Appendix

Attentional Communication Algorithm

For completeness, we provide the ATOC algorithm as below.

Algorithm 1 Attentional communication

```
Initialize critic network, actor network, communication channel and attention classifier
Initialize target networks
Initialize replay buffer $\mathcal{R}$
Initialize queue $\mathcal{D}$
for episode = $1, \ldots, \mathcal{M}$ do
    for $t = 1, \ldots, \mathcal{T}$ do
        Get thought $h_t^i = \pi_{\mathrm{I}}\left(o_t^i; \theta^\mu\right)$ for each agent $i$
        Each agent $i$ decides whether to initiate communication based on $h_i$ every $T$ timesteps
        for $i$ in initiators do
            $(\tilde{h}_t^i, \tilde{h}_t^1, \ldots, \tilde{h}_t^j) = g(h_t^i, h_t^1, \ldots, h_t^j; \theta^g)$, where agent 1 to $j$ in $i$'s group
        end for
        Select action $a_t^i = \pi_{\mathrm{II}}(h_t^i, \tilde{h}_t^i; \theta^\mu) + \mathcal{N}_t^i$ for each agent $i$ with communication
        Select action $a_t^i = \pi_{\mathrm{II}}(h_t^i; \theta^\mu) + \mathcal{N}_t^i$ for each agent $i$ without communication
        Execute action $a_t^i$, obtain reward $r_t^i$, and get new observation $o_{t+1}^i$ for each agent $i$
        for $i$ in initiators do
            Get action $\bar{a}_t^j = \pi_{\mathrm{II}}(h_t^j; \theta^\mu)$ for each agent $j$ in $i$'s group $G_i$
            Compute difference of mean Q values with and without communication:
                $\Delta Q_t^i = \frac{1}{|G_i|}\left(\sum_{j \in G_i} Q\left(o_t^j, a_t^j | \theta^Q\right) - \sum_{j \in G_i} Q\left(o_t^j, \bar{a}_t^j | \theta^Q\right)\right)$
            Store $\left(h_t^i, \Delta Q_t^i\right)$ in $\mathcal{D}$
        end for
        Store transition $(O_t, A_t, R_t, O_{t+1}, C_t)$ in $\mathcal{R}$
        Sample a random minibatch of $N$ transitions from $\mathcal{R}$
        Sample agents without communication
        Update the critic $\theta^Q$ by minimizing $\mathcal{L}(\theta^Q)$
        Update the actor $\theta^\mu$ using sampled policy gradients $\nabla_{\theta^\mu} J\left(\theta^\mu\right)$
        Sample agents with communication
        Update the critic $\theta^Q$ by minimizing $\mathcal{L}(\theta^Q)$
        Update the actor $\theta^\mu$ using sampled policy gradients $\nabla_{\theta^\mu} J\left(\theta^\mu\right)$
        Update the communication channel $\theta^g$ using sampled thoughts gradients $\nabla_{\theta^g} J\left(\theta^g\right)$
        Update the target networks: $\theta' = \tau\theta + (1 - \tau)\,\theta'$
    end for
    Get $\Delta\hat{Q}_i \in [0, 1]$ by min-max normalization for each $\Delta Q_i$ in $\mathcal{D}$
    Update the attention classifier $\theta^p$ by minimizing log loss:
        $\mathcal{L}(\theta^p) = -\Delta\hat{Q}_i \log(p\left(h^i|\theta^p\right)) - (1 - \Delta\hat{Q}_i) \log\left(1 - p\left(h^i|\theta^p\right)\right)$
    Empty the queue $\mathcal{D}$
end for
```

Video

The video of the experiments is given by this link `https://goo.gl/X8wBqw`. The video shows the games of cooperative navigation and cooperative pushball. As the performance of different methods cannot be visually differentiated in predator-prey, predator-prey is not included. The video aims to highlight different behaviors of agents trained using different methods. So, we show the games of ATOC, DDPG, and ComNet agents in cooperative navigation and ATOC and DDPG agents in cooperative pushball.