-based Multi-intersection Traffic Signal Control: Nhận thấy được tiềm năng của RL trong bài toán điều hướng giao thông đơn lẻ, các nhà nghiên cứu đã hướng tới việc thiết kế một chiến lược mà có thể điều khiển nhiều giao lộ hơn dựa trên bước đệm này. Một trong những cách tiếp cận đầu tiên là mô hình hóa hành động (action) giữa các learning agent cùng nhau với tối ưu hóa tập trung (centralized optimization) [16, 4]. Tuy nhiên, phương pháp này lại cần sự "trao đổi" giữa các agent trong toàn bộ một mạng, điều này dẫn đến sự tốn kém về chi phí tính toán. Sử dụng các agent RL phi tập trung là một cách tiếp cận khác để điều khiển tín hiệu giao thông của hệ thống nhiều giao lộ (multi-intersection) [25, 11, 17]. Vì mỗi một agent sẽ cho ra một chiến lược nhất định dựa trên những thông tin nó nắm được cộng hưởng với thông tin từ các giao lộ lân cận mà không cần một quyết định thống nhất nên các phương pháp phi tập trung có thể khả thi hơn nhiều. Khi ta áp các bộ điều khiển giao thông giao lộ vào hệ thống, các hệ thống phi tập trung sẽ dễ dàng mở rộng quy mô hơn. PressLight cũng được thiết kế dựa trên tư tưởng này.

3. PRELIMINARIES

3.1. Definition 3.1 (Làn đến (Incoming lane) và làn đi (Outgoing lane) của một giao lộ)

Làn đến (Incoming lane) của một giao lộ là một làn mà các phương tiện bắt đầu di chuyển vào giao lộ. Một làn đi của một giao lộ là một làn mà các phương tiện giao thông rời khỏi giao lộ. Làn đi và làn đến được kí hiệu lần lượt là L_{in} and L_{out}.

3.2. Definition 3.2 (Traffic movement – Các chuyển động giao thông)

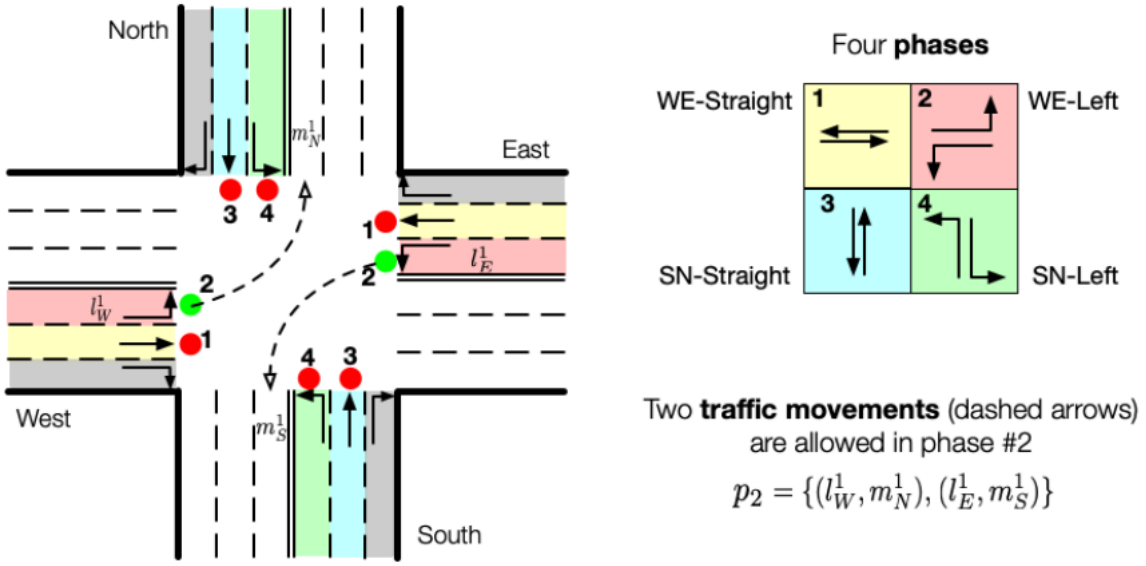
Một chuyển động giao thông được định nghĩa là lượng giao thông chuyển giao từ một làn đến sang một làn đi. Một chuyển động từ làn l sang làn m được ký hiệu là (l, m).

3.3. Definition 3.3 (Tín hiệu và pha của chuyển động)

Một tín hiệu của chuyển động được định nghĩa dựa trên chuyển động giao thông (traffic movement). Trong đó, tín hiệu đèn xanh ám chỉ chuyển động tương ứng là được cho phép, ngược lại là tín hiệu đèn đỏ khi chuyển động đó bị cấm. Tín hiệu chuyển động được ký hiệu là $a(l, m)$. Trong đó:

- $\boldsymbol{a(l, m) = 1}$ biểu thị cho tín hiệu đèn xanh cho chuyển động $\boldsymbol{(l, m)}$
- $\boldsymbol{a(l, m) = 0}$ biểu thị cho tín hiệu đèn đỏ cho chuyển động $\boldsymbol{(l, m)}$

Một pha chuyển động được ký hiệu $p = \{(l, m) | a(l, m) = 1\}$ với $l \in L_{in}$ và $m \in L_{out}$.



Hình 1. Pha và chuyển động giao thông trong bài toán điều hướng giao thông.

Trong Hình 1, ta thấy có 12 làn đến và 12 làn đi. Có 8 tín hiệu giao thông (đèn xanh, đèn đỏ). Gồm 4 pha chuyển động giao thông trong giao lộ: WE-Straight (đi theo đường thẳng từ West sang East), SN-Straight (đi theo đường thẳng từ South sang North), SN-Left (đi theo ngã rẻ bên trái từ West sang East), WE-Left (đi theo ngã rẻ bên trái từ South sang North). Cụ thể, hai chuyển động giao thông được cho phép trong pha WE-Left.

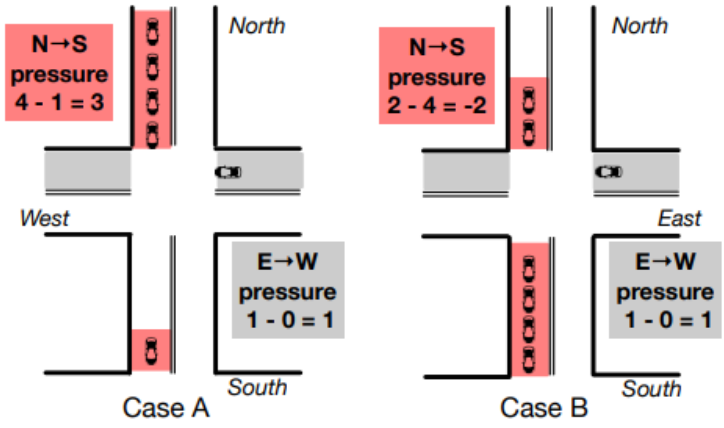
3.4. Definition 3.4 (Áp lực của chuyển động, áp lực của giao lộ - Pressure of movement, pressure of intersection)

Pressure of a movement được định nghĩa là sự chênh lệch số lượng phương tiện giao thông giữa làn đến và làn đi. Mật độ phương tiện của một làn được tính bằng $\frac{x(l)}{x_{max}(l)}$ với $x(l)$ là số lượng phương tiện trên làn l, $x_{max}(l)$ là số lượng phương tiện tối đa có thể đi trên làn l. Áp lực của chuyển đông (l, m) được tính băng:

$$w(l,m) = \frac{x(l)}{x_{max}(l)} - \frac{x(m)}{x_{max}(m)} \quad (1)$$

Áp lực của một giao lộ i được tính bằng tổng trị tuyệt đối của áp lực của toàn bộ chuyển động giao thông:

$$P_i = \left| \sum_{(l,m) \in i} w(l,m) \right| \quad (2)$$



Hình 2. Minh họa cho áp lực ở từng làn trong một giao lộ

Ở Hình 2, áp lực của giao lộ ở Case A là $|3 + 1| = 4$, trong khi đó ở Case B là $|-2 + 1| = 1$. Với đại lượng này, ta có thể nhận xét được mức độ mất cân bằng giữa mật độ xe ở làn đến và xe ở làn đi. Áp lực càng lớn, phân phối của làn đường đó càng mất cân bằng.

4. METHOD

Có ba đại lượng cần chú ý cho một agent điều khiển giao lộ như sau: state (trạng thái), action (hành động), reward (điểm thưởng).

4.1. Agent Design

- **State – Trạng thái:** Trạng thái được định nghĩa cho một giao lộ tương đương với định nghĩa quan sát trong multi-agent RL. Trạng thái bao gồm pha hiện thời p, số lượng phương tiện ở mỗi làn đi $x(m)$ $(m \in L_{out})$ và số lượng các phương tiện trong mỗi làn con (segment) của mỗi làn đến $x(l)_k$ $(l \in L_{in}, k = 1 \dots K)$. Trong bài báo này, mỗi làn được chia thành 3 làn con (segment) như nhau $(K = 3)$.
- **Action – Hành động:** Tại chu kì t, mỗi agent chọn một pha p như là một hành động a_t từ tập các hành động (action set) **A**, ám hiệu rằng tín hiệu giao thông nên được đặt theo pha p. Trong PressLight, mỗi agent sẽ có tối đa 4 hành động được cho phép tương ứng với 4 pha

ở Hình 1. Mỗi một hành động đề cử a_i sẽ được thể hiện như là một one-hot-vector. Tuy nhiên trong thế giới thực thì không chỉ có mỗi 4 pha như vậy mà còn nhiều hơn.

- **Reward – Điểm thưởng:** Reward r_i được tính như sau:

$$r_i = -P_i \quad (3)$$

Với P_i là áp lực của giao lộ i, được định nghĩa trong công thức (2). P_i là đại lượng cho thấy sự mất cân bằng mật độ phương tiện giữa làn đến và làn đi. Bằng cách tối thiểu hóa P_i thì ta sẽ được phân phối các phương tiện trong hệ thống là như nhau. Nhờ vậy tín hiệu đèn xanh sẽ được sử dụng một cách hiệu quả để có được một lời giải tối ưu sau cùng. Để ổn định quá trình huấn luyện, tác giả đã sử dụng bộ nhớ phát lại (replay memory) như được mô tả trong[18] bằng cách thêm các mẫu dữ liệu mới vào và thỉnh thoảng xóa các mẫu cũ. Theo từng chu kì, agent sẽ lấy mẫu từ bộ nhớ và sử dụng chúng để cập nhật network.

4.2. Learning Process

PressLight sử dụng Deep Q-Network (DQN) để làm hàm ước lượng nhằm tính toán Q-value function.

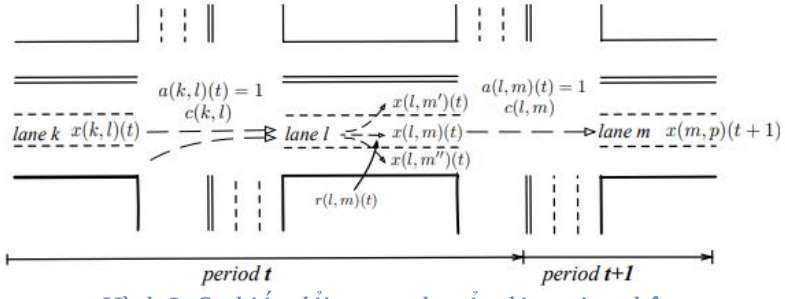
5. JUSTIFICATION OF RL AGENT.

Bảng 1: Tổng hợp các kí hiệu

Notation	Meaning
L_{in}	Tập các làn đến trong một giao lộ
L_{out}	Tập các làn đi trong một giao lộ
(l, m)	Một chuyển động giao thông từ làn l sang làn m
$x(l, m)$	Số lượng phương tiện rời làn l sang làn m
$x(l)$	Số lượng phương tiện trên làn l
$x(l)_k$	Số lượng phương tiện trên làn con thứ k của làn l (k-th segment of l)
$x_{max}(m)$	Số lượng phương tiện tối đa cho phép trên làn m
$r(l, m)$	Tỉ số thay đổi của chuyển động giao thông từ làn l sang làn m (Turning ratio)
$c(l, m)$	Lượng phương tiện rời đi của chuyển động (l, m) (Discharging rate)
$a(l, m)$	= 1 nếu bật đèn xanh cho chuyển động (l, m) = 0 nếu ngược lại

5.1. Justification of State Design

5.1.1. General description of traffic movement as a Markov chain – Mô tả khái quát chuyển động giao thông như dây chuyền Markov



Hình 3. Sự biến đổi trong chuyển động giao thông

Hình 3 cho thấy sự liên kết của một chuyển động giao thông riêng biệt với mỗi làn đến $l \in L_{in}$ và mỗi $m \in Out_l$ với Out_l là tập các đường ra khỏi l. Đặt $x(l,m)(t)$ là số lượng phương tiện tại đầu chu kì t, $X(t) = \{x(m,l)(t)\}$ là state – trạng thái chuyển động của mạng, được xem như là o^t. Có hai biến độc lập với $X(t)$:

- **Turning ratio** $r(l,m)$: $r(l,m)$ là một biến phân phối ngẫu nhiên và độc lập cho biết tỉ trọng số xe đi vào làn m từ làn l so với tổng số xe trên làn l.
- **Discharging rate** $c(l,m)$: Với mỗi chuyển động (l,m), tỉ lệ xả hàng đợi $c(l,m)$ là một biến không âm, bị ràng buộc, phân phối ngẫu nhiên và độc lập, tức là $c(l,m) < C(l,m)$ với $C(l,m)$ là tốc độ dòng bão hòa.

Tại thời điểm kết thúc chu kì t, một hành động $A^t = \{(l,m)|a^t(l,m)\}$ phải được chọn từ tập các hành động $\boldsymbol{A^t}$ như là một hàm của X^t để sử dụng cho chu kì $(t+1)$, như Hình 3 ta thấy agent đã bật đèn xanh cho chuyển động từ l sang m.
Với mỗi (l,m) và t, khai triển của $x(l,m)$ bao gồm lượng phương tiện nhận vào và lượng phương tiện thoát ra như công thức sau:

$$\begin{aligned} &x(l,m)(t+1) \\ &= x(l,m)(t) + \sum_{k \in In_l} \min\left[c(k,l) \cdot a(k,l)(t), x(k,l)(t)\right] \cdot r(l,m) \\ &- \min\{c(k,l) \cdot a(k,l)(t), x(k,l)(t)\} \cdot 1(x(m) < x_{max}(m)) \end{aligned} \quad (4)$$

Với In_l là tập các làn đường dẫn vào l. Khi l là làn đến (incoming lane), sẽ có đến $x(k,l)$ phương tiện di chuyển từ k sang nếu $a(k,l)(t) = 1$ (đèn xanh) và những phương tiện đó sẽ tham gia vào (l,m) nếu $r(l,m) = 1$. Khi chuyển động (l,m) được thực hiện, tức là, $a(l,m) = 1$, sẽ có đến $x(l,m)$ phương tiện rời làn l và chuyển hướng sang làn m nếu không vi phạm bất kì ràng buộc hay trở ngại nào, ví dụ là $x(m) < x_{max}(m)$.
Giả sử trạng thái khởi tạo $X(1) = x(l,m)(1)$ là một biến ngẫu nhiên có giới hạn. Vì $A(t) = a(l,m)(t)$ là hàm của trạng thái hiện thời $X(t)$, và $c(l,m)$ và $r(l,m)$ là hai biến độc lập với $X(1), \dots, X(t)$, ta thấy $X(t)$ là một dây chuyền Markov.

5.1.2. Specification with proposed state definition – Chi tiết về định nghĩa trạng thái đề xuất

Công thức của chuyển động giao thông có thể được chuyển hóa từ cấp độ làn (lane-level) sang cấp độ làn con (segment-level). Ta có $x(l_1)$ là số lượng phương tiện trên segment l_1 gần với giao lộ nhất và $x(l_2)$ là số lượng phương tiện trên segment l_2. Giả sử, các phương tiện chuyển làn ngay khi chúng đi vào làn đó, tức là, $x(l,m) = x(l)$, và tất cả phương tiện ở l_{i+1} đi vào làn con tiếp theo l_i trong thời điểm t, quá trình chuyển động trên làn con gần nhất với giao lộ được viết như sau:

$$x(l)_1(t+1) = x(l)_1(t) + x(l)_2(t) - \min\{c(l,m) \cdot a(l,m)(t), x(l)_1(t)\} \cdot 1(x(m) < x_{max}(m)) \quad (5)$$

Công thức cho các segment khác sẽ được triển khai tương tự.
Với một giao lộ i, $c(l,m)$ là một hằng số đặc trưng vật lý của mỗi chuyển động, tuy nhiên $x(l)_1$, $x(l)_2$ và $x(m)$ được cung cấp cho RL agent trong định nghĩa trạng thái. Do đó, trạng thái được đề xuất sẽ có thể mô tả toàn diện được tính động của cả một hệ thống giao thông.

5.2. Jusification for Reward Design

5.2.1. Stabilization on traffic movements with proposed reward - Ổn định chuyển động giao thông với reward đề xuất.

Mục tiêu của RL agent được chứng minh là sẽ ổn định độ dài hàng chờ, từ đó tối đa hóa thông lượng của hệ thống và tối thiểu hóa thời gian di chuyển của các phương tiện.

Definition 5.2: Movement process stability – Sự ổn định của quá trình chuyển động
Một quá trình chuyển động $X(t) = \{x(l,m)(t)\}$ là ổn định về giá trị trung bình (với u là chiến lược kiểm soát ổn định) nếu đối với một số $M < \infty$, điều sau là đúng:

$$\sum_{t=1}^{T} \sum_{(l,m)} E[x(l,m)(t)] < M, \quad \forall T \tag{6}$$

Với E là kỳ vọng. Sự ổn định chuyển động trong giá trị trung bình có nghĩa là dây chuyền lặp lại dương và có phân phối xác suất trạng thái ổn định duy nhất cho tất cả T.

Definition 5.3: Max-pressure control policy [19]
Tại mỗi chu kì t, agent sẽ chọn hành động với áp lực (pressure) tối đa tại mỗi trạng thái X: $\tilde{A} * (X) = argmax_{\tilde{A} \in \boldsymbol{A}} \theta(\tilde{A}, X)$, với $\tilde{A}$ được định nghĩa như sau:

$$\theta(\tilde{A}, X) = \sum_{(l,m):a(l,m)=1} \tilde{w}\,(l,m)$$

Với $\tilde{w}(l,m) = x(l) - x(m)$ là áp lực của từng chuyển động. Để phân biệt giữa chiến lược max-pressure và chiến lược RL, các kí hiệu sẽ được thêm dấu ngã, ví dụ: $\tilde{A}$

Với một chiến lược RL điều khiển tối ưu, agent chọn hành động A với giá trị tối ưu $Q(A,X)$ tại mỗi trạng thái X:

$$A^*(X) = \underset{A \in \boldsymbol{A}}{argmax}\, Q(A,X) \tag{7}$$

Với $Q_t(A,X) = E[r_{t+1} + \gamma r_{t+2} + \cdots | A, X]$ thể hiện cho tổng reward tối đa tại trạng thái X khi chọn hành động A tại thời điểm t (để tối giản thì Công thức (7) sẽ không chứa t). Sự khác nhau giữa định nghĩa của pressure trong RL reward với max-pressure là agent RL sử dụng trọng số pressure xem xét x_{max} trong Công thức (1). Giả sử tất cả các làn đều có cùng $x_{max}(l)$, thì kết quả ổn định vẫn đúng với $x(l)$ chuẩn hóa.
Định lý: Xét việc mở rộng hàng đợi vật lý trong đường chính, hành động A^* được chọn bởi chiến lược RL của tác giả cũng mang đến chuyển động có tính ổn định.
Để chứng minh cho định lý trên, ta xét trong ngữ cảnh ở đường chính như sau:

- Giả sử số lượng phương tiện tối đa x_{max} của làn bên m^{side} là vô hạn, do đó số hạng thứ hai trong Công thức (1) bằng 0. Từ đó ta có $w\left(l, m^{side}\right) = \frac{x(l)}{x_{max}(l)} > 0$:.
- Khi làn đi (outgoing lane) m^{main} dọc theo đường chính bị bão hòa (đầy), số hạng thứ hai của Công thức (1) sẽ xấp xỉ bằng 1 vì sự mở rộng hàng đợi. Do đó, ta có:

$$w\left(l, m^{main}\right) \approx \frac{x(l)}{x_{max}(l)} - 1 < 0$$

Khi chúng ta xét sự mở rộng của hàng đợi vật lý trong đường chính, $w\left(l, m^{side}\right) > w\left(l, m^{main}\right)$, chiến lược điều khiển sẽ hạn chế việc xếp hàng tràn ngược lại vì nó cấm nhiều phương tiện lao vào

giao lộ ở phía trước và cản trở sự di chuyển của các phương tiện trong các pha khác. Công thức (6) bây giờ có thể được đặt lại là: $M \leq \sum_{t=1}^{T} \sum_{(l,m)} x_{max}(m)$.

5.2.2. Connection to throughput maximization and travel time minimization - Kết nối để tối đa hóa thông lượng và giảm thiểu thời gian di chuyển

Cho biết trước rằng một quá trình chuyển động giao thông của một giao lộ là ổn định, do đó hệ thống cũng ổn định. Trong đường chính mà không có quay đầu xe (U-turn), phương tiện di chuyển từ làn m sang làn l sẽ không đi ngược từ làn l về làn m, tức là, ta có $x(l, m)$ và $x(m, l)$, duy nhất chỉ một trong hai được tồn tại trong mạng. Từ đó, các hành động mà agent RL chọn sẽ không gây ra tình trạng kẹt xe hay chặn đứng mạng, do đó có thể tận dụng khoảng thời gian đèn xanh bật hiệu quả hơn. Trong khoảng chu kì T cho trước, agent RL có thể tạo ra thông lượng tối đa và tối thiểu hóa được thời gian di chuyển của tất cả phương tiện trong hệ thống.

6. EXPERIMENT

Các thực nghiệm sẽ được thực hiện trên CityFlow [20], một giả lập giao thông mã nguồn mở mà hỗ trợ cho bài toán điều hướng tín hiệu giao thông với quy mô lớn. Sau khi nạp dữ liệu vào bộ giả lập, một phương tiện sẽ di chuyển đến vị trị của nó dựa theo thiết lập của môi trường. Bộ giả lập sẽ cung cấp trạng thái (state) cho phương pháp điều hướng tin hiệu và vận hành các hành động từ các phương pháp điều khiển đó.

6.1. Dataset Discription

Cả dòng dữ liệu giao thông ảo (synthetic data) và dữ liệu đời thực (real-world data) đều được sử dụng. Trong bộ dữ liệu, mỗi phương tiện được mô tả bằng (o, t, d) với o là vị trí gốc, t là thời gian, d là điểm đến. Hai vị trí o và d đều là những vị trí trên mạng của đường đi. Bộ giả lập sẽ nhận dữ liệu giao thông làm dữ liệu đầu vào. Tất cả dữ liệu đều chứa làn hai chiều và làn động.

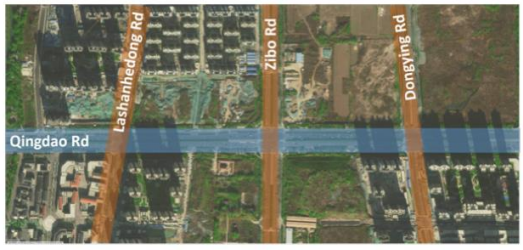
- **Dữ liệu ảo**: Bốn cấu hình khác nhau được test và thể hiện trong Bảng 2. Dữ liệu này được tổng hợp từ phân tích thống kê từ các mô hình giao thông ở Jinan và Hangxhou.
- **Dữ liệu đời thực**: Dữ liệu giao thông được thu thập từ ba thành phố để đánh giá hiệu suất của mô hình, bao gồm: Đại lộ Beaver ở State College, USA; Đường Quinyao ở Jinan, Trung Quốc; Bốn đại lộ ở Manhattan, thành phố New York, USA. Hình 4 minh họa cho các quang cảnh trên. Thống kê chi tiết sẽ được trình bày ở Bảng 3.

Bảng 2: Cấu hình của dữ liệu synthetic

Config	Demand Pattern	Arrival rate (phương tiện/giờ/đường)	Volume
1. Light-Flat	Flat	Đường chính: 600	(Light)
2. Light-Peak	Peak	Đường bên: 180	
3. Heavy-Flat	Flat	Đường chính: 1400	(Heavy)
4. Heavy-Peak	Peak	Đường bên: 420	

Bảng 3: Thống kê dữ liệu real-world

Dataset	Arrival rate (số phương tiện/giờ)				Số lượng giao lộ
	Mean	Std	Max	Min	
Quindao Rd., Jinan	3338.23	221.58	2748	3864	3
Beaver Ave., State College	2982.33	359.70	2724	3491	5
8-th Ave., NYC	6790.04	32.34	4968	7536	16
9-th Ave., NYC	4513.06	25.88	4416	6708	16
10-th Ave., NYC	6083.90	25.61	2892	5016	16
11-th Ave., NYC	4030.79	24.08	2472	4536	16



(a) Qingdao Road in Jinan, China:
a 3-intersection arterial with bidirectional traffic
on both the arterial and the side streets.

(b) Beaver Avenue in State College, Pennsylvania, USA:
a 5-intersection arterial with unidirectional traffic on the arterial
and bidirectional traffic on the side streets.

(c) 8-th, 9-th, 10-th and 11-th Avenue in New York City, USA:
four 16-intersection arterials with uni-directional traffic
on both the arterial and the side streets.

Hình 4. Minh họa cho dữ liệu đời thực (real-world data)

6.2. Experimental Settings

6.2.1. Enviromental Settings

Mỗi mạng lưới đường khác nhau đều được cấu hình. Tốc độ di chuyển trên các làn được đặt ở mức 40km/h. Phương tiện luôn luôn có thể rẽ phải nếu không có tắt nghẽn giao thông.

6.2.2. Evaluation Metric

Thời gian di chuyển trung bình (average travel time) ở đại lượng giây sẽ là thước đo để đánh giá mô hình. Độ đo được tính bằng cách tính trung bình thời gian di chuyển của toàn bộ phương tiện trong hệ thống.

6.3. Kết quả thực nghiệm

Bảng 4: Kết quả thực nghiệm của PressLight

	Dữ liệu synthetic				Dữ liệu real-world					
	LightFlat	LightPeak	HeavyFlat	HeavyPeak	Qingdao Rd., Jinan	Beaver Ave., State College	8th Ave., NYC	9th Ave., NYC	10th Ave., NYC	11th Ave., NYC
PressLight	59.96	61.34	160.84	184.51	54.87	92.00	223.36	149.01	161.21	140.82

7. CONCLUSION

Đồ án cuối kì của nhóm em nhằm mô tả và đánh giá được hiệu năng của một phương pháp học tăng cường (Reinforcement Learning method), cụ thể là PressLight. Trong đó, những kiến thức nền tảng và ứng dụng của học tăng cường được trình bày một cách rõ ràng. Trong PressLight, điểm thưởng (reward) và trạng thái (state) đều được tái thiết lại nhằm tường minh và hiệu quả hơn để giải quyết bài toán điều hướng giao thông. Đây sẽ là một nền tảng đủ vững để có thể làm bước đệm cho công cuộc nghiên cứu bài toán Traffic Signal Control trong tương lai.

ACKNOWLEDGMENT

Sau một học kì CS106 vô cùng thú vị và bổ ích, chúng em muốn dành một lời cảm ơn đến Thầy. Những chia sẻ, kinh nghiệm vô giá mà thầy đã mang đến trong học kì vừa qua sẽ giúp chúng em chuẩn bị tốt hơn những hành trang để tiếp tục trên con đường nghiên cứu và học tập sau này. Một lần nữa, em xin chân thành gửi một lời cảm ơn sâu sắc đến Thầy và chúc Thầy sẽ luôn mạnh khoẻ, luôn tràn đầy năng lượng để có thể tiếp tục truyền những ngọn lửa đam mê đến những thế hệ sau này như cách thầy đã làm với chúng em.

REFERENCES

[1] Monireh Abdoos, Nasser Mozayani, and Ana LC Bazzan. 2013. Holonic multi-agent system for traffic signals control. Engineering Applications of Artificial Intelligence 26, 5 (2013), 1575–1587.

[2] Baher Abdulhai, Rob Pringle, and Grigoris J Karakoulas. 2003. Reinforcement learning for true adaptive traffic signal control. Journal of Transportation Engineering 129, 3 (2003), 278–285.

[3] Itamar Arel, Cong Liu, T Urbanik, and AG Kohls. 2010. Reinforcement learning based multi-agent system for network traffic signal control. IET Intelligent Transport Systems 4, 2 (2010), 128–135.

[4] Elise Van der Pol and Frans A Oliehoek. 2016. Coordinated deep reinforcement learners for traffic light control. Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016).

[5] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. 2018. IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2496–2505.

[6] MA Wiering. 2000. Multi-agent reinforcement learning for traffic light control. In Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000). 1151–1158.

[7] Florence Boillot, Sophie Midenet, and Jean-Claude Pierrelee. 2006. The real-time urban traffic control system CRONOS: Algorithm and experiments. Transportation Research Part C: Emerging Technologies 14, 1 (2006), 18–38.

[8] Li Li, Yisheng Lv, and Fei-Yue Wang. 2016. Traffic signal timing via deep reinforcement learning. IEEE/CAA Journal of Automatica Sinica 3, 3 (2016), 247–254.

[9] Jean-Jacques Henry, Jean Loup Farges, and J Tuffal. 1984. The PRODYN real time traffic algorithm. In Control in Transportation Systems. Elsevier, 305–310.

[10] Xiaoyuan Liang, Xunsheng Du, Guiling Wang, and Zhu Han. 2018. Deep reinforcement learning for traffic light control in vehicular networks. arXiv preprint arXiv:1803.11115 (2018).

[11] Suvrajeet Sen and K Larry Head. 1997. Controlled optimization of phases at an intersection. Transportation science 31, 1 (1997), 5–17.

[12] Noe Casas. 2017. Deep Deterministic Policy Gradient for Urban Traffic Light Control. arXiv preprint arXiv:1703.09035 (2017).

[13] Li Li, Yisheng Lv, and Fei-Yue Wang. 2016. Traffic signal timing via deep reinforcement learning. IEEE/CAA Journal of Automatica Sinica 3, 3 (2016), 247–254.

[14] Seyed Sajad Mousavi, Michael Schukat, Peter Corcoran, and Enda Howley. 2017. Traffic Light Control Using Deep Policy-Gradient and Value-Function Based Reinforcement Learning. arXiv preprint arXiv:1704.08883 (2017).

[15] Guanjie Zheng, Yuanhao Xiong, Xinshi Zang, Jie Feng, Hua Wei, Huichu Zhang, Yong Li, Kai Xu, and Zhenhui Li. 2019. Learning Phase Competition for Traffic Signal Control. CoRR abs/1905.04722 (2019). arXiv:1905.04722

[16] Lior Kuyer, Shimon Whiteson, Bram Bakker, and Nikos Vlassis. 2008. Multi-agent reinforcement learning for urban traffic control using coordination graphs. Machine learning and knowledge discovery in databases (2008), 656–671.

[17] Samah El-Tantawy, Baher Abdulhai, and Hossam Abdelgawad. 2013. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto. IEEE Transactions on Intelligent Transportation Systems 14, 3 (2013), 1140–1150.

[18] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. Nature 518, 7540 (2015), 529.

[19] Pravin Varaiya. 2013. Max pressure control of a network of signalized intersections. Transportation Research Part C: Emerging Technologies 36 (2013), 177–195.

[20] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. 2019. CityFlow: A Multi-Agent Reinforcement Learning Environment for Large Scale City Traffic Scenario. (2019).