Cooperation Learning in Time-Varying Multi-Agent Networks

Vasanth Reddy Baddam[1], Almuatazbellah Boker[2], Hoda Eldardiry[1]

[1]Department of Computer Science, [2]Department of Electrical and Computer Engineering
Virginia Tech

Motivation and Objectives

We propose a Multi-Agent Systems (MAS) co-ordination framework for complex and dynamic environments, where agents' neighbors vary over time. Our proposed approach Cooperation Learner in MultiAgent Networks (CooLMAN) has a number of features:

- It captures information flow in a dynamic environment using temporal indexing
- Agents can achieve optimal policy and stability by the system-enabled timed interaction and coordination
- Providing trained weights that can be deployed to larger swarms in a scalable manner.

Proposed Approach

Heat Diffused Critic:

1. MAS network is represented as a graph network
2. Observations are accommodated as the nodes and the edges as the parameter sharing channels.

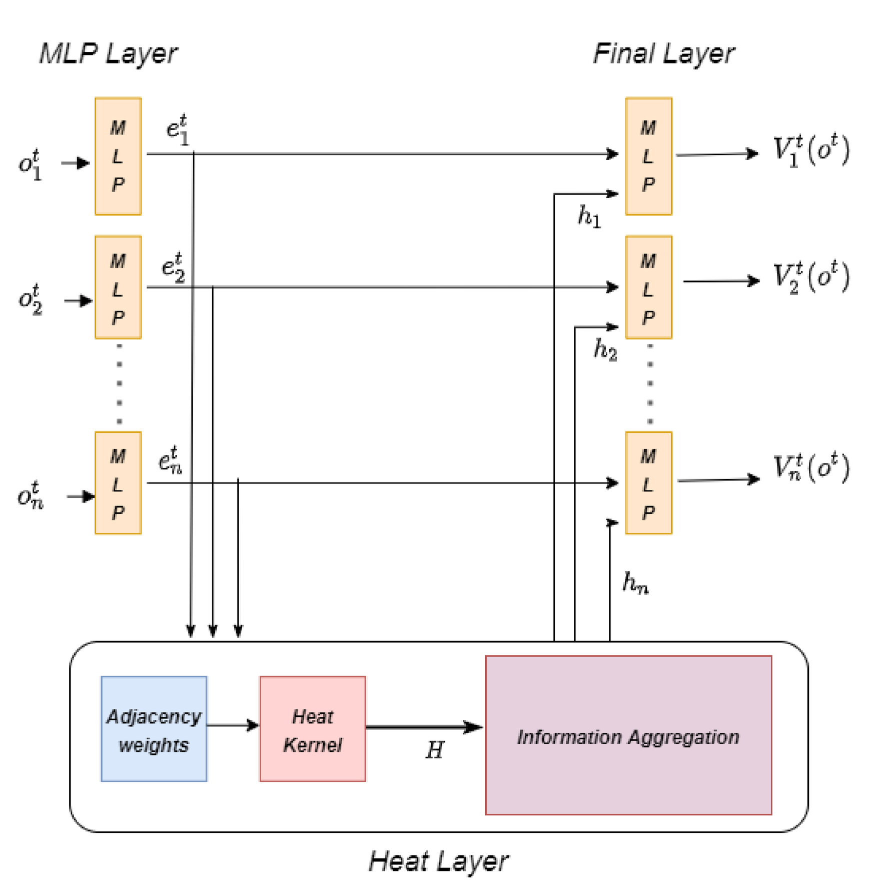

Figure 1: Illustrates the main architecture of the proposed heat diffused critic model, which perceives the local observations of all the agents and gives out the state values for each agent

VIRGINIA TECH

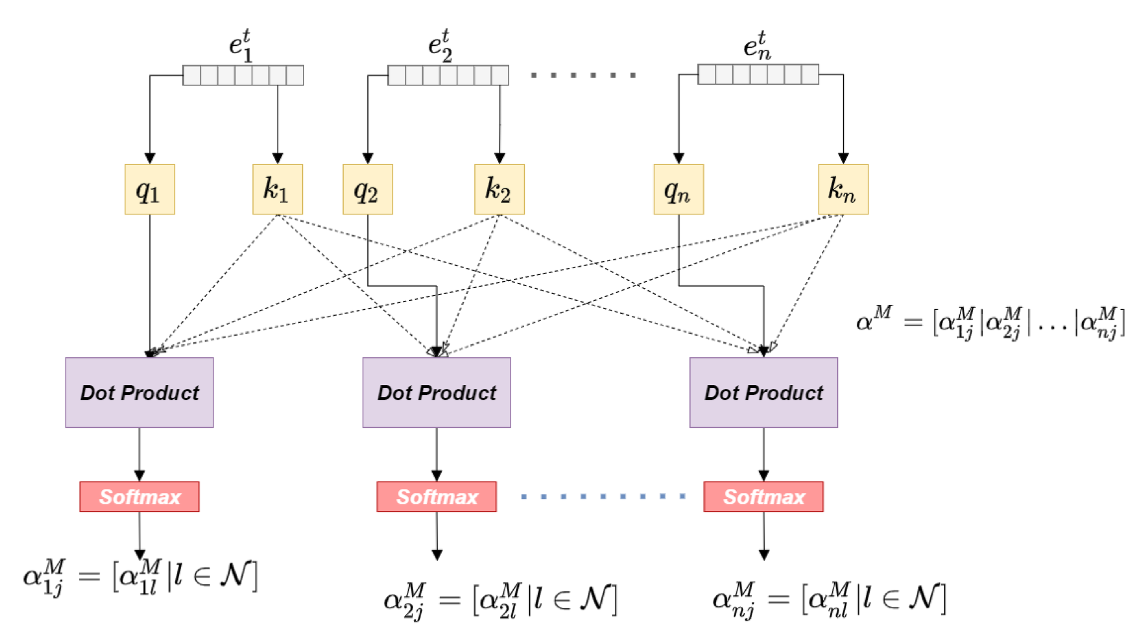

Figure 2: Illustrates the dot-product mechanism outputs the adjacency matrix elements and the same would be followed for M communication channels

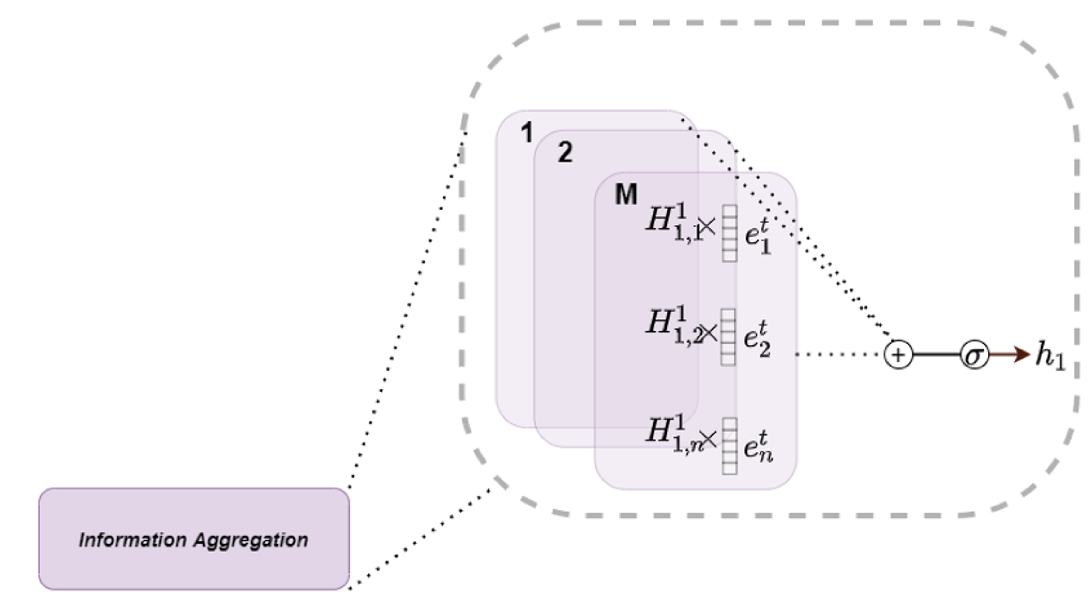

Figure 3: *Illustrates the dot-product mechanism outputs the adjacency matrix elements and the same would be followed for M communication channels*

Heat Layer

Adjacency matrix is given as:

$$\alpha_{ij} = \frac{\exp(\frac{e_{ij}}{d})}{\sum_{k \in \mathcal{N}_i} \exp(\frac{e_{ik}}{d})}, \tag{1}$$

Heat Kernel is given as:

$$\frac{\partial H^t}{\partial t} = -\hat{\mathcal{L}}^t H^t \tag{2}$$

where $\mathcal{L}^t = D^t - A^t$. D is the diagonal matrix and A is the adjacency matrix.

The solution to the above equation is given as:

$$H^t_{v_i,v_j} = \sum_{l=1}^{n} \exp^{-\lambda_l^t t} \phi_l^t(v_i)\phi_l^t(v_j) \tag{3}$$

where $\phi_l^t(v_i)$ is the eigenvector of ith node.

Training

Critic Training:

$$J^{critic}(\theta) = \sum_{i=1}^{\mathcal{N}} \mathbb{E}_{<o_t,r,o_{t+1}>\sim\mathcal{D}} \left[(y_i^{td} - V_{\theta_i}(o_t))^2\right], \tag{4}$$

where $y_i^{td} = A_i^{GAE} + V_{\varphi_i}(o_{t+1})$ and A^{GAE} is the Generalized Advantage Estimate is given as:

$$A_t^{GAE} = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{5}$$

Actor Training:

$$J_i^{CLIP}(\Phi_i) = \mathbb{E}\left[min(r(\Phi_i)A_i^{GAE}, clip(r(\Phi_i), 1-\epsilon, 1+\epsilon)\right). \tag{6}$$

Table 1: *Mean reward for the different number of agents during testing. $\mathcal{N}$ is the size of the neighborhood set of the agents. Map size is the metric given for $n \times n$ grid. Mean reward for agents 50 and 100 during testing using the weights trained on 8 agents are omitted as the number of agents does not fit in the given map size.*

			During testing				
			Number of agents(Mean reward)				
Trained on	Map Size	$\mathcal{N}$	8	14	20	50	100
8 Agents($\mathcal{N}=2$)	12	2	1.21	1.03	1.10	-	-
		3	1.08	1.06	1.12	-	-
		4	1.07	1.08	1.14	-	-
14 Agents($\mathcal{N}=4$)	21	2	-	0.95	0.90	0.83	0.74
		3	-	0.93	0.87	0.82	0.73
		4	-	0.95	0.84	0.80	0.74

Results

Table 2: *Statistics for different models.*

Stats	CooLMAN	DGN [1]	DQN [2]
Hit	412	289	4
Expired	2	3	381
mean reward	1.14	0.76	-0.015

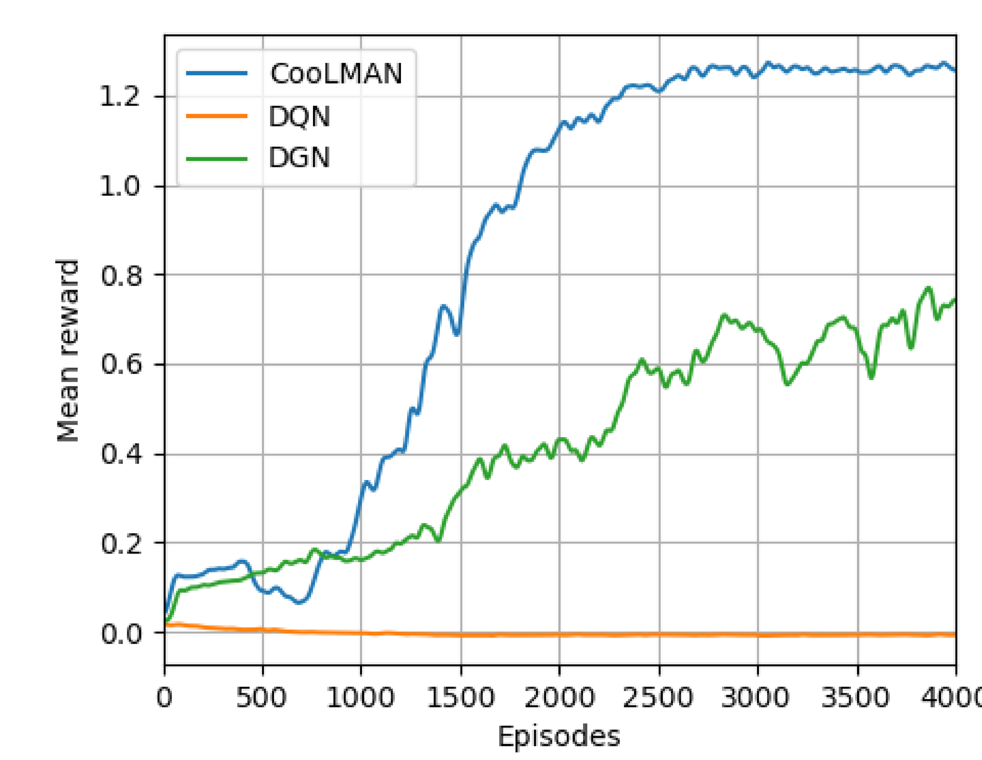

Figure 4: Mean reward for different models

Conclusion

By modelling time-varying edges as the function of heat diffusion, we enable dynamic communication channel between agents. Thus our model fits the dynamic nature of the complex time-varying multi-agent systems. In comaprison, CooLMAN to DGN and DQN, our model overcomes

References

[1] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu.
Graph convolutional reinforcement learning.
In *ICLR*, 2019.

[2] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente.
Multiagent cooperation and competition with