1. Framework

Actor-critic or DQN

<https://www.zhihu.com/question/49787932>

Centralised training:

Solve non-stationarity.

Drawback

One agent learns a useful policy, but a second agent is discouraged from learning because its exploration would hinder the first agent and lead to worse team reward. (one agent premature)

Global reward & Local reward

Counterfactual baseline:

A new advantage function.

Motivation: Actor not only follow global reward, but also take into account agent’s own contribution.

Value-Decomposition Networks:

Only receive joint reward, learn individual Q-function automatically through network.

Motivation: reward shaping is difficult and not precise.

Top-K

Collaboration & Competition

Bi-directional RNN:

The dependency of our agents are built upon the internal layers, rather than directly from the actions.

Grouping (based on geometry):

Capture intra-group behavior.

Multi-task

Policy Ensembles.

Continual learning.

Dec-POMDP MT-MARL:

For each agent and each task, learn its specialised DRQN. Then, use these DRQN to collect data. Finally, supervised learning on these data to learn a unified network.