ICML
International Conference
On Machine Learning

Deconfounded Value Decomposition for Multi-Agent Reinforcement Learning

Jiahui Li[1], Kun Kuang[1*], Baoxiang Wang[23], Furui Liu[4], Long Chen[1], Changjie Fan[5], Fei Wu[1], Jun Xiao[1]

[1]ZJU

[2]CUHK,Shenzhen

[3]AIRS

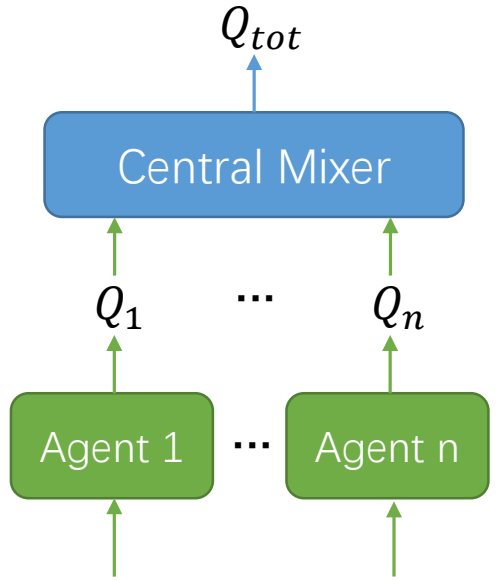
[4]Huawei Noah's ARK LAB

[5]Fuxi AI Lab, NetEase Games

VD Methods in MARL

In the CTDE paradigm, value decomposition (VD) methods have shown strength on challenging tasks. *(e.g. QMIX, QPLEX, RODE)*

They design the central mixer where the local value functions are composed into the joint value function, and the whole framework can be updated via one time backpropagation.

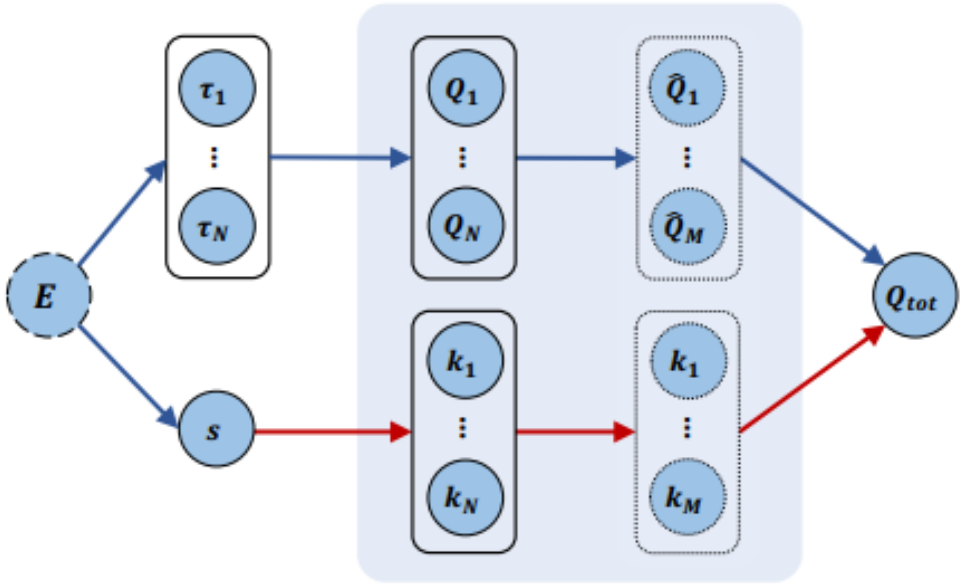


Credit Assignment in VD methods

Credit assignment is one of the main challenges in VD methods which aims to deduce the contributions of individual agents from the overall success, and is usually designed as a module (red lines) embedded in the central mixer.

Then, **the joint value function** is computed via:

$$Q_{tot} = \sum_{j=1}^{M} k_j \hat{Q}_j$$

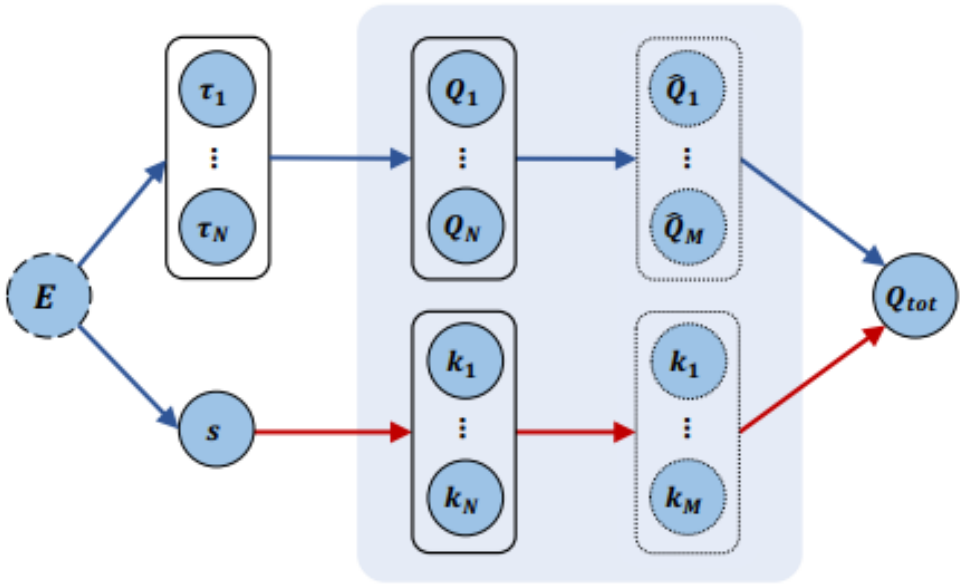


In VD methods, first, local value functions $\mathrm{Q} = \{\mathrm{Q}_1, \ \ldots, \mathrm{Q}_N\}$ are estimated via local trajectories $\tau = \{\tau_1, \ \ldots, \tau_N\}$.
Then, in the central mixer credits $\mathrm{K} = \{\mathrm{k}_1, \ \ldots, \mathrm{k}_N\}$ are estimated by a trainable network.
Finally, the joint value function Q_{tot} are computed according to Q and K.

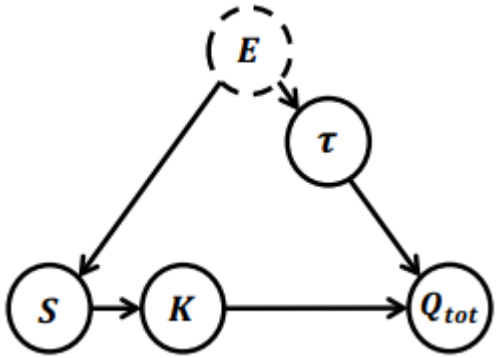
Confounding Effect in VD Methods

However, the environment E is an unobserved confounder as the common cause factor of the global state s and the joint value function Q_{tot}.

In fact, there is a backdoor path $s \leftarrow E \rightarrow \tau \rightarrow Q_{tot}$. , which is harmful to traditional VD methods.



Workflow



Causal Graph

Proposed Causal Graph

One possible approach to address the confounding bias in Fig. 1 is backdoor adjustment which is computed as $P(\mathrm{Q}_{tot}|do(s)) = \sum_{\tau} P(\mathrm{Q}_{tot}|s,\tau)P(\tau)$

It is, however, intractable to estimate the right-hand side by sampling $\tau \sim \mathrm{P}(\tau)$, as the environment is complicated and uncontrollable in general.

Hence, we propose a new causal graph in Fig. 2. We set up a new variable G and create a new path $\tau \rightarrow G \rightarrow K$

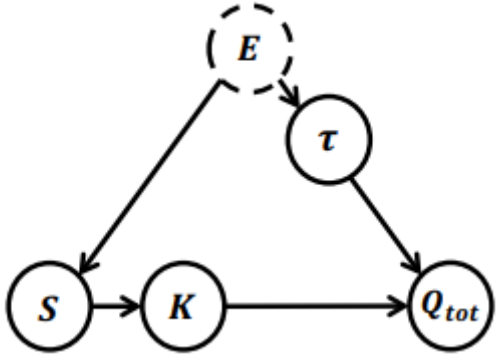


Fig. 1 Causal Graph

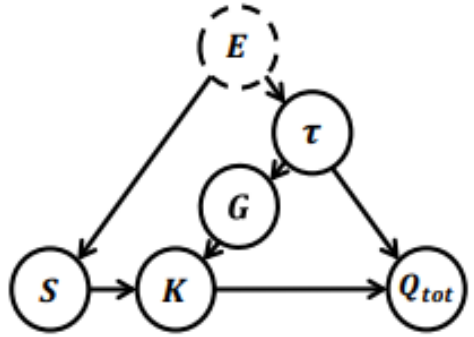


Fig. 2 Proposed Causal Graph

De-confounded Training for MARL

Path $\tau \rightarrow G \rightarrow K$ can help decompose the confounding bias on learning credits assignment into two parts: one is $s \leftarrow E \rightarrow \tau \rightarrow G \rightarrow K$, and the other is $K \leftarrow G \leftarrow \tau \rightarrow Q_{tot}$.

In the new causal graph, G serves as the proxy confounder, and $P(Q_{tot}|do(s))$ is achieved via $P(K|do(s))$ and $P(Q_{tot}|do(K))$, where

$$P(K|do(s)) = \sum_G P(K|s,G)P(G),$$
$$P(Q_{tot}|do(K)) \approx \sum_G P(Q_{tot}|K,G)P(G).$$

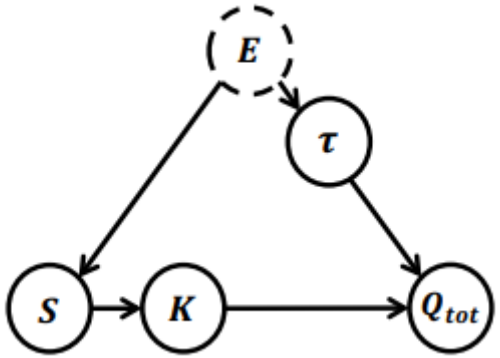


Fig. 1 Causal Graph

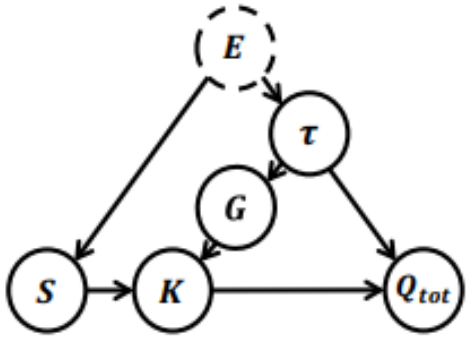


Fig. 2 Proposed Causal Graph

Our method is general enough to be applied to various VD methods and improve their performance significantly.

Overall Architecture

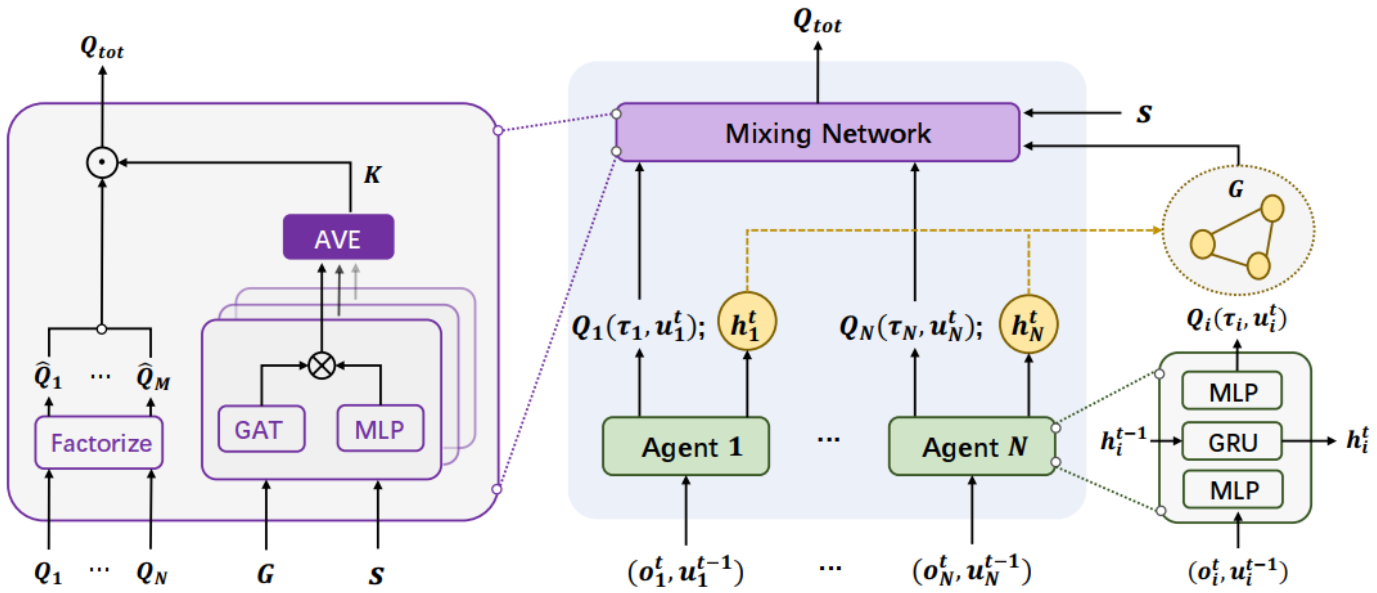


Figure 2. The framework of our method. First, each local agent models a value function conditions on its local observation-action history. Then, we construct a trajectory graph via hidden states in their RNNs. In the mixing network, local value functions $\{Q_1, \ldots, Q_N\}$ will be factorized into $\{\hat{Q}_1, \ldots, \hat{Q}_M\}$, and the graph as well as the global state are used to estimate the credits. Finally, the joint value function is computed via credits K and factorized value functions $\{\hat{Q}_1, \ldots, \hat{Q}_M\}$. The whole framework is trained via TD-loss.

First backdoor adjustment

$$K^d := P\left(K \middle| s, G^d\right) = |f_s(s)G^d|$$

$$K = \frac{1}{D}\sum_{d=1}^{D} K^d$$

Second backdoor adjustment

$$Q_{tot} = \frac{1}{D}\sum_{d=1}^{D}\sum_{j=1}^{M} k_j^d \hat{Q}_j = \sum_{j=1}^{M} k_j \hat{Q}_j$$

Experiments

StarCraft II

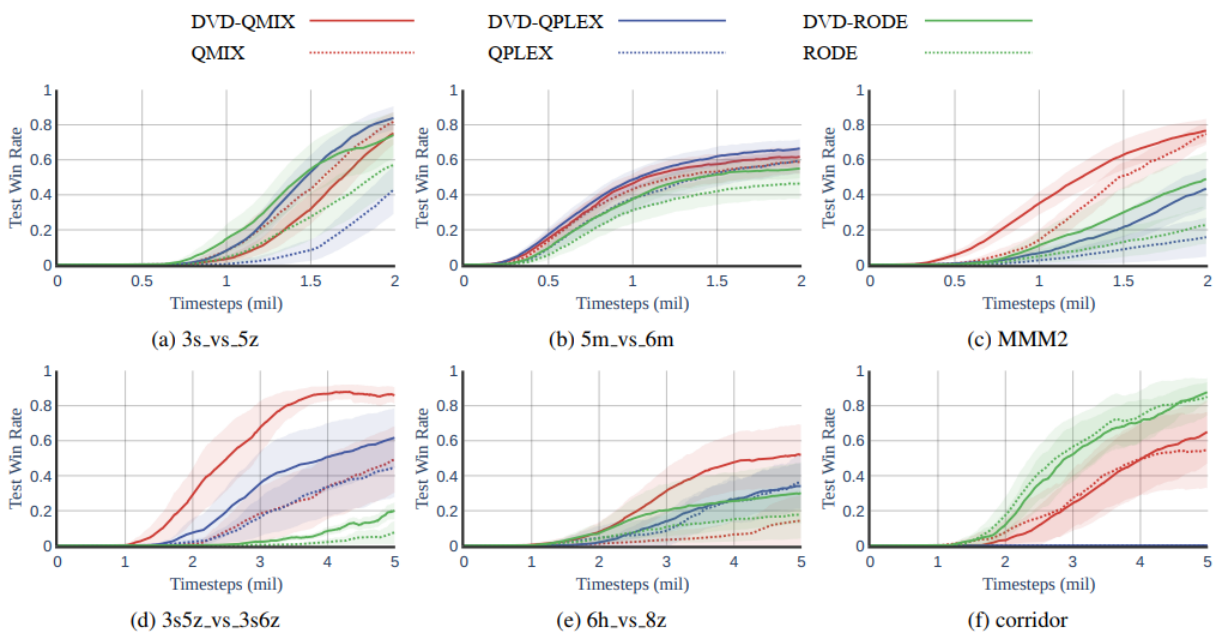


Figure 3. Performance comparison with baselines on the StarCraft II micro management benchmark.

MACO

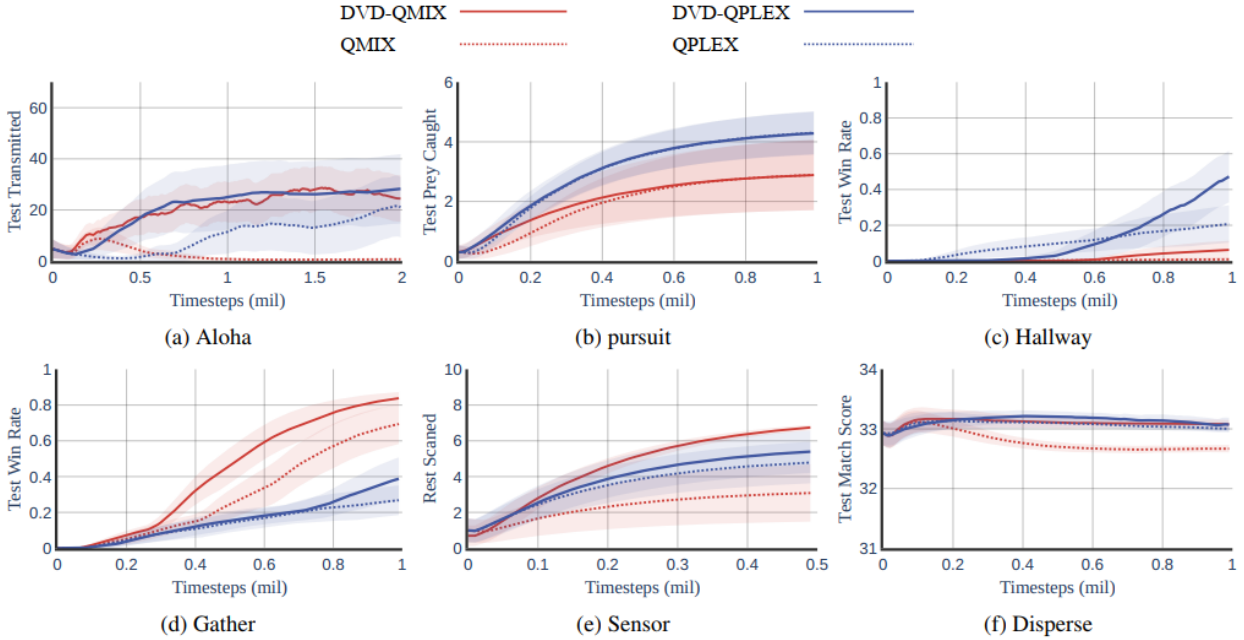


Figure 4. Performance comparison with baselines on the multi-agent coordination challenge benchmark.

Thanks!

ICML
International Conference
On Machine Learning

Deconfounded Value Decomposition for Multi-Agent Reinforcement Learning

Jiahui Li[1], Kun Kuang[1*], Baoxiang Wang[23], Furui Liu[4], Long Chen[1], Changjie Fan[5], Fei Wu[1], Jun Xiao[1]

[1]ZJU [2]CUHK,Shenzhen [3]AIRS [4]Huawei Noah's ARK LAB [5]Fuxi AI Lab, NetEase Games