A unified network for Multi-task

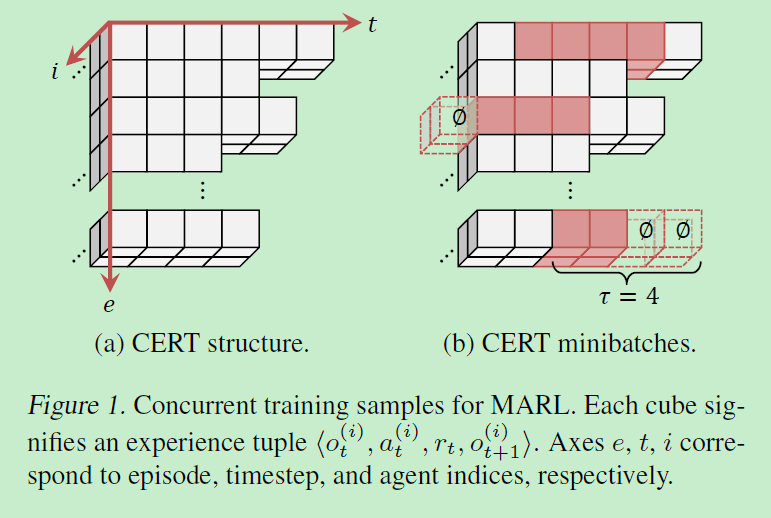
5. Approach

5.1.1 Hysteretic Q-learning

This approach uses two learning rates: nominal learning rate, a, is used when the TD-error is non-negative; a smaller learning rate, b, is used otherwise (where 0 < a < b < 1).

The result is hysteresis (lag) of Q-value degradation for actions associated with positive past experiences that occurred due to successful cooperation. Agents are, therefore, robust against negative learning due to teammate exploration and concurrent actions.

5.1.2 CONCURRENT EXPERIENCE REPLAY TRAJECTORIES (CERTS)



5.2 Multi-task

A. Data collection

Agents use each specialized DRQN (from Phase I) to execute actions in corresponding tasks, resulting in a set of regression CERTs fMRg (one per task), each containing sequences of regression experiences.

B. Supervised learning

Use tempered Kullback-Leibler (KL) divergence loss.

The motivation behind loss function (6) is that low temperatures (0 < T < 1) lead to sharpening of specialized DRQN action-values, Qbt, ensuring that the distilled DRQN ultimately chooses similar actions as the specialized policy it was trained on.

Drawback

We have to train networks for each agent and each task. In some cases we can just directly use these networks instead of using a unified network.