Xây dựng hệ thống điều khiển đèn giao thông bằng Deep Reinforcement Learning

Nguyễn Đỗ Đức Anh, Đỗ Lê Duy

Ngày 18 tháng 6 năm 2019

Mục lục

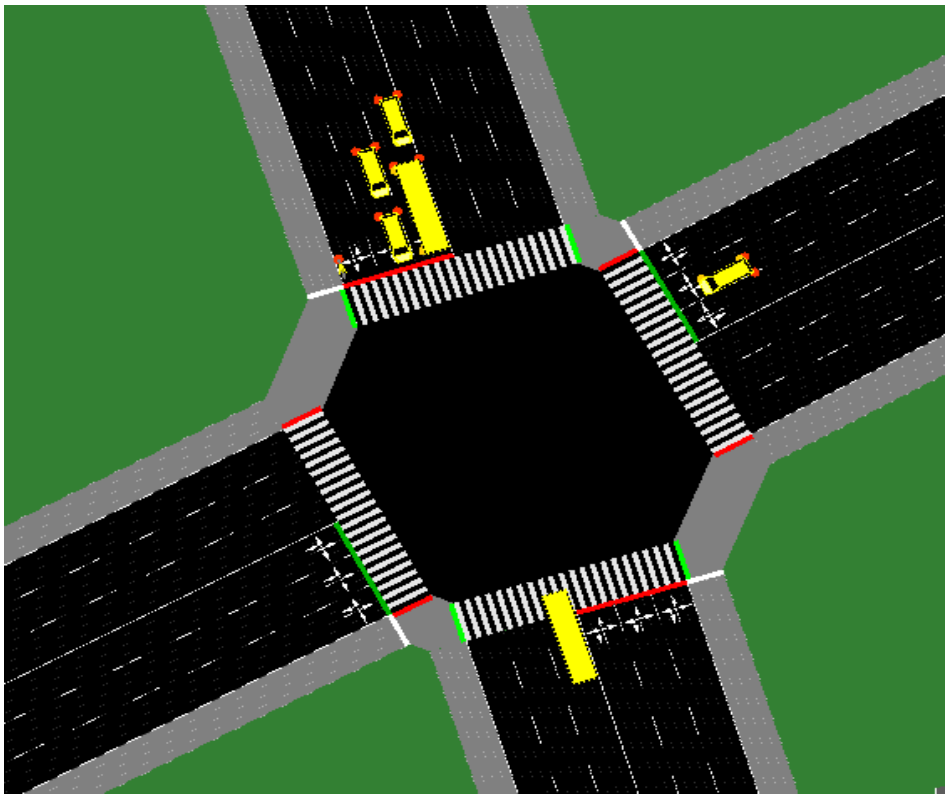
Mục lục

Trạng thái

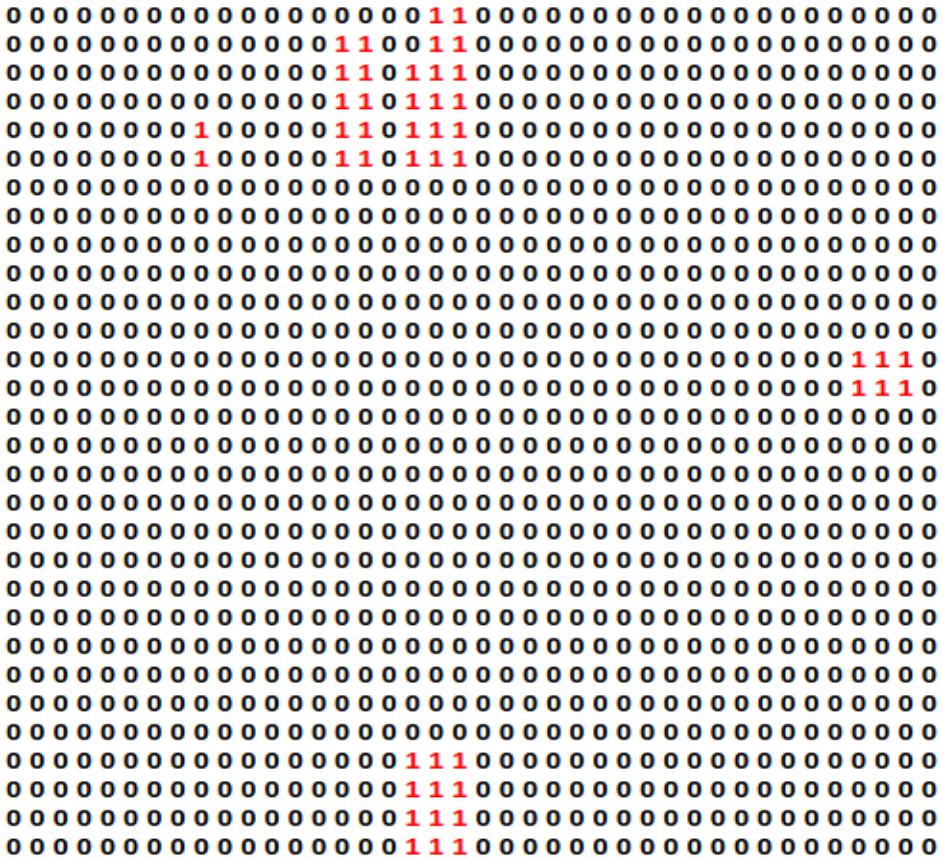
Mô hình DePGVF	**Mô hình DeTLC**
Trạng thái: • Tốn nhiều thời gian huấn luyện. • Không thử nghiệm trên thực tế được. • 128 x 128 x 4	**Trạng thái**: • Tốn nhiều chi phí tính toán xây dựng ma trận. • Có thể thử nghiệm trên thực tế. • 60 x 60 x 2

Trạng thái

Map 1-N:



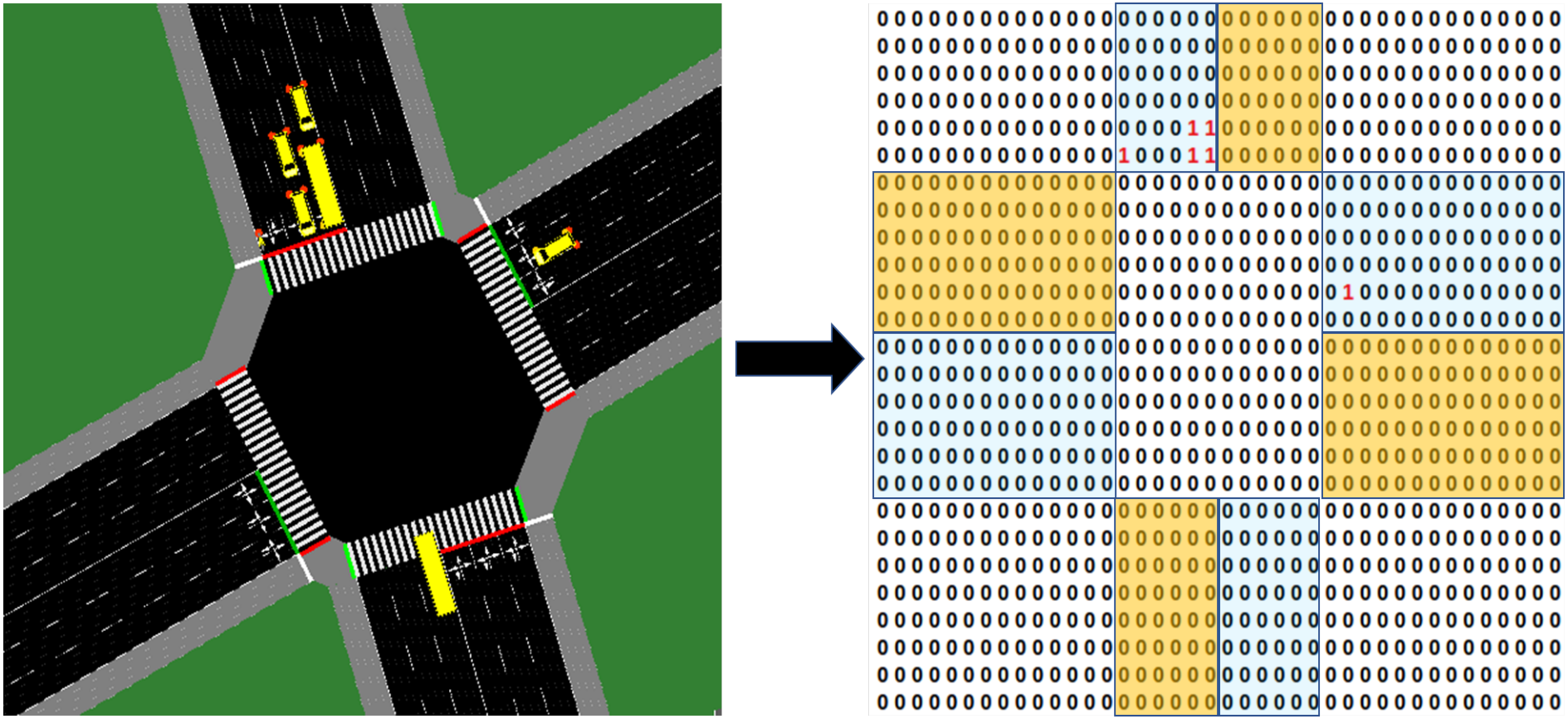
Hình: Ngã tư SUMO



Hình: Ma trận 1-N

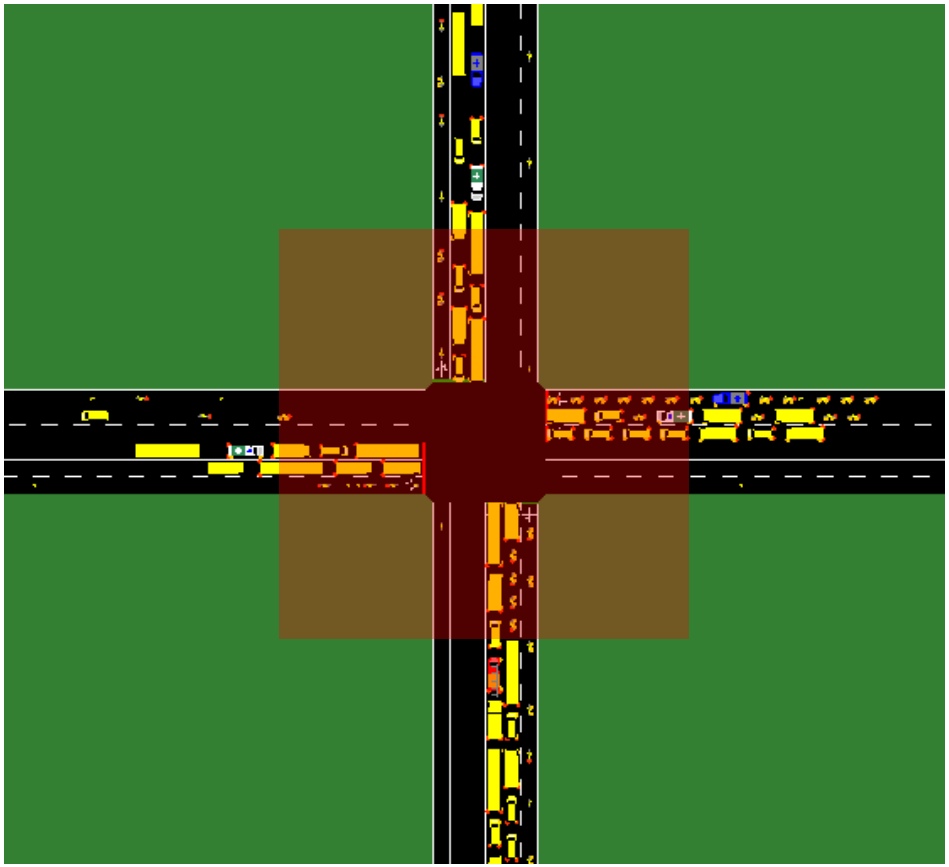
Trạng thái

Map 1-1:

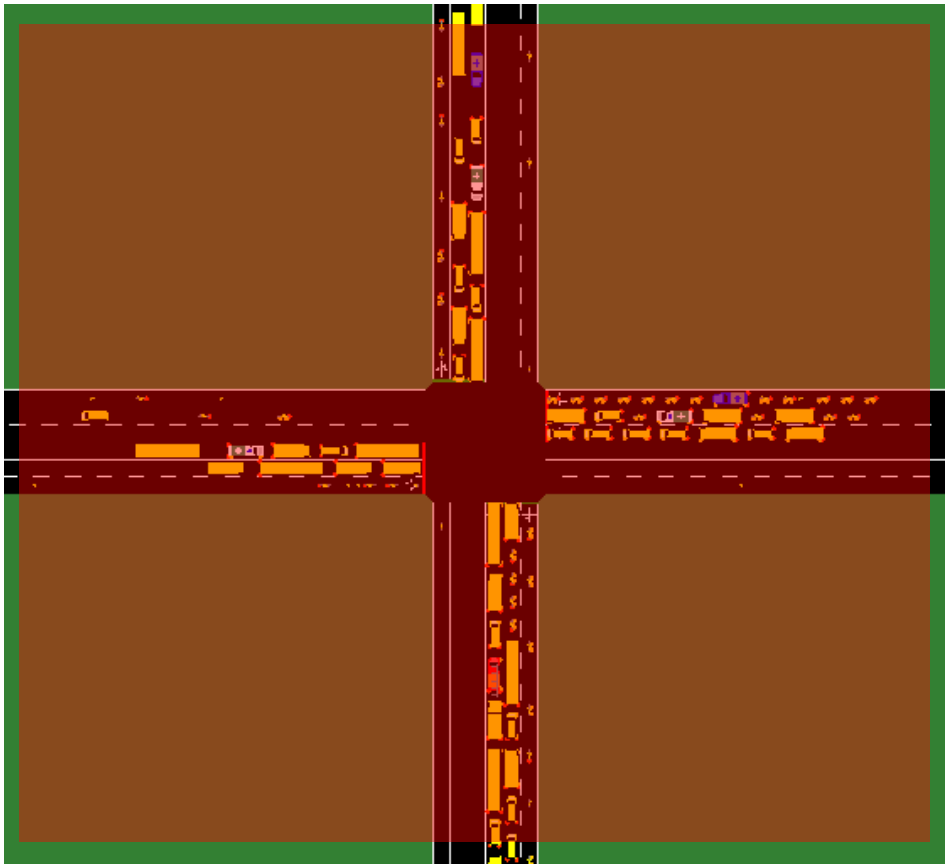


Hình: Ma trận 1-1

Trạng thái



Hình: Không gian ma trận map 1-N



Hình: Không gian ma trận map 1-1

Trạng thái


Số lượng xe được map ít
Map 1-N
Chi phí tính toán cao
Phân biệt kích thước, loại xe.

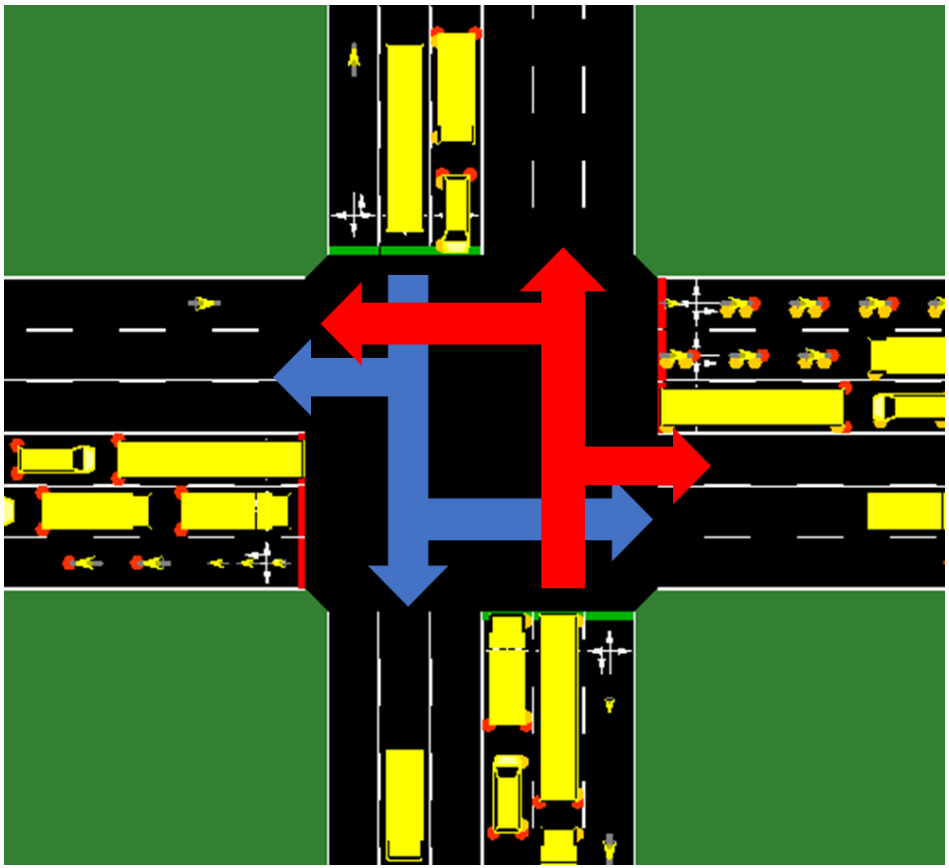

Số lượng xe được map nhiều
Map 1-1
Ít chi phí tính toán
Không phân biệt kích thước, loại xe

Mục lục

Hành động

Quy ước

1. **Status**: trạng thái giao thông

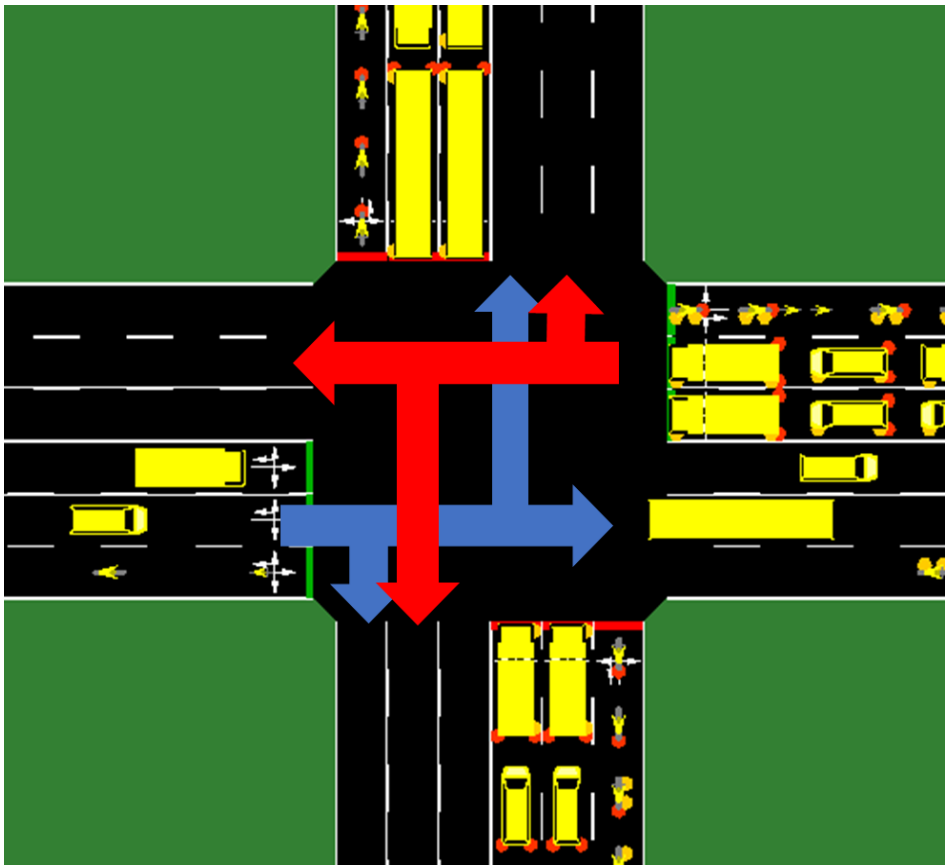


Hình: Bắc - Nam

Hành động

Quy ước

1. **Status**: trạng thái giao thông



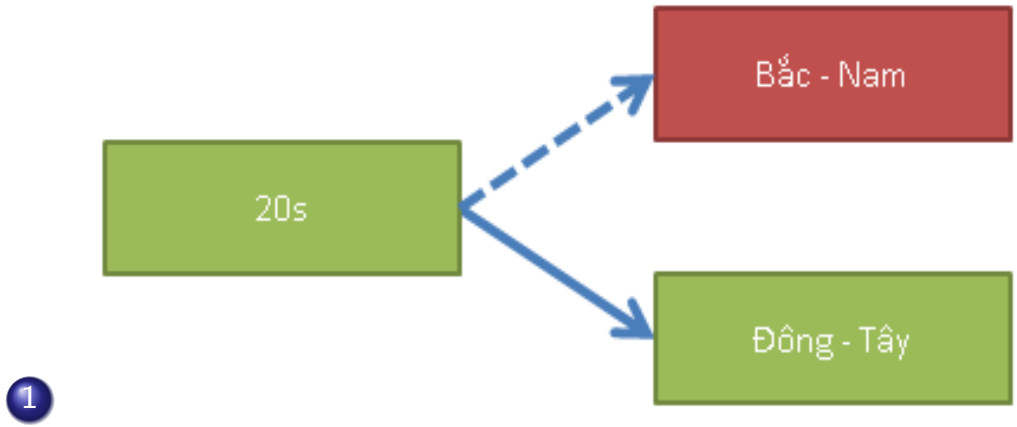
Hình: Đông - Tây

Hành động

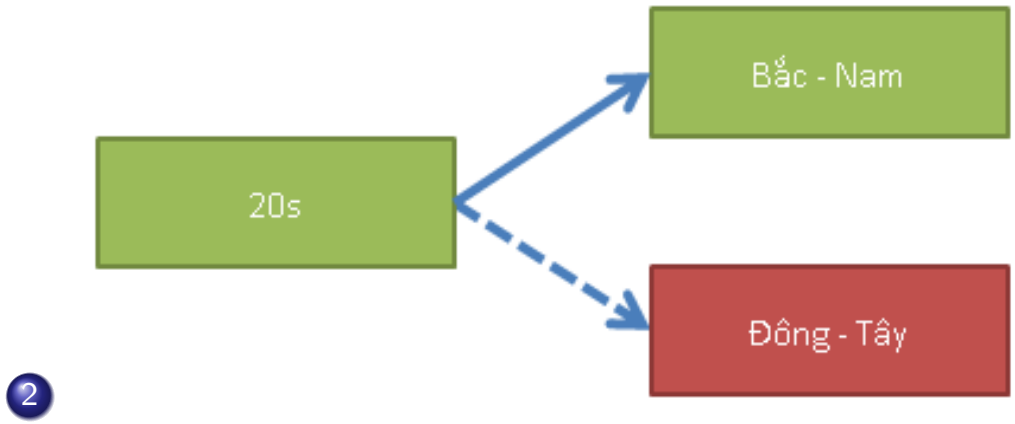
Quy ước

1. **Status**: trạng thái giao thông
2. **Phase**: Thời gian diễn ra của một status.
3. **Chu kì**: vòng liên tiếp các status.

Mô hình DePGVF

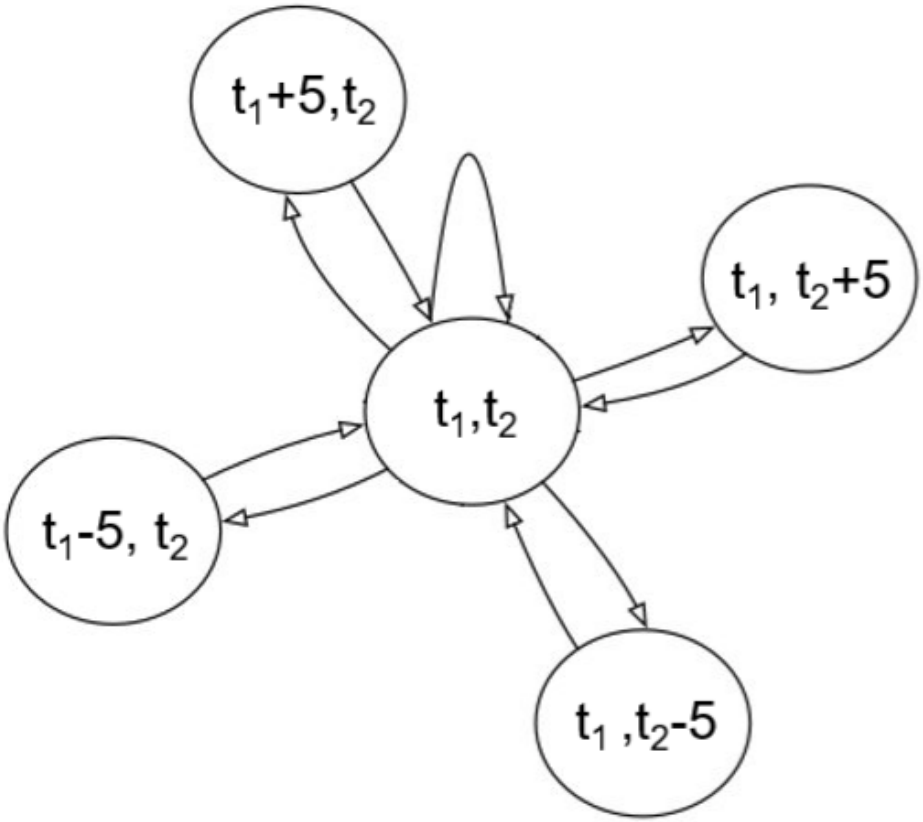


Hình: Chọn chiều Đông - Tây



Hình: Chọn chiều Bắc - Nam

Mô hình DeTLC



Hình: Chính sách 5 hành động

- $t_1 = 33, t_2 = 33 \rightarrow [33, 33]$
- action:
  $[[0, 0], [5, 0], [-5, 0], [0, 5], [0, -5]]$
- tentative action:
  $[[1, 1, 1, 1, 1], [1, 1, 0, 0, 0],$
  $[1, 0, 1, 0, 0], [1, 0, 0, 1, 0],$
  $[1, 0, 0, 0, 1]]$
- $0 \leq t \leq 60$

Hành động

Mô hình DePGVF	Mô hình DeTLC
Hành động: • Linh hoạt giải quyết tình huống. • Không biết trước thời gian chờ.	**Hành động**: • Biết trước thời gian chờ. • Thời gian ít thay đổi $\rightarrow$ an toàn và ổn định.

Hành động

Đèn vàng

Đảm bảo tính an toàn, ra hiệu cho xe dừng lại trước tín hiệu đèn đỏ.

$$T_{yellow} = \frac{v_{max}}{a_{dec}} \tag{1}$$

Mục lục

Phần thưởng

Mô hình DePGVF

t: chu kì thứ t

D_t: tổng thời gian chờ tức thời thời điểm t.

$$r_t = D_{t-1} - D_t \tag{2}$$

- $r_t > 0$: giảm tổng thời gian chờ giữa hai chu kì.
- $r_t < 0$: tăng tổng thời gian chờ giữa hai chu kì.
- $r_t = 0$: tổng thời gian chờ không thay đổi giữa hai chu kì.

Phần thưởng

Mô hình DeTLC

t: chu kì thứ t
W_t: tổng thời gian chờ từ $0 \rightarrow t$
N_t:tổng số xe từ $0 \rightarrow t$.

$$r_t = W_t - W_{t+1} \tag{3}$$

với

$$W_t = \sum_{i_t=1}^{N_t} w_{i_t,t} \tag{4}$$

- r_t luôn âm $\rightarrow$ Xe càng chờ lâu, tổng phần thưởng càng âm.
- $r_t = 0$ là giá trị cực đại của phần thưởng.

Phần thưởng

Mô hình DePGVF	Mô hình DeTLC
Phần thưởng: • Tổng phần thưởng chỉ là giá trị âm của tổng thời gian chờ của xe ở chu kì cuối cùng. • Không thể dùng để đánh giá.	**Phần thưởng:** • Tổng phần thưởng chỉ là giá trị âm của tổng thời gian chờ của tất cả xe tính từ thời điểm bắt đầu. • Có thể dùng để đánh giá.

Mục lục

Mạng nơ-ron

Deep-Q-Learning

$$Q_{target}(s,a) = r + \gamma \max_{a'} Q(s', a'; \theta^-). \quad (??)$$

Double Deep-Q-Learning

$$Q_{target}(s,a) = r + \gamma Q(s', \arg\max_{a'}(Q(s', a'; \theta)); \theta^-) \quad (5)$$

Mục lục

Thông số liên quan

Thông số	Giá trị
Kích thước Memory	20000
Kích thước minibatch	100
Starting ϵ	1
Ending ϵ	0
Số bước giảm từ Starting ϵ đến Ending ϵ	100 (eposide)
Số bước pre-training *tp*	0
Update rate α	0.001
Discount factor γ	0.75
Learning rate ϵ_r	0.0001
Leaky ReLU β	0.01

Bảng: Thông số mô hình hành động 2

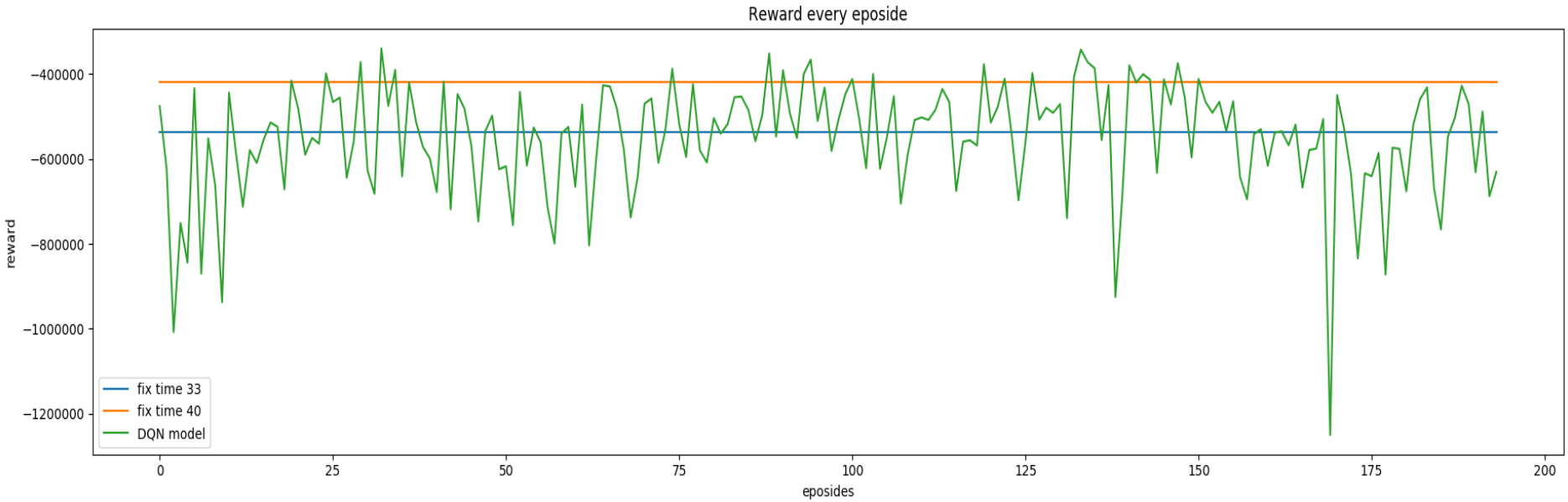
Thông số liên quan

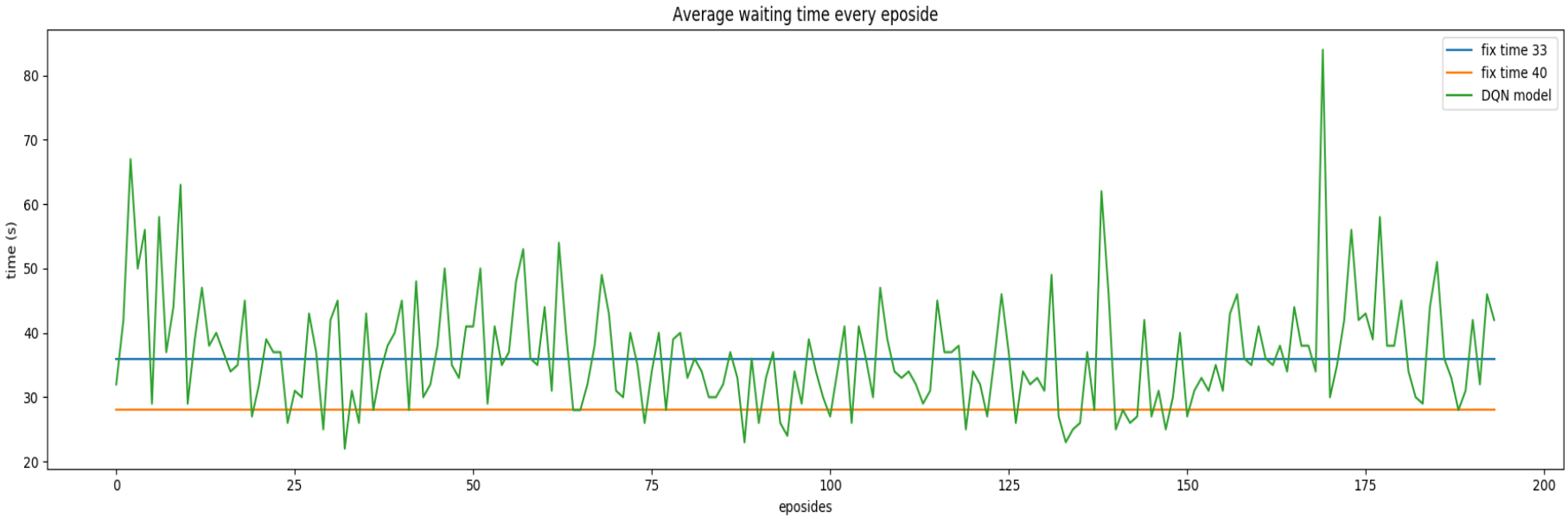
Thông số	Giá trị
Kích thước Memory	20000
Kích thước minibatch	64
Starting ϵ	1
Ending ϵ	0
Số bước giảm từ Starting ϵ đến Ending ϵ	10000 (bước huấn luyện)
Số bước pre-training t_p	2000
Update rate α	0.001
Discount factor γ	0.99
Learning rate ϵ_r	0.0001
Leaky ReLU β	0.01

Bảng: Thông số mô hình hành động 1

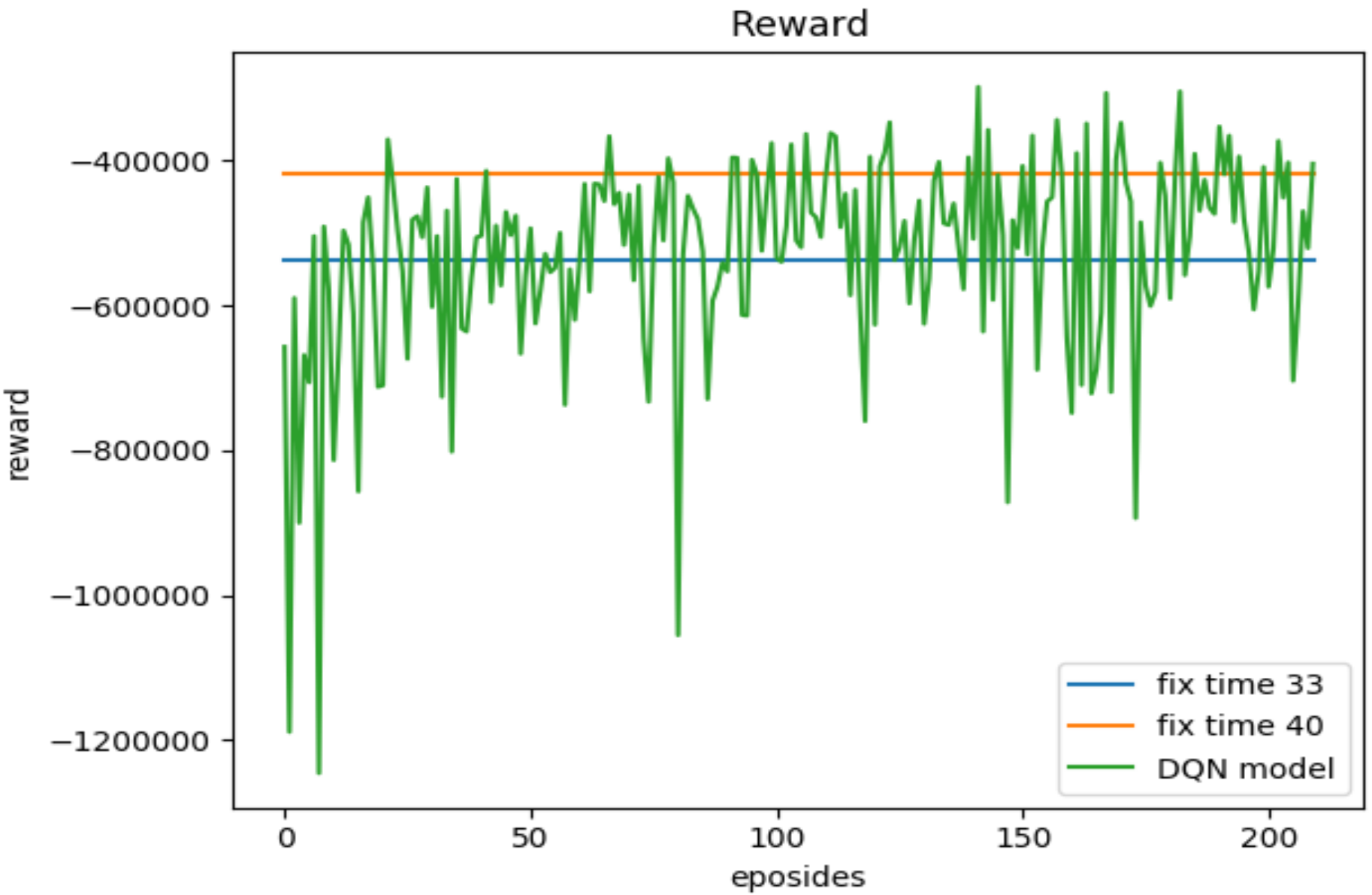
Mục lục

Kết quả huấn luyện mô hình DeTLC

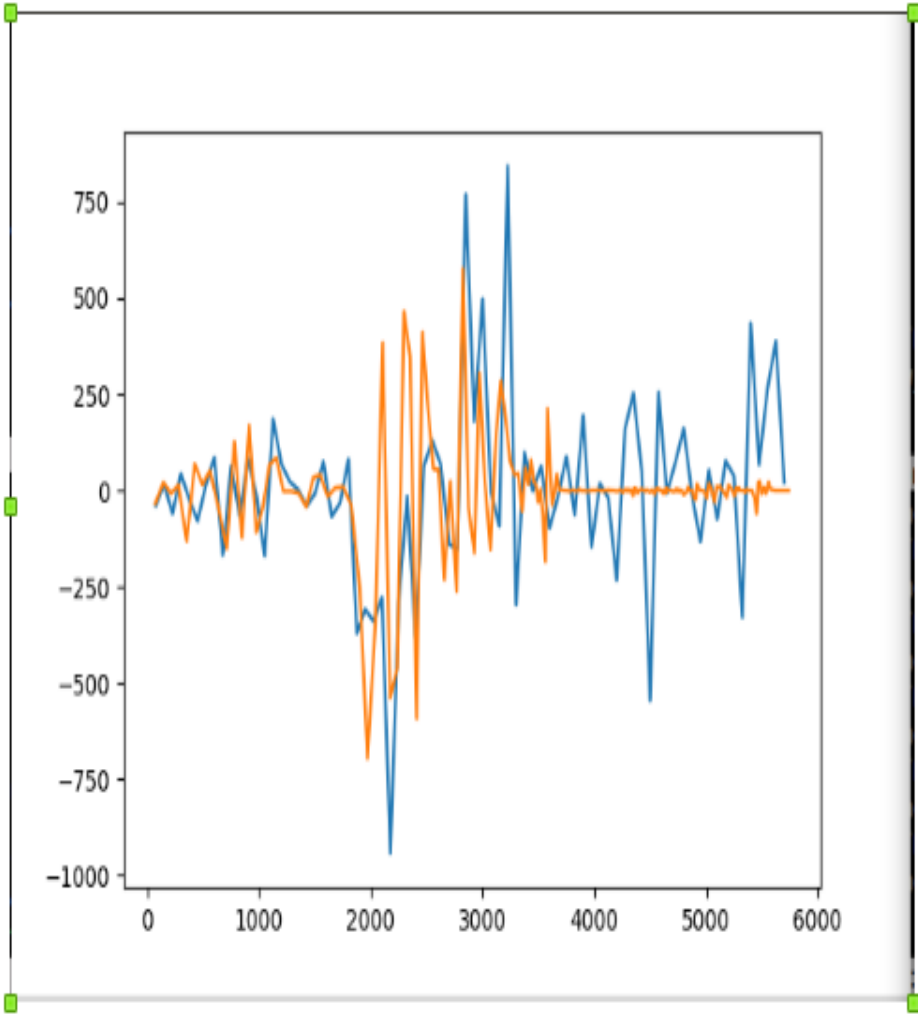


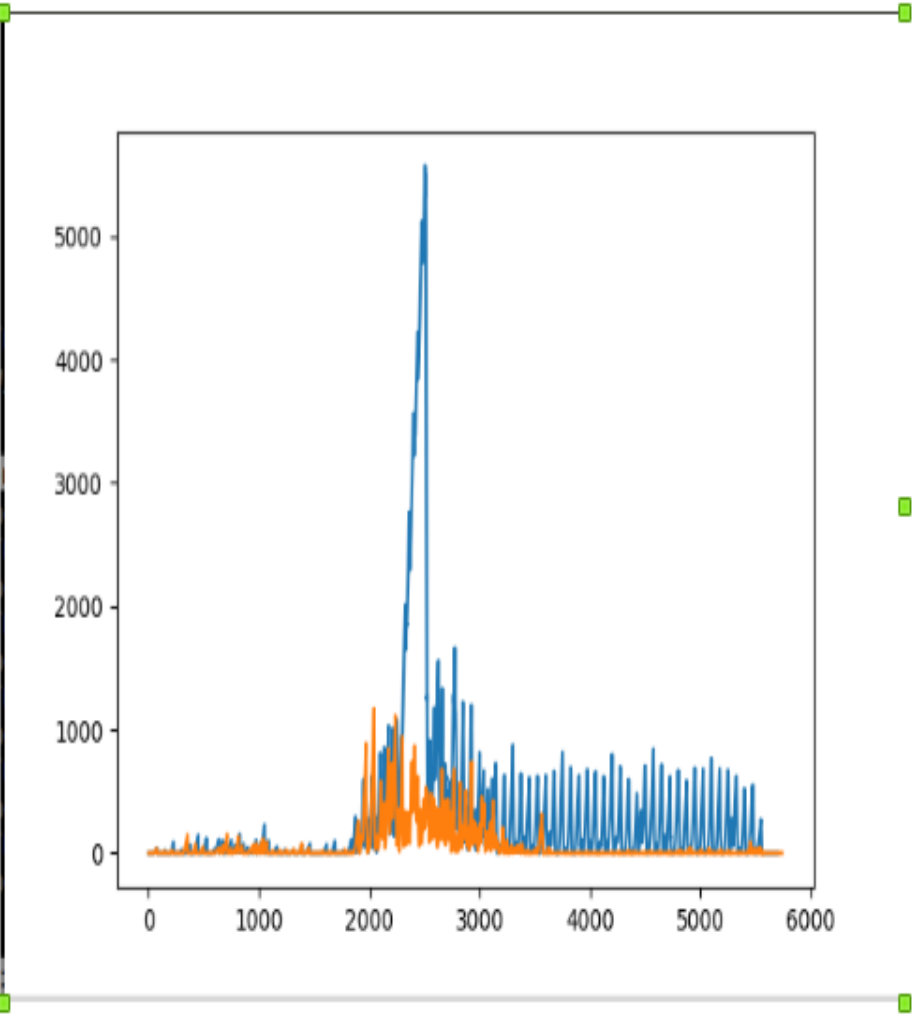


Kết quả huấn luyện mô hình DeTLC

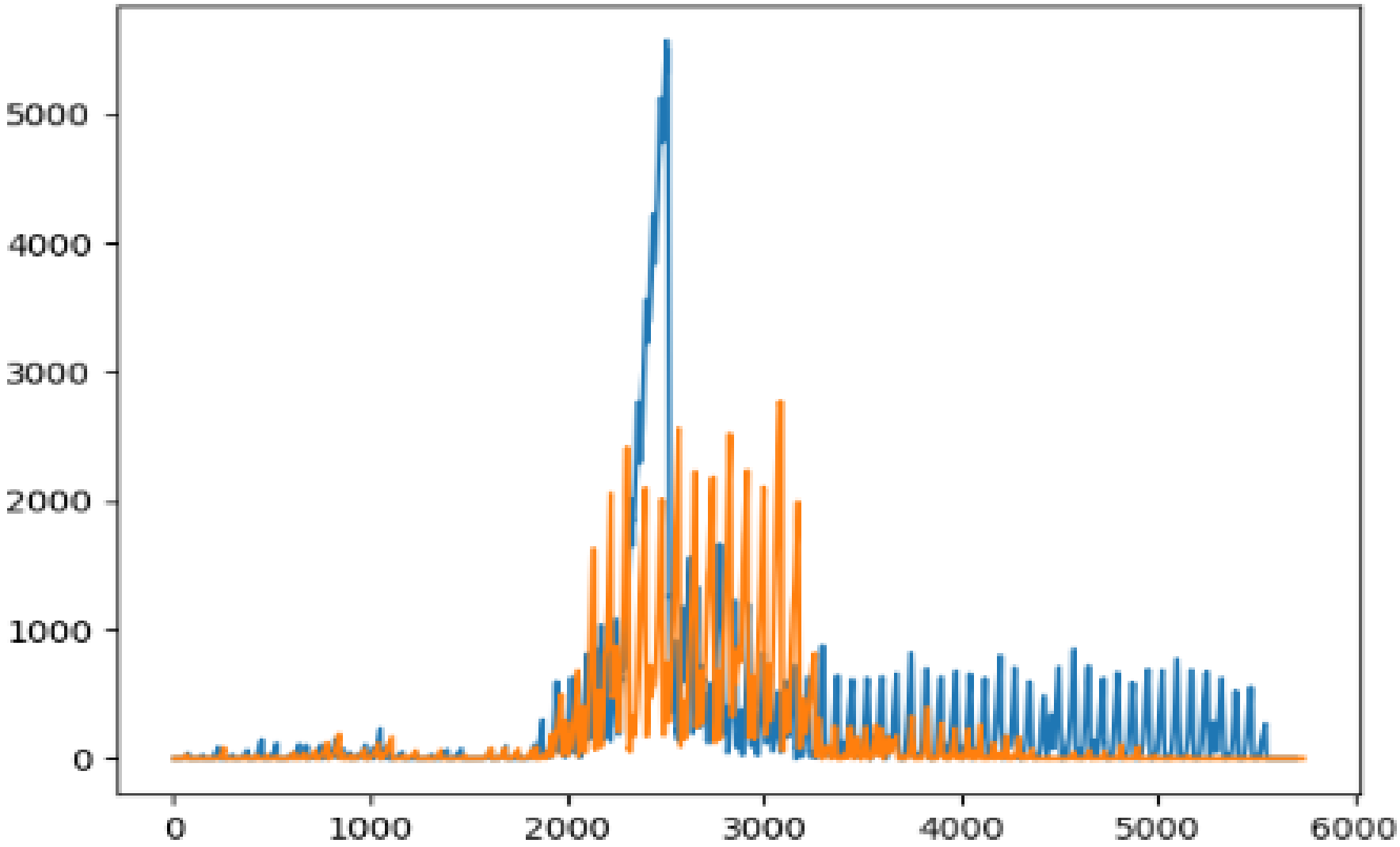


Kết quả huấn luyện mô hình DeTLC





Kết quả huấn luyện mô hình DeTLC



Kết quả huấn luyện mô hình DeTLC

Compare **TOTAL NEGATIVE REWARD:**

	Fixed System (33,4,33,4)	Model (20 green, 4 yellow)	Improvement
LOW	-4.165	-2.271	**45.5%**
HIGH	-69.740	-49.304	**29.3%**
NS	**-67.127**	**-15.095**	**77.6%**
EW	**-75217**	**-17.341**	**77%**

Compare **CUMULATIVE QUEUE LENGTH:**

	Fixed System (33,4,33,4)	Model (20 green, 4 yellow)	Improvement
LOW	12.685	8.580	**32.5%**
HIGH	1.159.607	955.999	**17.5%**
NS	447.550	135.826	**69.6%**
EW	262.515	121.607	**53.7%**