DQN Control Solution for CityBrain Challenge

Yitian Chen

Bigo Technology & Alibaba Group

August 19, 2021

Motivation: traffic congestion

Figure: **Tokyo**

- Become progressively more severe in many large cities.
- Significantly affect the lives of people.

- Is it because that the number of vehicles has exceeded the capacity of the city?
- Or we fail to utilize the road network at its maximum capacity?

Figure: **Beijing**

Task: traffic signals coordination

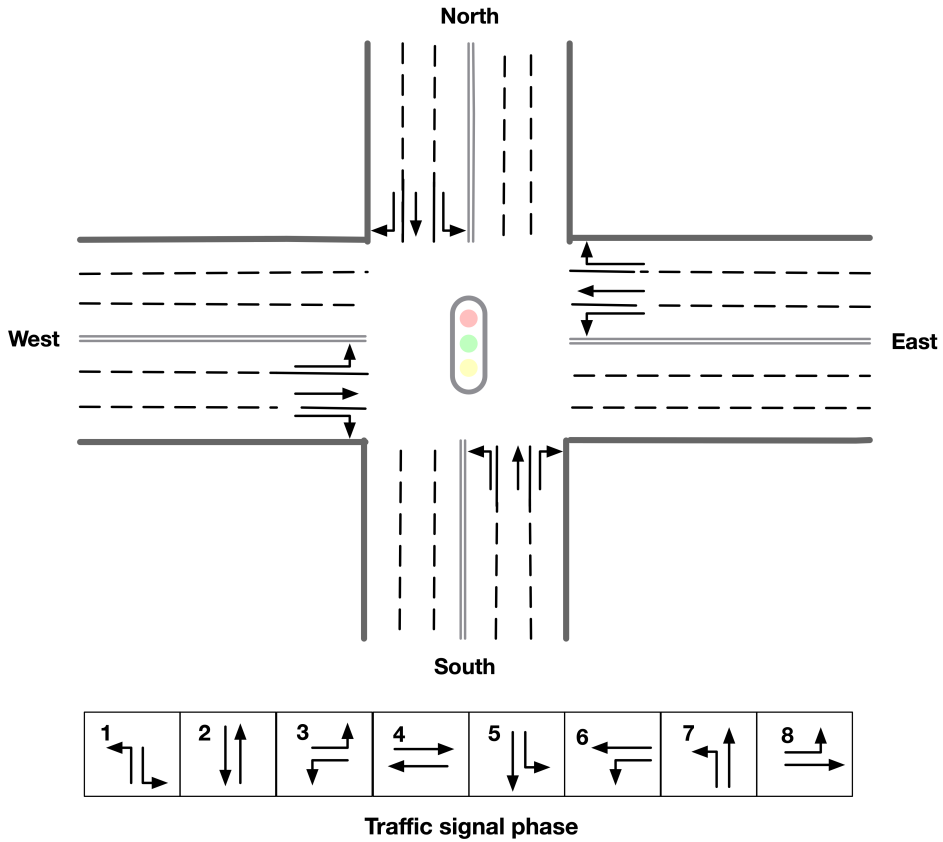
Task: Coordinating traffic signals with a self-designed agent to maximize the number of vehicles served while maintaining an acceptable delay.

Objective: Maximizing total number of vehicles served via optimizing traffic signal setting:

- For each intersection, select one from eight pre-defined signal phases every 10 seconds.

Environment:

1. A city-scale road network.
2. Traffic flow derived from real traffic data.



Challenges: complex real-world road network

1. The distribution of road lengths ranges from 37 meters to more than 4,000 meters.
2. Complex road scenarios: road with 3 (or 2) lanes; T intersection.
3. An action of traffic signal phase to be selected lasts for 10 seconds.
4. Challenges: how to extend the classical DQN framework [4, 1, 5, 7] to traffic signal control in a real-world road network with with unpredictable traffic flow changes?

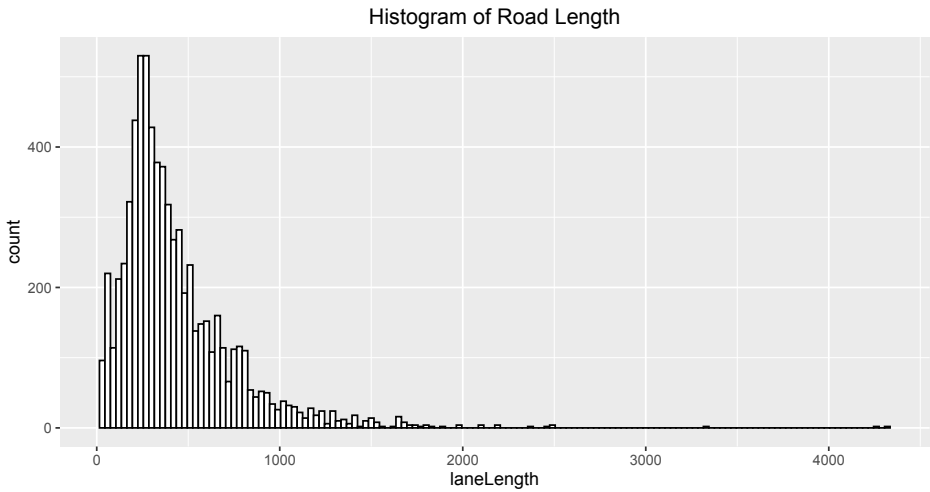


Figure: **Distribution of road lengths**

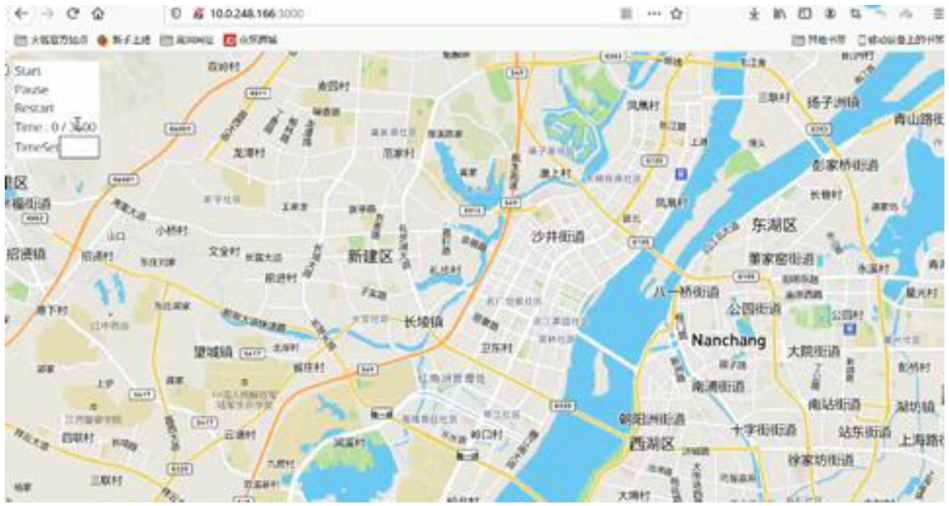


Figure: **Nanchang, Jiangxi Province**

Solution framework

1. How to extend the classical DQN framework [4, 1, 5, 7, 6] to traffic signal control in a real-world road network with complex traffic flows?
2. State design? Reward Design? Overall control scheme?

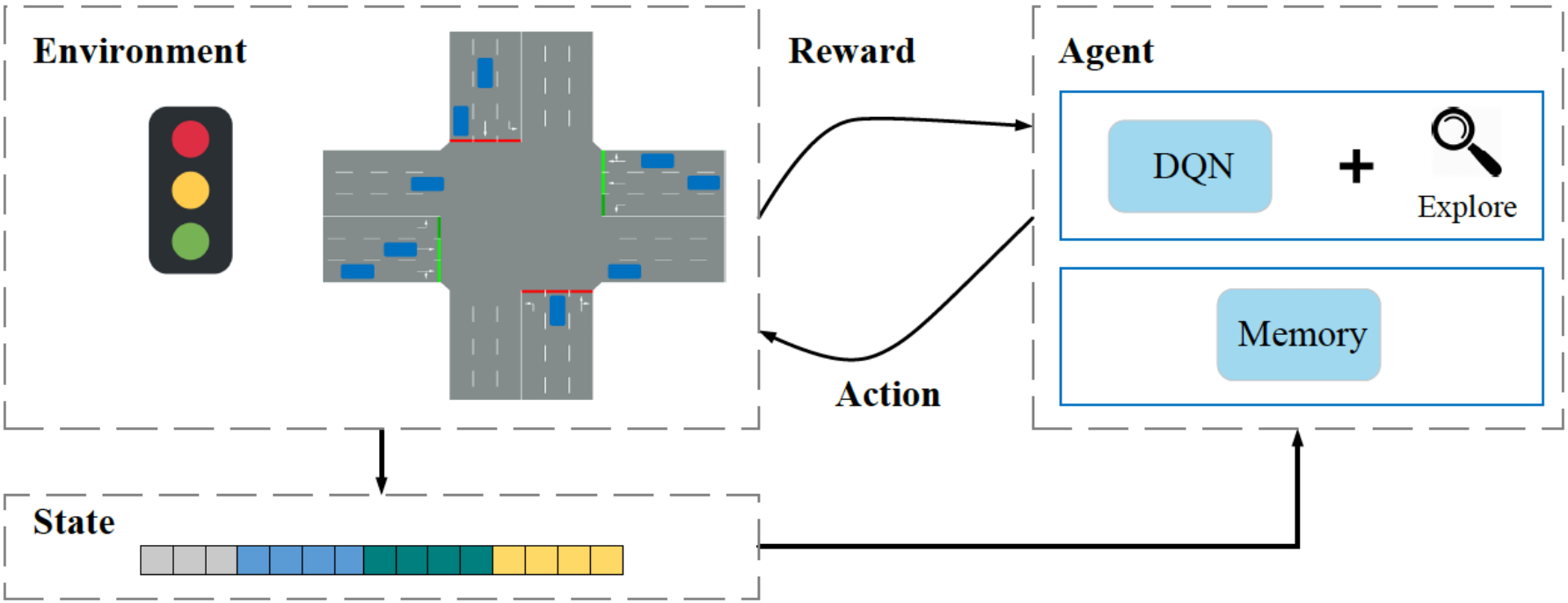


Figure: **Intellilight: DQN Control framework [6]**

Zone of influence

Define **Zone of influence** for the state and reward design:

1. For each intersection i, we define the segment of each lane where the distance to its center is less than k meters as the zone of influence with k and denote it as A_i^k.
2. Meanwhile, for each lane, the lane length is set as the distance capping of k.
3. We extract zone of influence with $k = 60$, $k = 100$, $k = 200$ for state design and $k = 100$ for reward design.

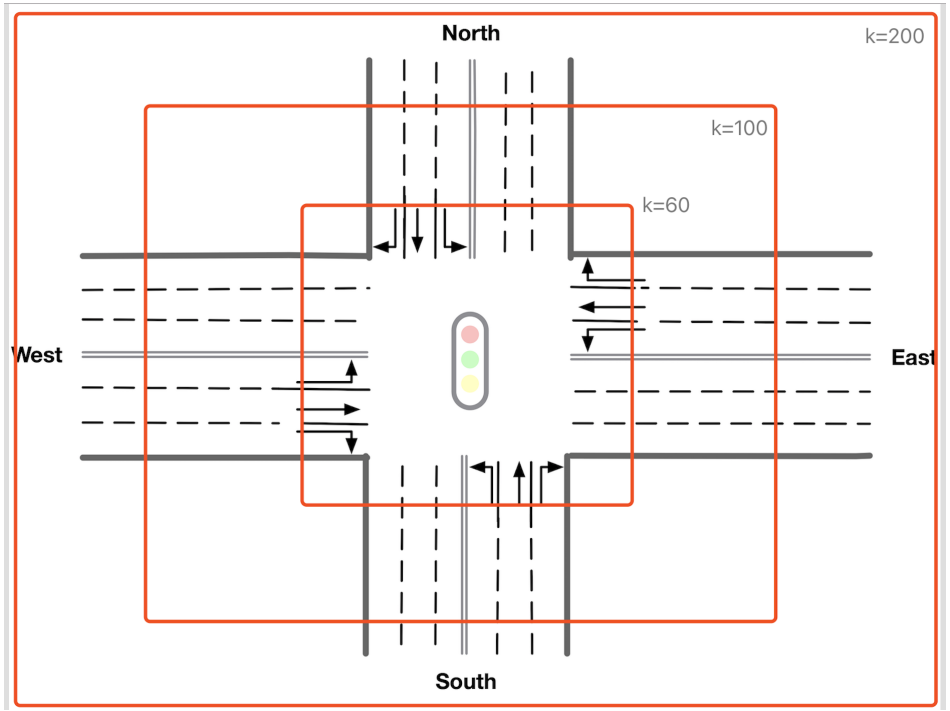


Figure: Zone of influence: we extract zone of influence with $k = 60$, $k = 100$, $k = 200$ for state design and $k = 100$ for reward design.

State design: features extraction by *zone of influence*

1. Statistic features extract from *zone of influence* with different k.
2. Get a better representation for all intersections with different road network scenarios using the same feature space.
3. Choose a maximum of k equal to 200, $k = 200$.

Features	Feature dimension		
	$k = 60$	$k = 100$	$k = 200$
Vehicle number	8	8	8
Delay index	8	8	8
Queue length	8	8	8
Pressure of vehicle number	8	8	8
Pressure of delay index	8	8	8
Pressure of queue length	8	8	8

Table: Statistic features used in the state design. For each pair (e.g., the feature "vehicle number" and zone of influence with distance $k = 60$), the feature dimension is 8, corresponding to 8 signal phases.

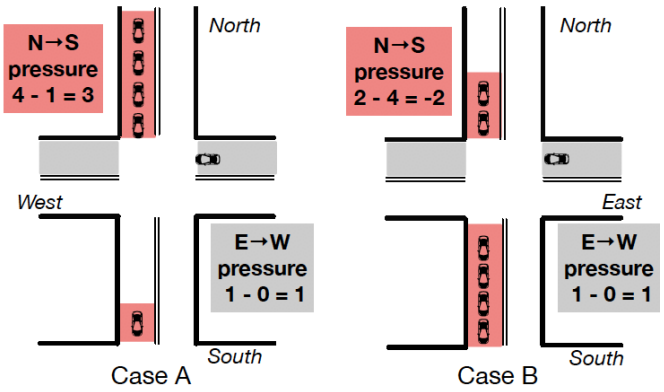
Reward design: classical approach

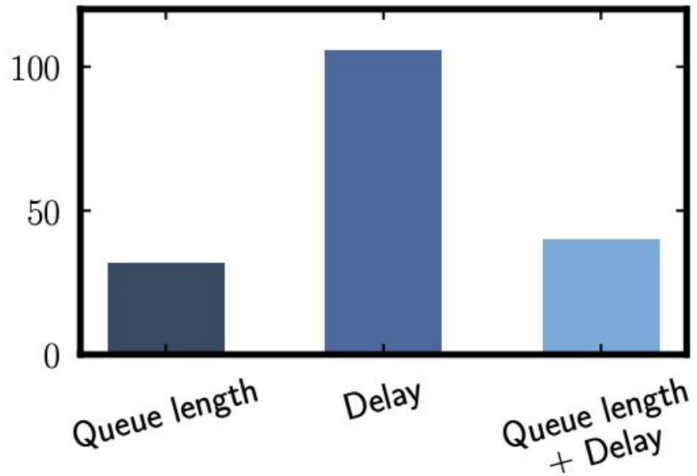
DQ: Delay+Queue length

- Queue length: the rewards of one intersection at time t is defined as the sum of all queue length at the next step $t + 1$.
- Delay: the same logic applied to the delay of the vehicle:

$$d(v) = 1 - \frac{speed(v)}{speedLimit(v)}$$

- DQ: Delay+Queue length.





Max pressure:

- A simple MP controller: Each time step (10s), actuate the signal phase with maximum difference in upstream and downstream queue lengths.

Newly-proposed reward function: Twin-DQ

Basic ideas: consider the rewards of upstream and downstream lanes separately:

1. Upstream: the actions of traffic signal affect the DQ of upstream lanes directly. Minimize the DQ of the upstream lanes.
2. Downstream: the DQ of downstream lanes can be more complicated. Minimize the difference between DQ at step t and DQ at the next step $t+1$.

$$R^{\text{Twin-DQ}}(t) = \underbrace{-\sum_{j=0}^{11}(d_j^{t+1}(A_i^k) + q_j^{t+1}(A_i^k))}_{Upstream}$$
$$\underbrace{-\sum_{j=12}^{23}(d_j^{t+1}(A_i^k) + q_j^{t+1}(A_i^k)) - (d_j^{t}(A_i^k) + q_j^{t}(A_i^k))}_{Downstream}$$

Model framework and training scheme

- The training framework follows the classical paradigm of Double Q-Network (double DQN) [3, 2] which includes an online network and a target network.
- The number of output dimensions is 8, corresponding to 8 candidate signal phases to be selected.
- The ϵ-greedy policy is adopted to train the model.
- The value of γ is set to 0.8, which aims at maximizing the total rewards of the next 5 rounds.

Parameter	Value
greenSec	20 seconds
Gamma	0.8
Model update frequency	1
Target model update frequency	17
Learning rate	5e-5
epsilon	0.2
epsilonMin	0.01
Loss function	Huber loss (smooth-L1)

Control scheme

Default control scheme for phase selection:

- Default setting: the green time exceeds 30 (or 20) seconds.
- Cons: it's observed that there are situations in which current phase goes with no vehicles in upstream lanes.

A more sophisticated approach:

Condition 1 : The green time of the current signal phase exceeds 30 seconds.

Condition 2 : The total queue length of the current signal phase of upstream lanes equals zero (zone of influence with $k = 60$).

Condition 3 : The total queue length of the current signal phase of downstream lanes is ≥ 8 (zone of influence with $k = 60$).

Condition 4 : The queue pressure of the current signal phase is ≤ -5 (zone of influence with $k = 60$).

Warning: The DQN agent is allowed to choose the current phase.

Experiment: reward function-control scheme

Reward function	TP1	TP2	TP3
DQ	46,619/1.400	48,201/1.409	48,201/1.403
Pressure	47,747/1.403	48,201/1.402	48,201/1.400
Twin-DQ	47,747/1.400	48,201/1.400	48,590/1.418

Table: Evaluation results of number of vehicles served and delay index of different reward function-control scheme pairs on the default round3_flow0 traffic data.

1. TP1: TP1 is the same as condition 1: the green time of the current signal phase exceeds 30 seconds.
2. TP2: TP2 includes the condition 1 and condition 2. If one of these two conditions above is met, the DQN-agent is triggered to begin the phase selection process.
3. TP3: TP3 includes all conditions of control scheme, if any of these conditions is met, the DQN-agent is triggered to begin the phase selection process.

Experiment: leader-board results

Strategy	Number of vehicles served	Delay index
Rule-based	313,035	1.402
DQN-single	313,035	1.403
DQN-ensemble	314,467	1.404
DQN-ensemble + Rule-based	317,954	1.401

Table: Experiment results in the leader-board. The rule-based agent is mainly based on the calculation of vehicle density near the center of the intersections. The "Twin-DQ" reward function is used to train the DQN model.

Conclusion

1. We present a DQN-based framework for the "City Brain Challenge" competition.
2. We described our overall analysis and details of the DQN-based framework for real-time traffic signal control.
3. Our main improvements come from two points:
   - A newly-proposed reward function:"Twin-DQ".
   - Control scheme: a suite of control conditions.
4. Applying heuristic rules to revise the DQN control actions in some cases can improve the performance of the DQN-based control framework.
5. The codes of our solution is available at Github: `https://github.com/oneday88/kddcup2021CBCBingo`

Thank you!

References I

Abdoos, M., Mozayani, N., and Bazzan, A. L.

Traffic light control in non-stationary environments based on multi agent q-learning.

In *2011 14th International IEEE conference on intelligent transportation systems (ITSC)* (2011), IEEE, pp. 1580–1585.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al.

Human-level control through deep reinforcement learning.

nature 518, 7540 (2015), 529–533.

Van Hasselt, H., Guez, A., and Silver, D.

Deep reinforcement learning with double q-learning.

In *Proceedings of the AAAI conference on artificial intelligence* (2016), vol. 30.

References II

Wei, H., Chen, C., Zheng, G., Wu, K., Gayah, V., Xu, K., and Li, Z.

Presslight: Learning max pressure control to coordinate traffic signals in arterial network.

In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019), pp. 1290–1298.

Wei, H., Zheng, G., Gayah, V., and Li, Z.

A survey on traffic signal control methods.

arXiv preprint arXiv:1904.08117 (2019).

Wei, H., Zheng, G., Yao, H., and Li, Z.

Intellilight: A reinforcement learning approach for intelligent traffic light control.

In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), pp. 2496–2505.

References III

Zheng, G., Xiong, Y., Zang, X., Feng, J., Wei, H., Zhang, H., Li, Y., Xu, K., and Li, Z.

Learning phase competition for traffic signal control.

In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (2019), pp. 1963–1972.