Neighborhood Cognition Consistent Multi-Agent Reinforcement Learning

Hangyu Mao,[1,2] Wulong Liu,[2] Jianye Hao,[3,2] Jun Luo[2]
Dong Li,[2] Zhengchao Zhang,[1] Jun Wang,[4] Zhen Xiao[1]
[1]Peking University, [2]Noah's Ark Lab, Huawei
[3]Tianjin University, [4]University College London
{hy.mao, zhengchaozhang, xiaozhen}@pku.edu.cn
{liuwulong, haojianye, jun.luo1, lidong106}@huawei.com
jun.wang@cs.ucl.ac.uk

北京大学 PEKING UNIVERSITY

HUAWEI 华为技术有限公司

诺亚方舟实验室 Noah's Ark Lab

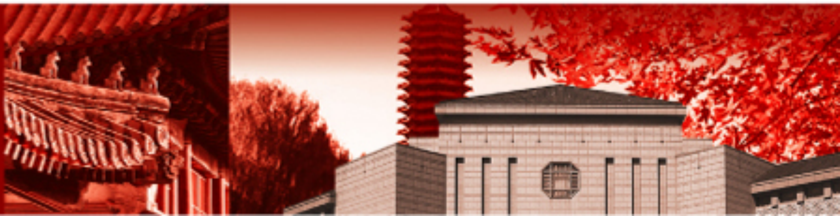
天津大学 TIANJIN UNIVERSITY

UCL 英国伦敦大学学院 University College London

北京大学

Outline

- **Motivation**
- Design
- Evaluation
- Conclusion



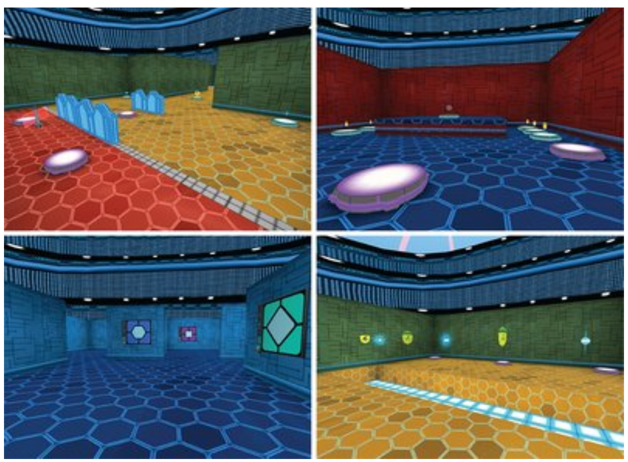
Many Stories of DRL

AlphaGo

Win the best human

Go → Go Zero → Zero

Reduce data center cooling bill by 40%



Playing Atari games

Berkeley helicopter

Relatively Backward of MARL

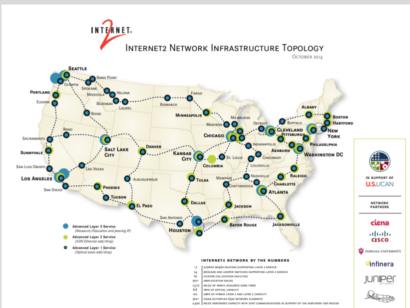
five decentralized cooperative agents

only one centralized agent

what is the next one

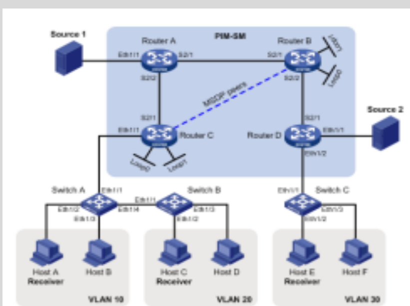
Focus of This Research

more agents in real systems



Abilene/Internet2 Network

Unmanned Aerial Vehicle



Smart Grid/WiFi Network

Autopilot/Unmanned Warship

possible answer:

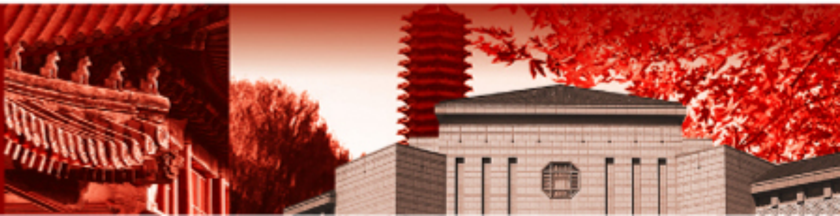
more agents

Motivation of Our Method

But how to coordinate more agents?

Cognitive Consistency

Neighborhood Cognitive Consistency (NCC)



Motivation of Our Method

We apply NCC to MARL to guarantee good agent cooperation.

(without disturbing by neighborhood formation)

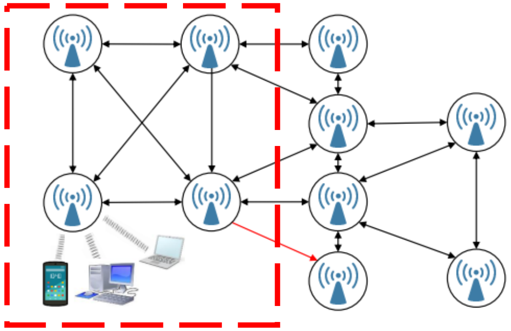
Outline

- Motivation
- **Design**
- Evaluation
- Conclusion

Key Definition

- Neighborhood
  - The neighboring agents linked with agent i are represented as $N(i)$, and each agent $j \in N(i)$ is within the **neighborhood** of agent i.
- Cognition
  - We define **cognition** of an agent as its understanding of the local environment.
  - It includes the observations of all agents in its neighborhood, as well as the high-level knowledge extracted from these observations (e.g., learned through deep neural networks).
- Neighborhood Cognitive Consistency
  - We define **NCC** as that the neighboring agents have formed similar cognitions about their neighborhood.
  - The similarity can be measured, e.g., by the similar distribution of cognition variables.

define MAS as graph

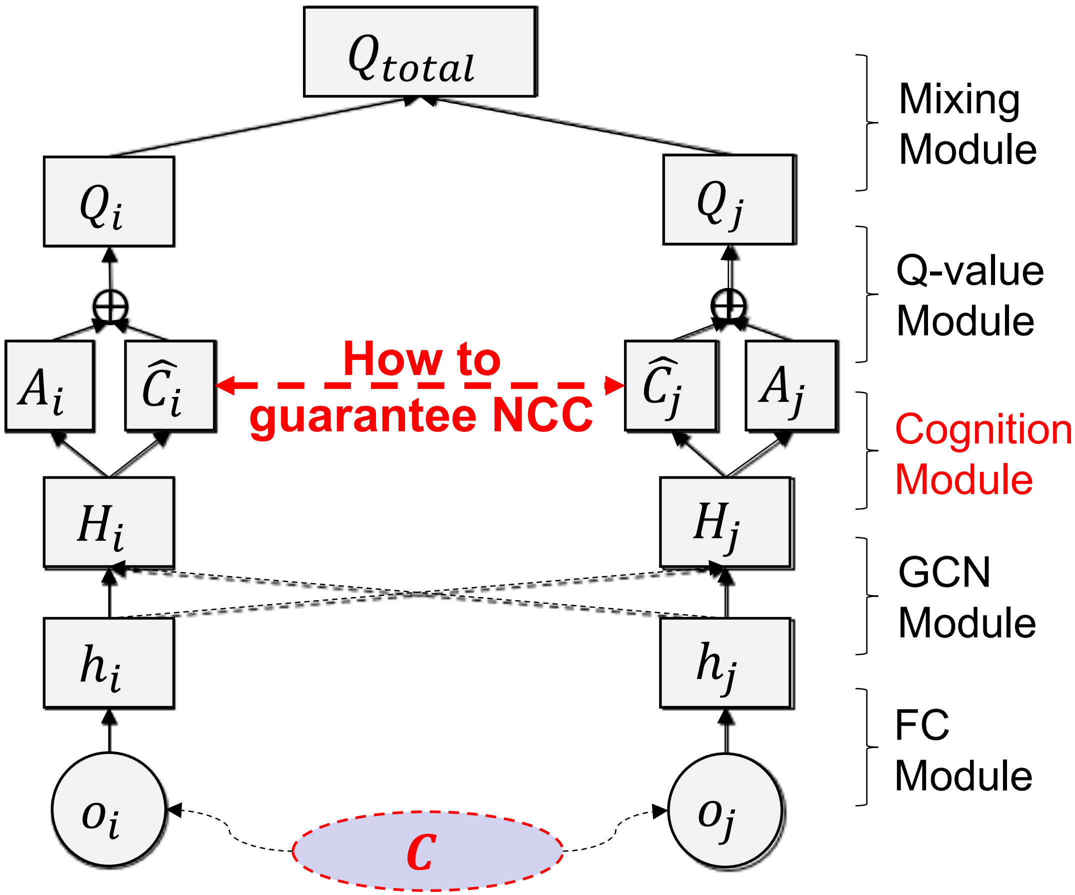


neighborhood

radio frequency
bandwidth
the rate of package loss
the number of band
the current number of users
download bytes in ten seconds
the upload coordinate speed (Mbps)
the download coordinate speed
Latency
etc.

observation → cognition

Overall Design of NCC-Q



When:
$\widehat{C}_i$ and C are consistent &
$\widehat{C}_j$ and C are consistent,
Then:
$\widehat{C}_i$ and $\widehat{C}_j$ will be consistent.

Every agent in the neighborhood tries to generate its own cognitive variable $\hat{C}$ *by variational inference such that $\hat{C}$ is consistent with C*.

Suppose every neighborhood has a ***true hidden*** *cognitive variable C*.

Guarantee of NCC

Assumption 1. For each neighborhood, there is a ***true hidden cognitive variable C*** to derive the observation o_j of each agent $j \in N(i) \cap \{i\}$.

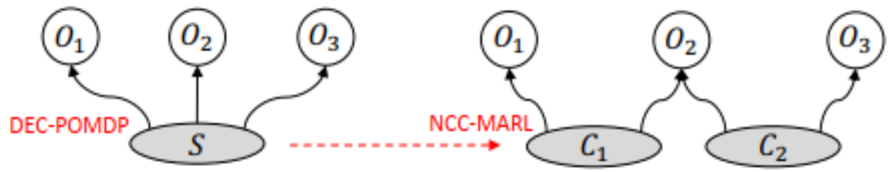


Figure 2: In NCC-MARL, the observations O_i are generated based on the hidden cognitive variable C_i instead of global state S. Here, agent 2 belongs to two neighborhoods.

In large-scale settings, decomposing S into individual cognitive variables for each neighborhood is more in line with the reality.

Guarantee of NCC

Assumption 2. If the neighboring agents can recover the true hidden cognitive variable C, they will eventually form consistent neighborhood cognitions and thus achieve better cooperations. In other words, the *learned* cognitive variable $\widehat{C_i}$ should be similar to the *true* cognitive variable C.

Assumption 2 can be formulated as a **variational inference problem**:

Supposing each agent i can only observe o_i [1], there exists a hidden process $p(o_i|C)$, and we would like to infer C by:

$$p(C|o_i) = \frac{p(o_i|C)p(C)}{p(o_i)} = \frac{p(o_i|C)p(C)}{\int p(x|C)p(C)dC} \qquad (7)$$

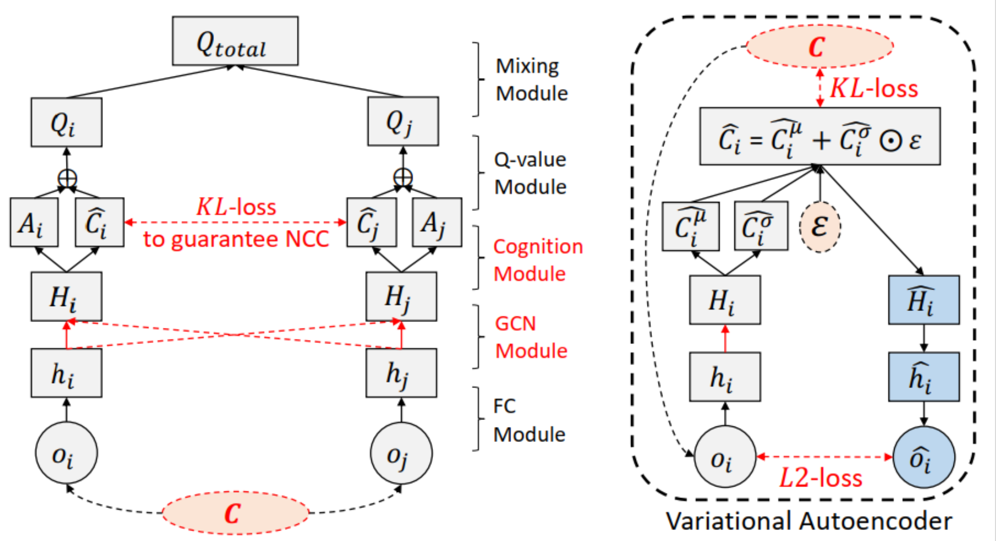
Directly computing Equation (7) is quite difficult, so we **approximate** $p(C|o_i)$ with $q(C|o_i)$ by minimizing KL-Divergence between them:

$$\min KL(q(C|o_i)||p(C|o_i)) = \max \mathbb{E}_{q(C|o_i)} \log p(o_i|C) - KL(q(C|o_i)||p(C)) \quad \textbf{(8)}$$

Guarantee of NCC

Assumption 2. If the neighboring agents can recover the true hidden cognitive variable C, they will eventually form consistent neighborhood cognitions and thus achieve better cooperations. In other words, the *learned* cognitive variable $\widehat{C_i}$ should be similar to the *true* cognitive variable C.

$$\min KL(q(C|o_i)||p(C|o_i)) = \max \mathbb{E}_{q(C|o_i)} \log p(o_i|C) - KL(q(C|o_i)||p(C)) \quad \textbf{(8)}$$



(a) **Left**: the network structure of NCC-Q. **Right**: the details of a single agent i.

Training Method

NCC-Q is trained by minimizing two loss functions. First, a temporal-difference loss (TD-loss) is shared by all agents:

$$L^{td}(w) = \mathbb{E}_{(\vec{o},\vec{a},r,\vec{o}')}[(y_{total} - Q_{total}(\vec{o},\vec{a};w))^2] \quad (11)$$

$$y_{total} = r + \gamma \max_{\vec{a}'} Q_{total}(\vec{o}', \vec{a}'; w^-) \quad (12)$$

This is analogous to the standard DQN loss shown in Equation 1 and 2. It encourages all agents to cooperatively produce a large Q_{total}, and thus ensures good agent cooperation at the whole team level as the training goes on.

Second, a cognitive-dissonance loss (CD-loss) is specified for each agent i:

$$L_i^{cd}(w) = \mathbb{E}_{o_i}[L2(o_i, \widehat{o_i}; w) + KL(q(\widehat{C_i}|o_i; w)||p(C))] \quad (13)$$

This is a mini-batch version of Equation 10. It ensures that cognitive consistency and good agent cooperation can be achieved at the neighborhood level as the training goes on.

The total loss is a combination of Equation 11 and 13:

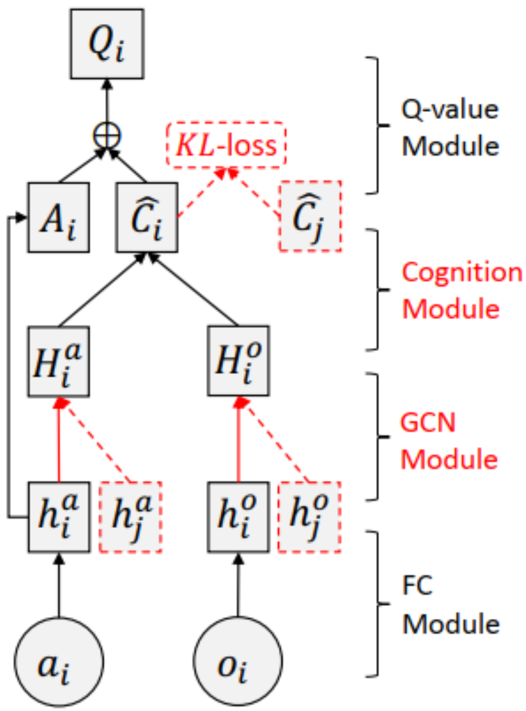
$$L^{total}(w) = L^{td}(w) + \alpha \Sigma_{i=1}^{N} L_i^{cd}(w) \quad (14)$$

Nevertheless, there are two remaining questions about the CD-loss $L_i^{cd}(w)$. (1) The true ***hidden*** cognitive variable C and its distribution $p(C)$ are unknown. (2) If there are multiple agent neighborhoods, how to choose a suitable $p(C)$ for each neighborhood.

In cases that there is only one neighborhood (e.g., the number of agents is small), we assume that $p(C)$ follows a unit Gaussian distribution, which is commonly used in many variational inference problems. However, if there are more neighborhoods, it is neither elegant nor appropriate to apply the same $p(C)$ for all neighborhoods. In practice, we find that the neighboring agents' cognitive distribution $q(\widehat{C_j}|o_j; w)$ is a good surrogate for $p(C)$. Specifically, we approximate the cognitive-dissonance loss by:

$$\begin{aligned} L_i^{cd}(w) &= \mathbb{E}_{o_i}[L2(o_i, \widehat{o_i}; w) + KL(q(\widehat{C_i}|o_i; w)||p(C))] \\ &\approx \mathbb{E}_{o_i}[L2(o_i, \widehat{o_i}; w) + \quad (15) \\ &\quad \tfrac{1}{|N(i)|} \Sigma_{j \in N(i)} KL(q(\widehat{C_i}|o_i; w)||q(\widehat{C_j}|o_j; w))] \end{aligned}$$

NCC-AC for Continuous Action



(b) The critic structure of NCC-AC.

Like NCC-Q, the critic of NCC-AC is trained by minimizing the combination of $L_i^{td}(w_i)$ and $L_i^{cd}(w_i)$ as follows:

$$L_i^{total}(w_i) = L_i^{td}(w_i) + \alpha L_i^{cd}(w_i) \tag{16}$$

$$L_i^{td}(w_i) = \mathbb{E}_{(o_i,\vec{o}_{-i},a_i,\vec{a}_{-i},r,o'_i,\vec{o}'_{-i})\sim D}[(\delta_i)^2] \tag{17}$$

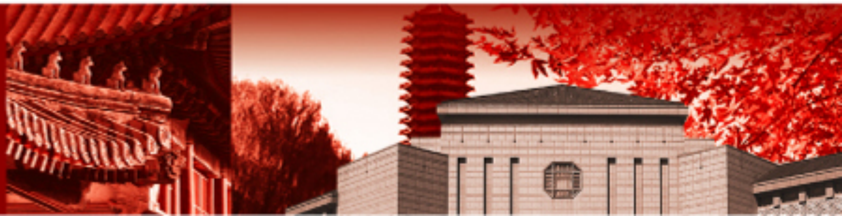
$$\begin{aligned}\delta_i &= r + \gamma Q_i(\langle o'_i, a'_i\rangle, \vec{o}'_{-i}, \vec{a}'_{-i}; w_i^-)|_{a'_j=\mu_{\theta_j^-}(o'_j)} \\ &\quad - Q_i(\langle o_i, a_i\rangle, \vec{o}_{-i}, \vec{a}_{-i}; w_i)\end{aligned} \tag{18}$$

$$\begin{aligned}L_i^{cd}(w_i) &\approx \mathbb{E}_{o_i}[L2(o_i, \widehat{o}_i; w_i) + L2(a_i, \widehat{a}_i; w_i) + \\ &\frac{1}{|N(i)|}\Sigma_{j\in N(i)} KL(q(\widehat{C}_i|o_i, a_i; w_i)||q(\widehat{C}_j|o_j, a_j; w_j))]\end{aligned} \tag{19}$$

As for the actor of NCC-AC, we extend Equation 5 into multi-agent formulation as follows:

$$\begin{aligned}\nabla_{\theta_i} J(\theta_i) = &\mathbb{E}_{(o_i,\vec{o}_{-i})\sim D}[\nabla_{\theta_i}\mu_{\theta_i}(o_i) * \\ &\nabla_{a_i} Q_i(\langle o_i, a_i\rangle, \vec{o}_{-i}, \vec{a}_{-i}; w_i)|_{a_j=\mu_{\theta_j}(o_j)}]\end{aligned} \tag{20}$$

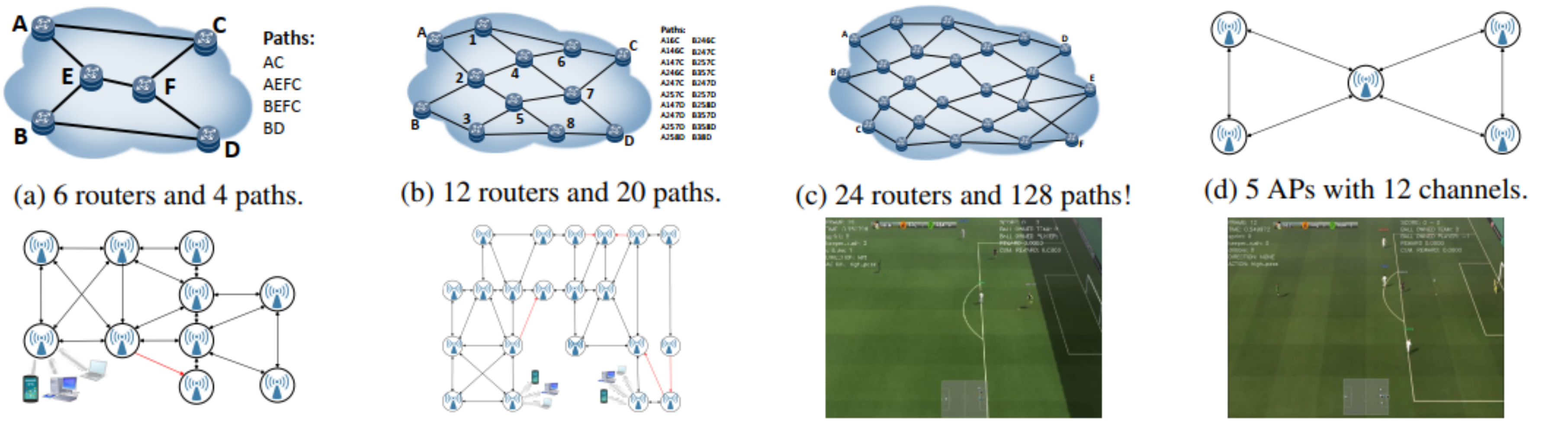
北京大學



Outline

- Motivation
- Design
- **Evaluation**
- Conclusion

Environments



(a) 6 routers and 4 paths. (b) 12 routers and 20 paths. (c) 24 routers and 128 paths! (d) 5 APs with 12 channels.

(e) 10 APs with 35 channels. (f) 21 APs with 70 channels! (g) 2-vs-2 scenario. (h) 3-vs-2 scenario.

Figure 3: The evaluation environments that are developed based on real-world scenarios. (a-c): The small, middle and large packet routing topologies. (d-f): The small, middle and large wifi configuration topologies. (g-h): The Google football tasks.

The natural topology between agents can be used to form neighborhoods, so we can evaluate our methods without disturbing by neighborhood formation.

Baselines

- Discrete Action
  - VDN
  - QMIX
  - Independent DQN (IDQN)
  - DGN

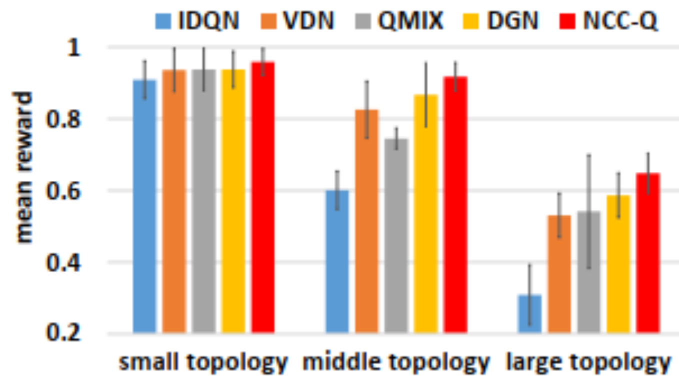
  → NCC-Q

- Continuous Action
  - MADDPG
  - ATT-MADDPG

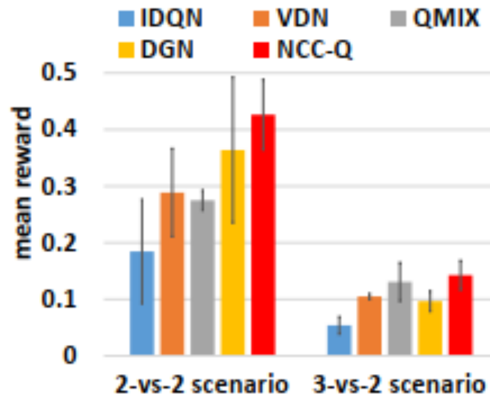
  → NCC-AC

Results

- Better performance
- Lower variance and more stable



(a) Wifi configuration tasks.



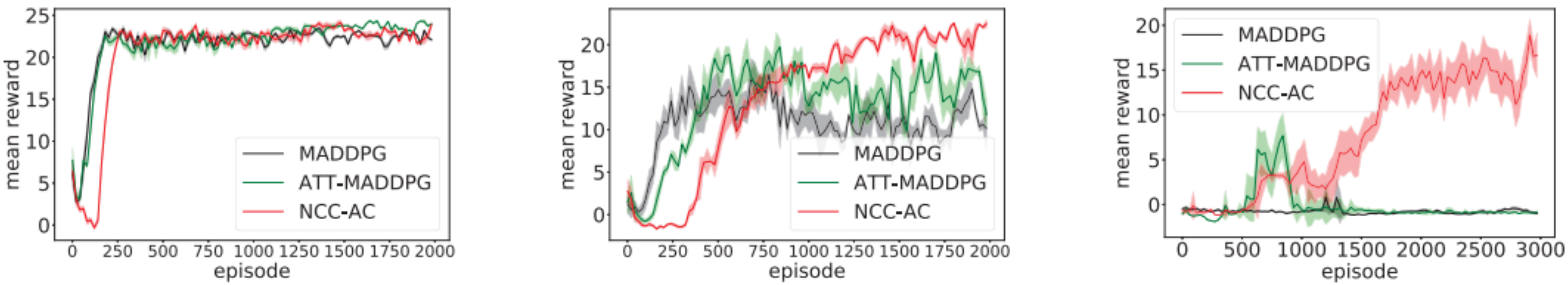
(b) Google football tasks.

Figure 5: The average results of wifi and football tasks.

北京大學

Results

- Better performance
- Better scalability



(a) Small topology: 6 routers, 4 paths.
(b) Middle topology: 12 routers, 20 paths.
(c) Large topology: 24 routers, 128 paths!

Figure 4: The average results of different packet routing scenarios.

Ablation Study

- Discrete Action
  - Graph-Q (w/o any CC)
  - GCC-Q (w/ Global CC)

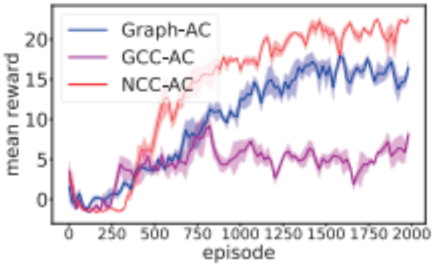
NCC-Q

- Continuous Action
  - Graph-AC (w/o any CC)
  - GCC-AC (w/ Global CC)

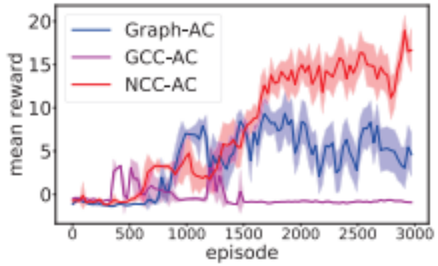
NCC-AC

Results

- Approaches with NCC work well in all scenarios.
- Methods with GCC or without any CC can only achieve good results in specific tasks.

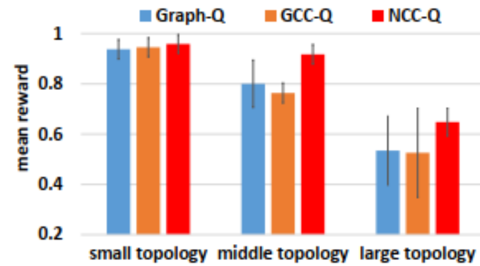


(a) For middle topology.

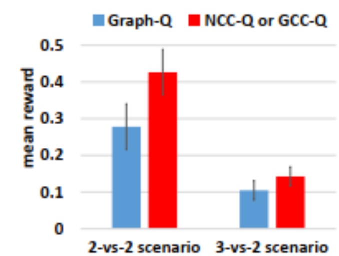


(b) For large topology.

Figure 6: The ablation results of packet routing tasks.

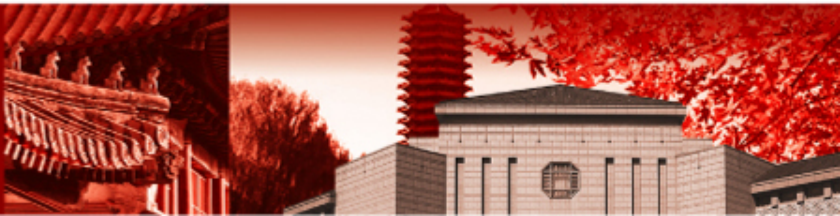


(a) Wifi configuration tasks.

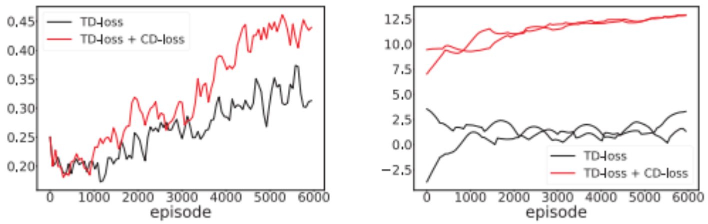


(b) Google football tasks.

Figure 7: The ablation results of wifi configuration and Google football tasks. For Google football, there is only one neighborhood, therefore GCC-Q is equivalent to NCC-Q.



Further Analysis

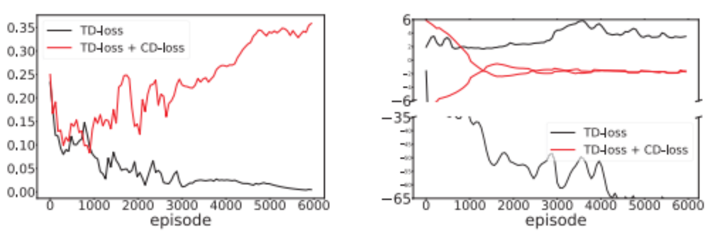


(a) The mean reward.　(b) The cognition value.

Figure 8: The results of different loss settings for the 2-vs-2 football scenario with “game_difficulty=0.6”. In Figure (b), the cognition value stands for the arithmetic mean of all elements in variable $\widehat{C_i}$; besides, there are two curves belonging to two agents for each loss setting.

In ***low-difficulty*** scenario, the proposed “CD-loss” plays a critical role to ***accelerate*** the formation of cognitive consistency and thus better cooperation.

There is usually a close relationship (e.g., positive correlation) between agent cooperation and agent cognitive consistency.

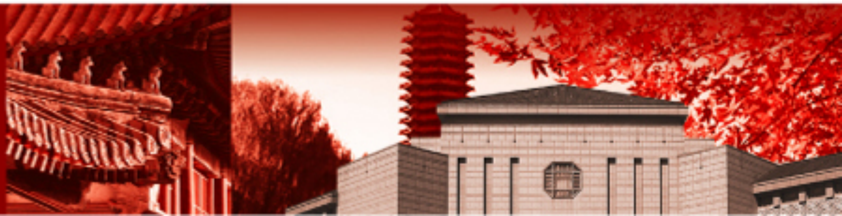


(a) The mean reward.　(b) The cognition value.

Figure 9: The results of different loss settings for the 2-vs-2 football scenario with “game_difficulty=0.9”.

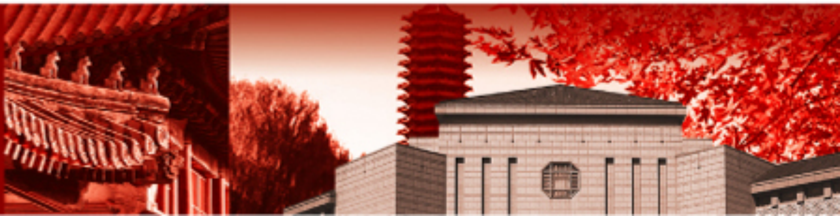
In ***high-difficulty*** scenario, the proposed “CD-loss” has the ability to ***guarantee*** the formation of cognitive consistency and thus better cooperation.

北京大學



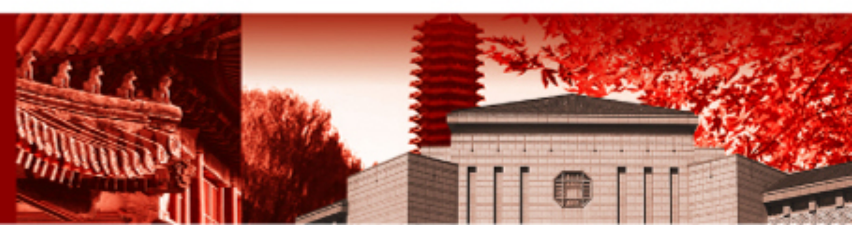
Outline

- Motivation
- Design
- Evaluation
- **Conclusion**



@methods

- Inspired by both social psychology and real experiences, this paper introduces two novel neighborhood cognition consistent reinforcement learning methods, NCC-Q and NCC-AC, to facilitate large-scale agent cooperation.
- Our methods assume a hidden cognitive variable in each neighborhood, then infer this hidden cognitive variable by variational inference. As a result, all neighboring agents will eventually form consistent neighborhood cognitions and achieve good cooperation.



@experiments

- We evaluate our methods on three tasks developed based on eight real-world scenarios.

- Extensive results show that they not only outperform the state-of-the-art methods by a clear margin, but also achieve good scalability in routing tasks.

- Moreover, ablation studies and further analyses are provided for better understanding of our methods.

Thanks for Listening!

Question?

Acknowledgments

The authors would like to thank Jun Qian, Dong Chen and Min Cheng for partial engineering support. The authors would also like to thank the anonymous reviewers for their comments. This work was supported by the National Natural Science Foundation of China under Grant No.61872397. The contact author is Zhen Xiao.