Multi-Agent Deep Reinforcement Learning for Traffic Signal Control

Paolo Fazzini

WINLOG @SMM 2019 20-21-22 November 2019 – Rende (CS) Calabria- Italy

Topics:

- Markov Decision Process
- Reinforcement Learning
- Multi-Agent Reinforcement Learning
- Deep Neural Networks
- Long Short-Term Memory Networks
- Sumo (Simulation of Urban Mobility)
- Adaptive Traffic Signal Control

Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Markov Decision Process

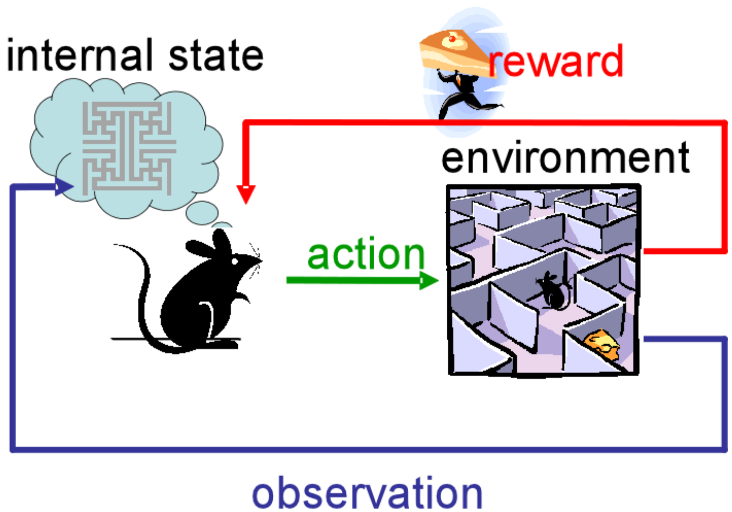
You want to go from the Church of St. Francis to the Belvedere.
Two paths take you there, but you don't know which path is the quickest.
We need to create a model to represent this problem.
This is called the Markov Decision Process.

$$P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$$

Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Reinforcement Learning

$$R_t = \sum_{i=t+1}^{\infty} \gamma_i \cdot r(s_i, a_i, s_{i+1})$$



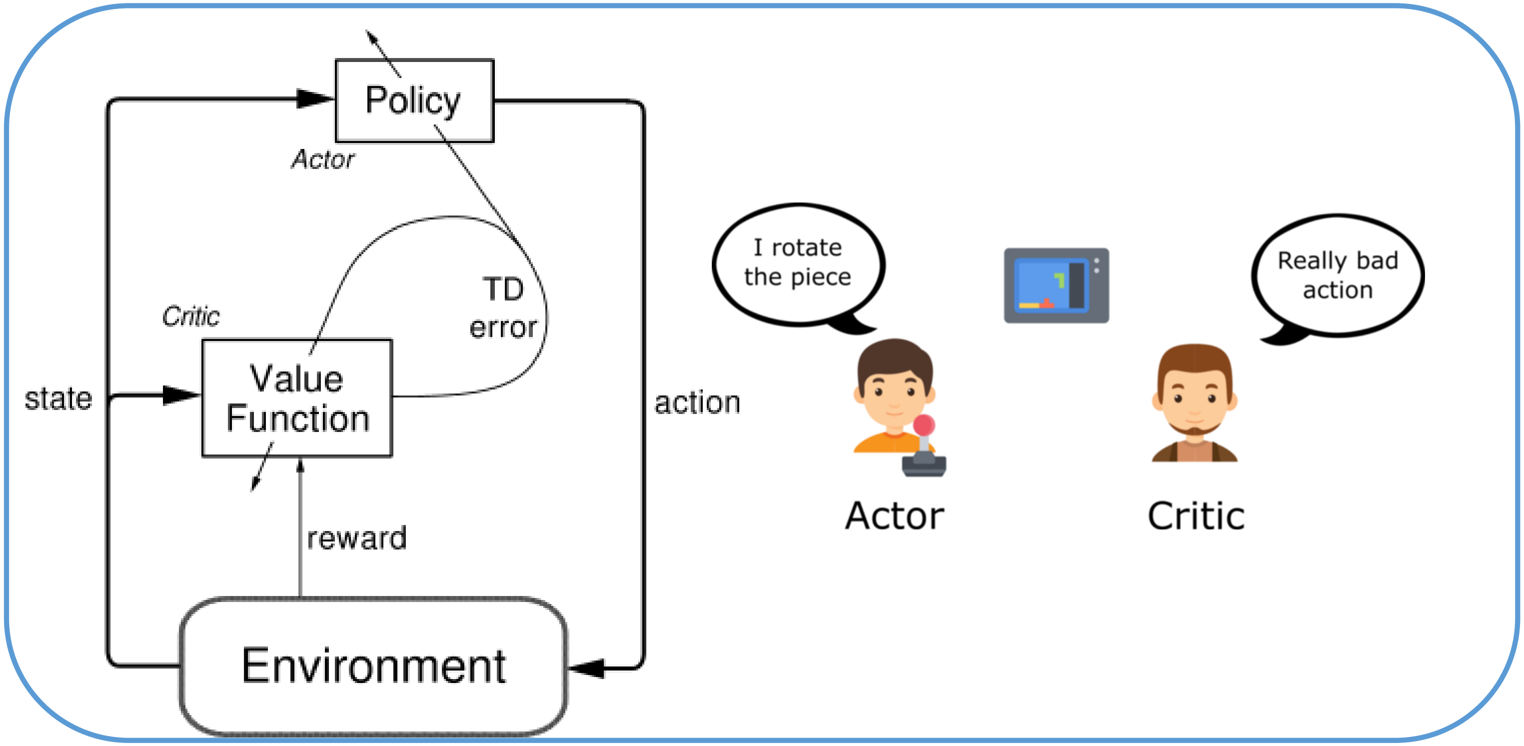
Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Reinforcement Learning

- SARSA
- Expected SARSA
- Q-Learning
- General Q-Learning
- QV-Learning
- Double Q-Learning
- Actor-Critic
- …

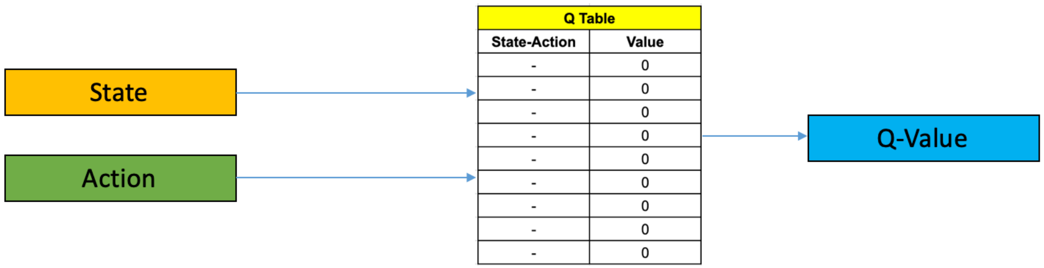
Reinforcement Learning

Advantage Actor-Critic (A2C)



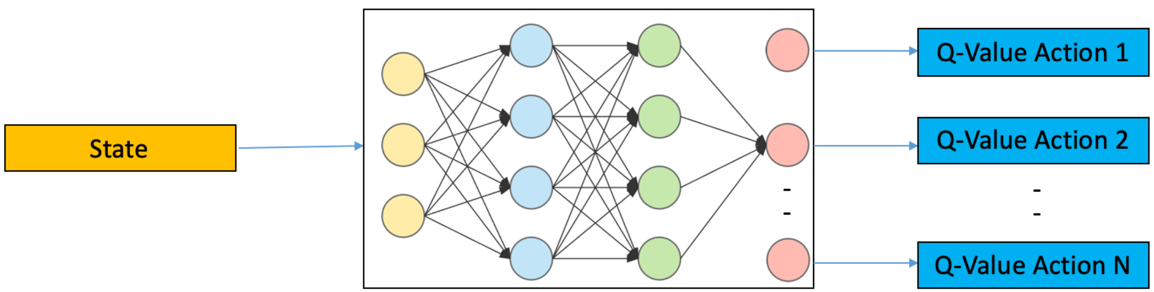
Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Deep Reinforcement Learning



Q Table	
State-Action	Value
-	0
-	0
-	0
-	0
-	0
-	0
-	0
-	0
-	0

Q Learning

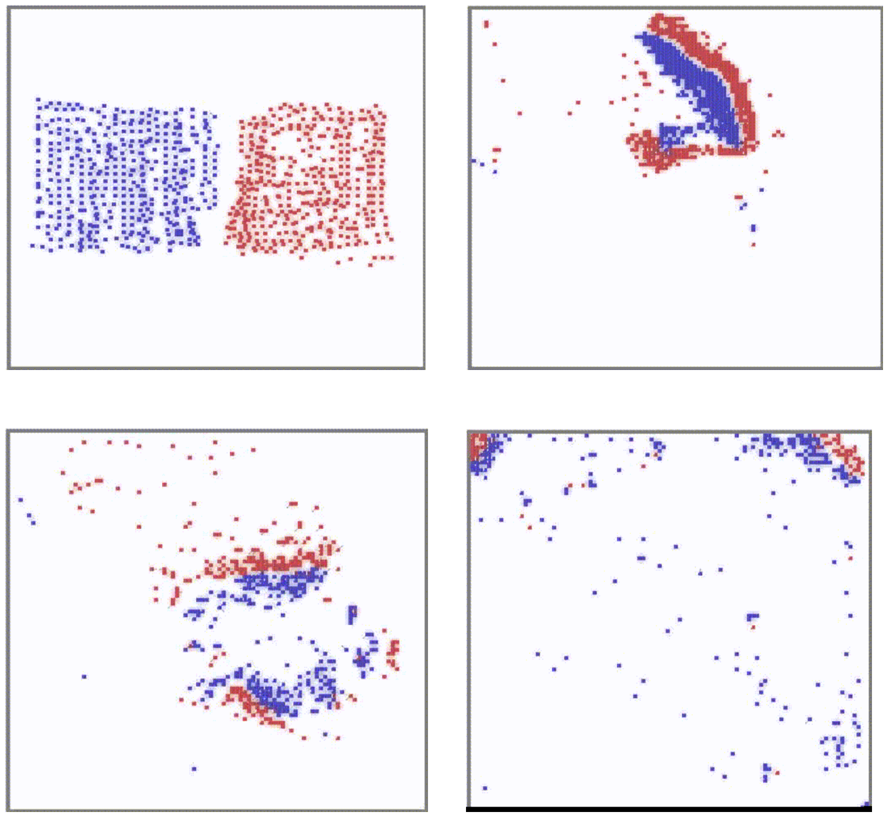


Deep Q Learning

Multi-Agent Reinforcement Learning

Type:

- Cooperative
- Competitive
- Mixed
-



Issues:

- Non Stationarity
- Partial Observability
- Training schemes
- Scalability

Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

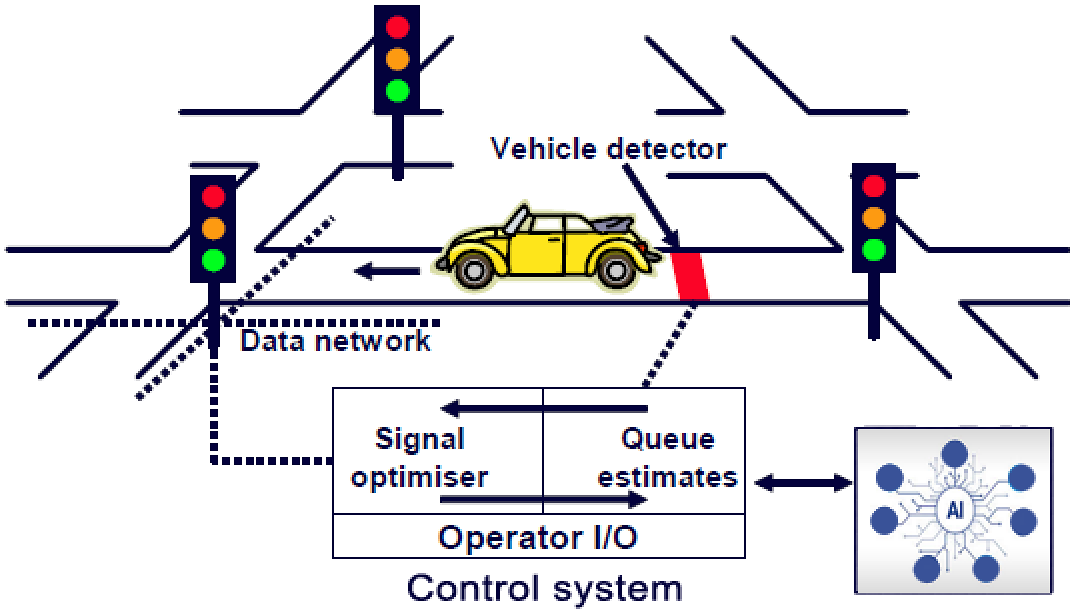
Multi-Agent Reinforcement Learning

tackling MARL with traditional RL is not straightforward. If all agents observe the true state we can model a cooperative multi-agent system as a single meta-agent. However, the size of this meta-agent's action space grows exponentially in the number of agents. Furthermore, it is not applicable when each agent receives different observations that may not disambiguate the state. Hence:

- Independent Deep Q-Learning (IDQL)
- Independent Deep Advantage AC (IA2C)
- Multi-agent Deep AC (MA2C)

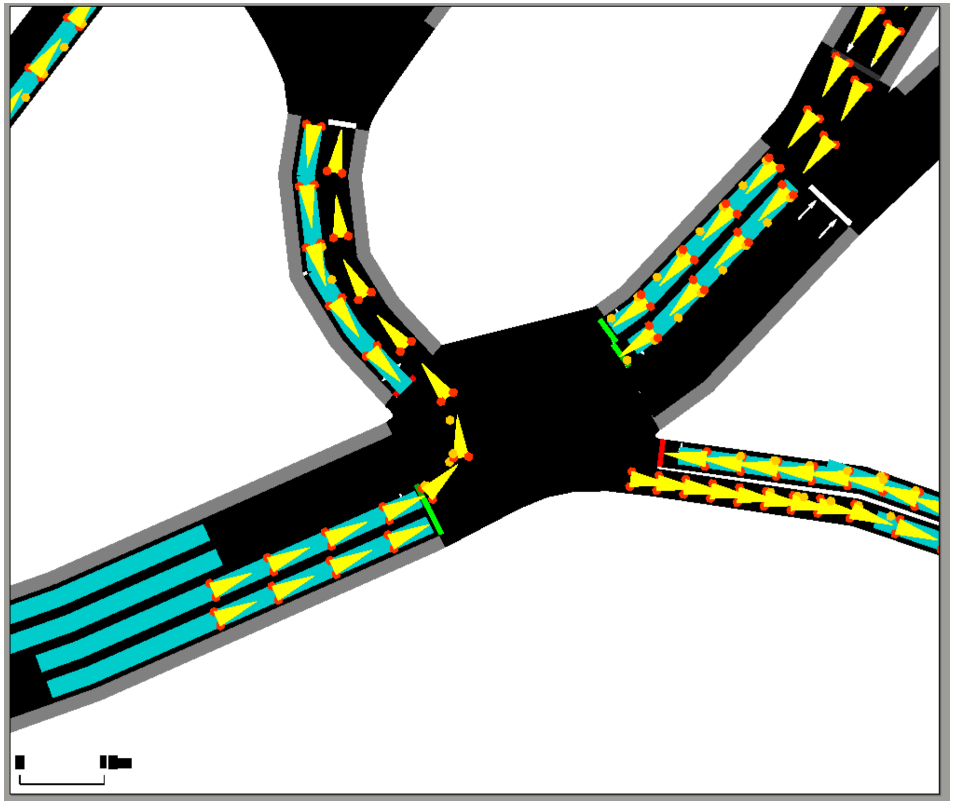
New challenges: now the environment becomes partially observable from the viewpoint of each local agent due to limited communication among agents

Traffic Signal Control



Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Traffic Signal Control



Traffic Signal Control

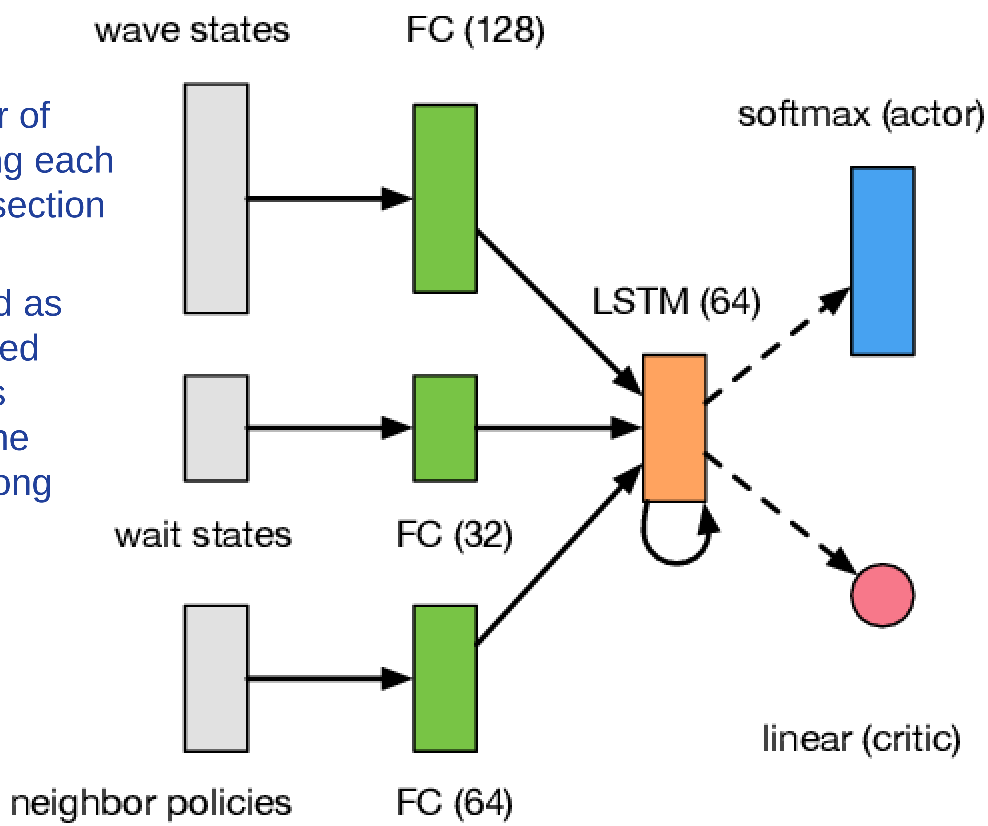
Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Traffic Signal Control

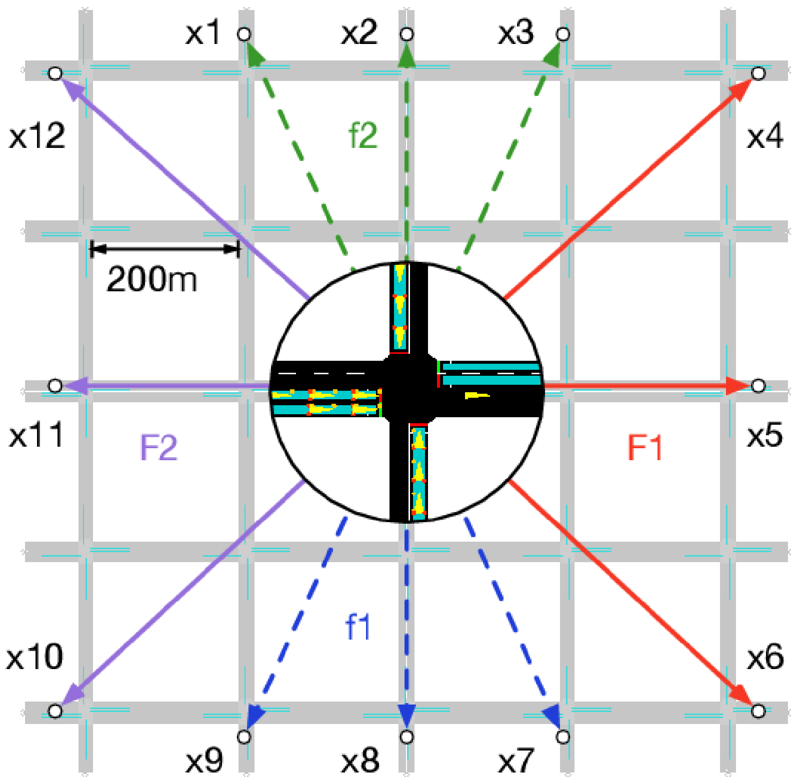
wave[veh.] measures the total number of waiting and approaching vehicles along each incoming lane, within 50m to the intersection

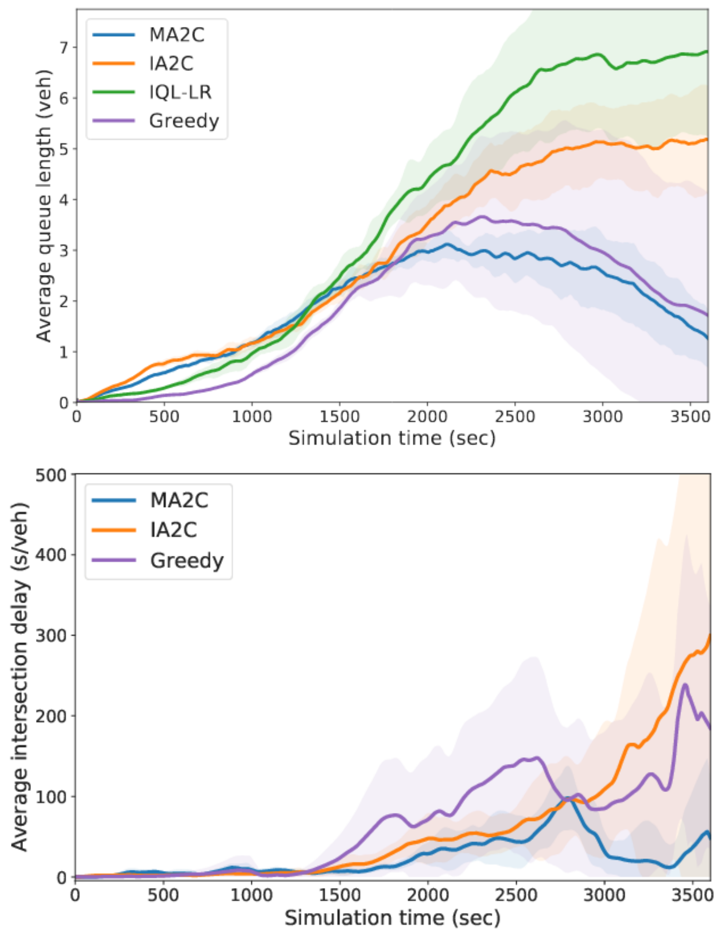
The waiting time of a vehicle is defined as the time (in seconds) spent with a speed below 0.1m/s since the last time it was faster than 0.1m/s. ***wait***[s]measures the cumulative delay of the first vehicle along each incoming lane

Neighbor policies are the policies or the closest Traffic Signal Controllers



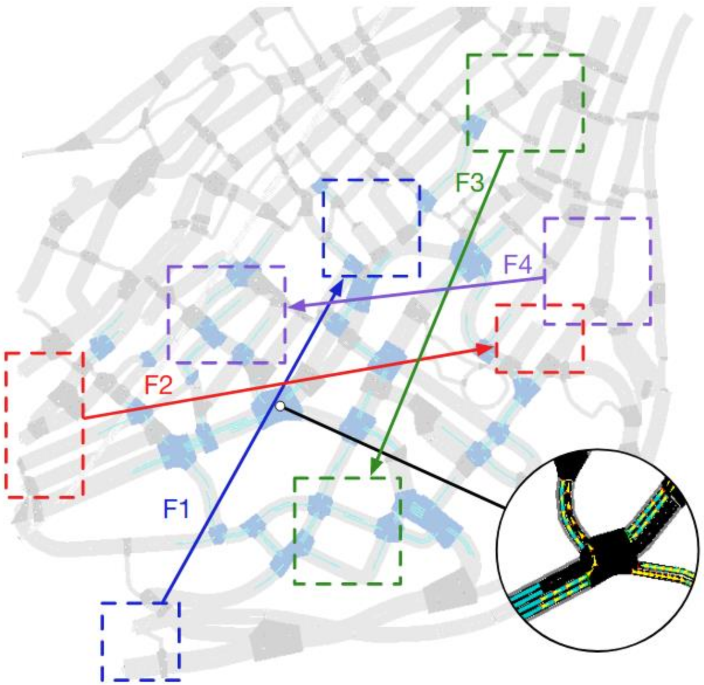
Traffic Signal Control

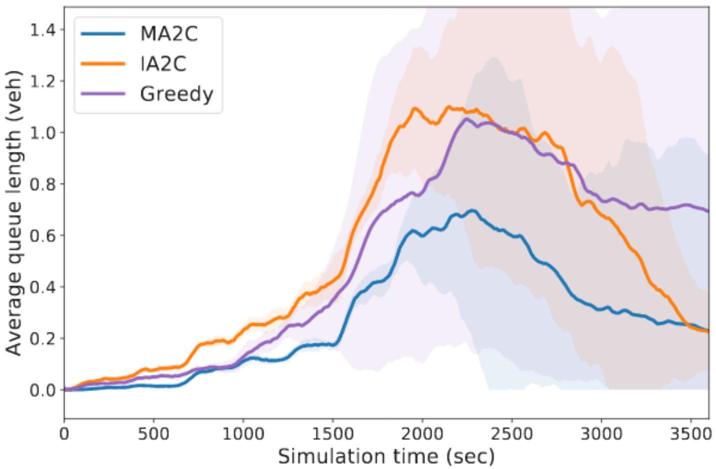


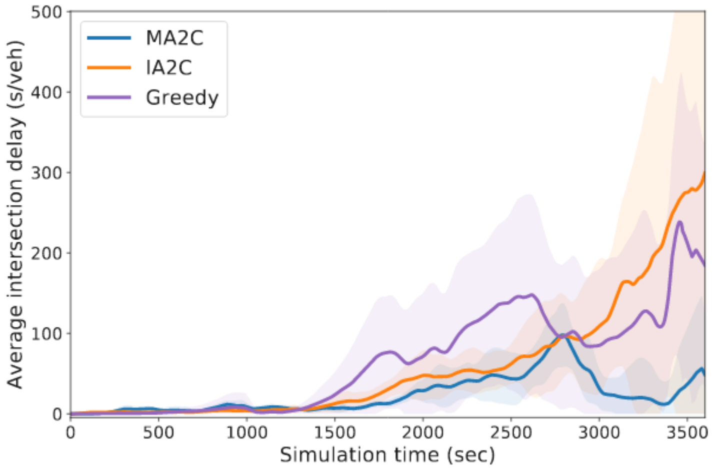


Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Traffic Signal Control







Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

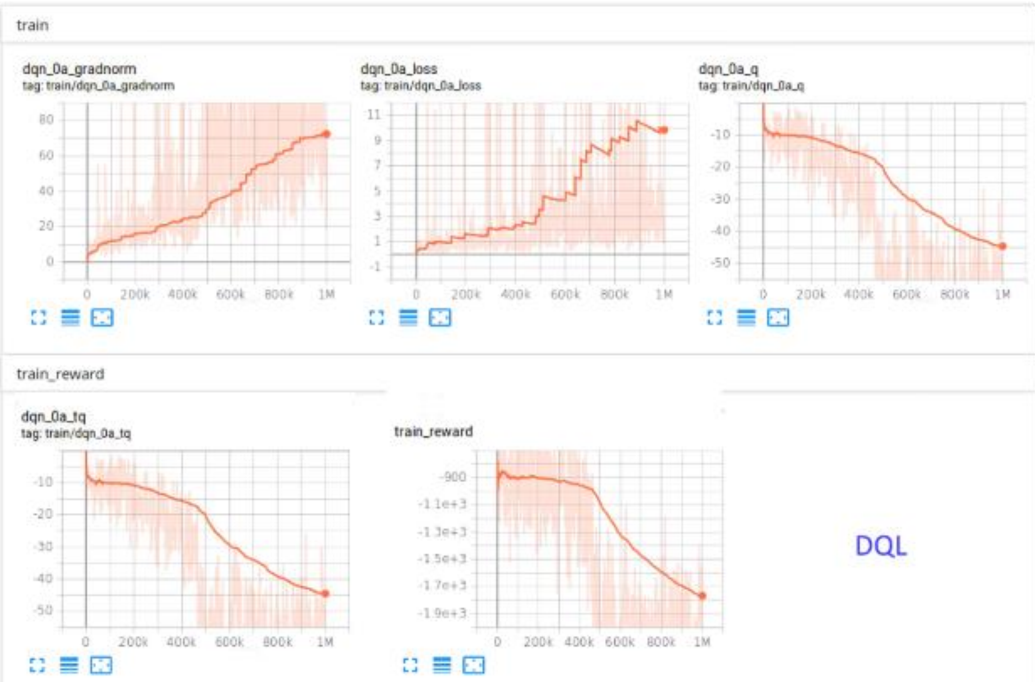
Traffic Signal Control

To do:

- City of Palermo/Other cities
- Deep Q-Learning performance analysis
- Double Deep Q-Learning and Experience Sampling
- Other Hyper/Meta-parameters
- Other Deep Learning Algorithms

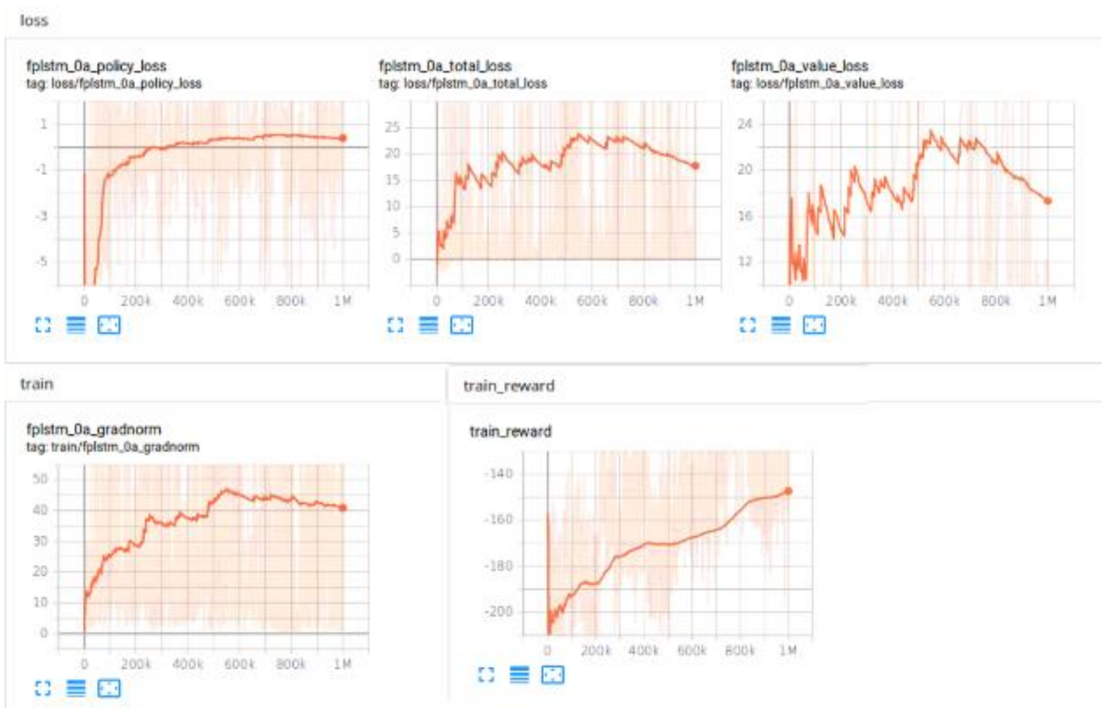
Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Traffic Signal Control



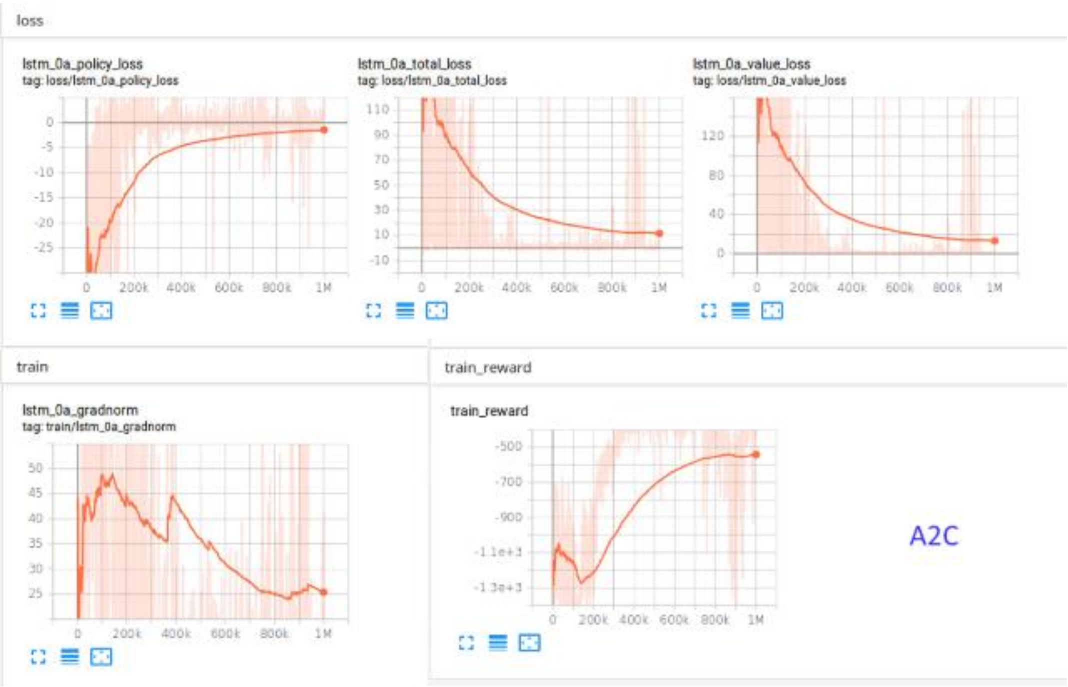
DQL

Traffic Signal Control



Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

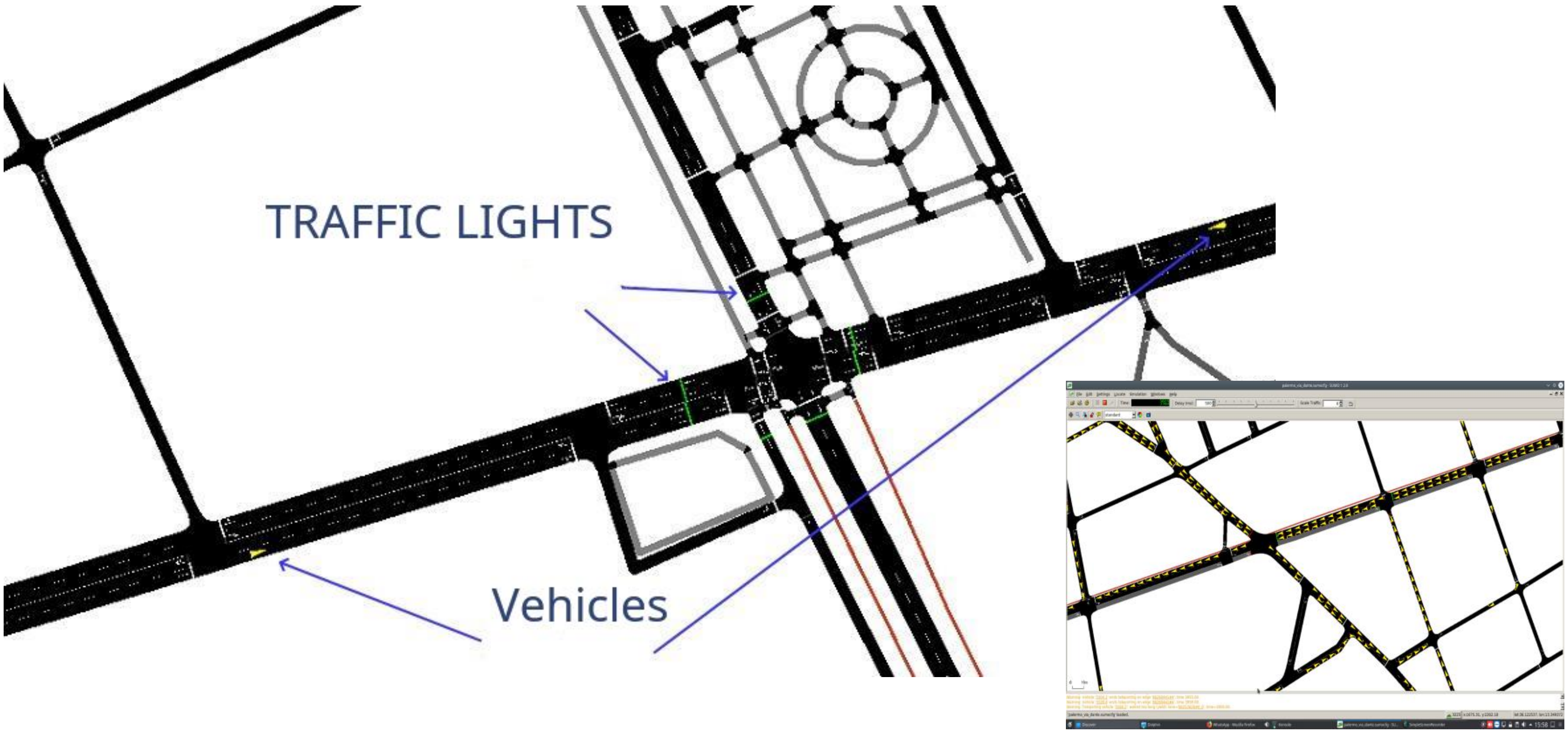
Traffic Signal Control



A2C

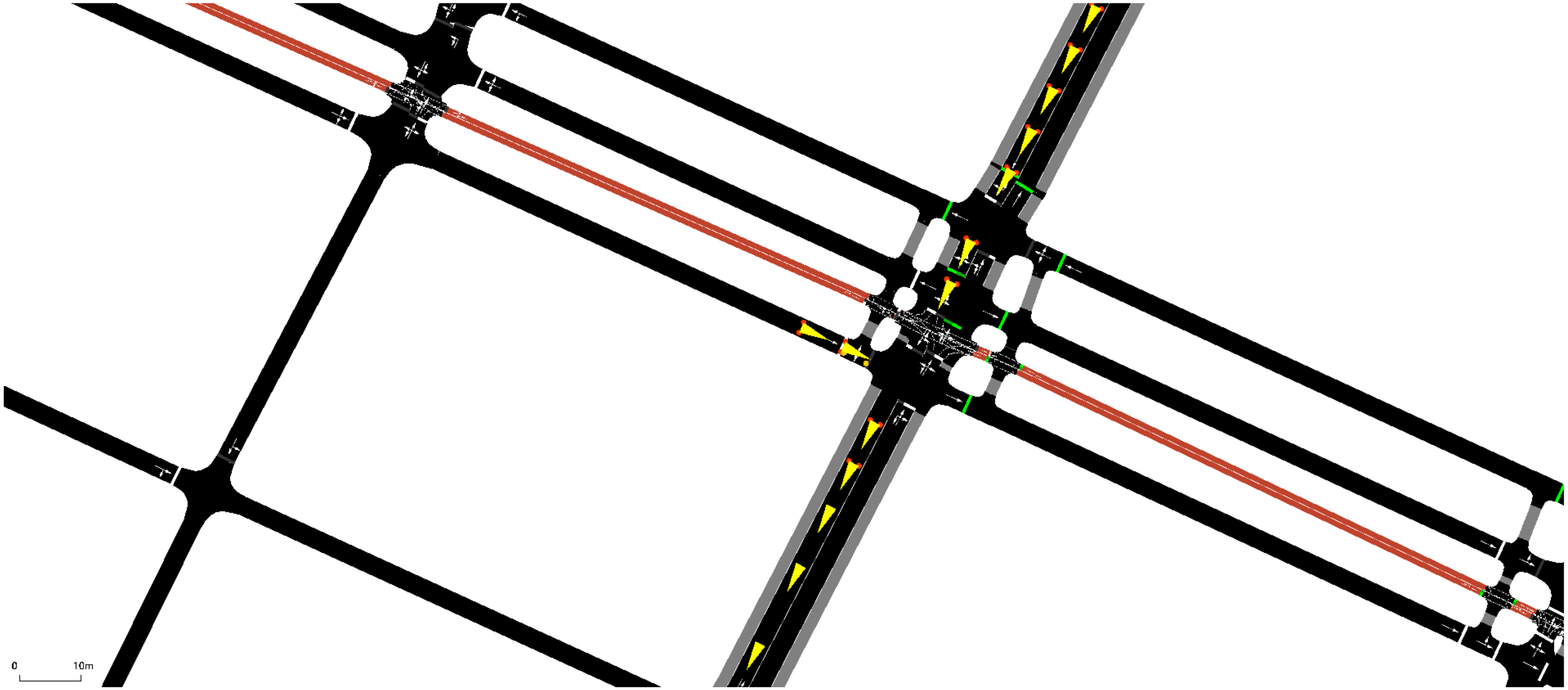
Traffic Signal Control

SUMO: Palermo (‘via Dante’ area)



Traffic Signal Control

SUMO: Torino ('Porta Nuova Station') area



Thank you for watching/listening

Reinforcement Learning

In RL, an agent interacts with its environment, typically modeled as a MDP **(S,A,p,r,γ)**, with state space **S**, actionspace **A**, and ***unknown*** transition dynamics **p(s’|s,a)**. At each discrete time step, the agent receives a reward **r(s,a,s’) ∈ R** for performing action **a** in states and arriving at the state **s**’. The goal of the agent is to maximize the expectation of the sum of discounted rewards, known as the return:

$$R_t = \sum_{i=t+1}^{\infty} \gamma_i \cdot r(s_i, a_i, s_{i+1})$$

which weighs future rewards with respect to the discount factor **γ∈[0,1)**.

Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Traffic Signal Control

$$\tilde{R}_{t,i} = \hat{R}_{t,i} + \gamma^{t_B - t} V_{w_i^-}(\tilde{s}_{t_B,\mathcal{V}_i}, \pi_{t_B-1,\mathcal{N}_i} | \pi_{\theta^-_{-i}}).$$

$$\mathcal{L}(w_i) = \frac{1}{2|B|} \sum_{t \in B} \left(\tilde{R}_{t,i} - V_{w_i}(\tilde{s}_{t,\mathcal{V}_i}, \pi_{t-1,\mathcal{N}_i}) \right)^2.$$

$$\mathcal{L}(\theta_i) = -\frac{1}{|B|} \sum_{t \in B} \Bigg(\log \pi_{\theta_i}(u_{t,i} | \tilde{s}_{t,\mathcal{V}_i}, \pi_{t-1,\mathcal{N}_i}) \tilde{A}_{t,i} - \beta \sum_{u_i \in \mathcal{U}_i} \pi_{\theta_i} \log \pi_{\theta_i}(u_i | \tilde{s}_{t,\mathcal{V}_i}, \pi_{t-1,\mathcal{N}_i}) \Bigg)$$

Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it

Bibliography

- Deep Reinforcement Learning for Multi-Agent Systems: A Review of Challenges, Solutions and Applications (Guyen et al – 2019)
- (web) https://www.analyticsvidhya.com/blog/2019/04/introduction-deep-q-learning-python/
- Multi-Agent Deep Reinforcement Learning for Large-scale Traffic Signal Control (Chu et al. - 2019)
- (web) https://becominghuman.ai/the-very-basics-of-reinforcement-learning-154f28a79071
- Off-Policy Deep Reinforcement Learning without Exploration (Fujimoto et al. - 2018)
- Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning (Foerster et al. - 2018)
- https://github.com/geek-ai/MAgent

Paolo Fazzini paolo.fazzini@iia.cnr.it www.iia.cnr.it