Credit Assignment and Fine-Tuning Enhanced Reinforcement Learning for Collaborative Spatial Crowdsourcing

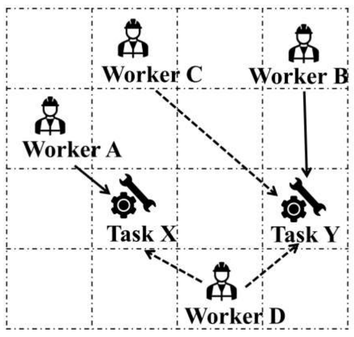
Wei Chen, Yafei Li*, Baolong Mei, Guanglei Zhu, Jiaqi Wu, Mingliang Xu

Zhengzhou University

Abstract: Collaborative spatial crowdsourcing (CSC) leverages distributed workers' collective intelligence to accomplish spatial tasks. A central challenge is to efficiently assign suitable workers to collaborate on these tasks. Although mainstream reinforcement learning (RL) methods have proven effective in task allocation, they face two key obstacles: delayed reward feedback and non-stationary data distributions, both hindering optimal allocation and collaborative efficiency. To address these limitations, we propose CAFE (credit assignment and fine-tuning enhanced), a novel multiagent RL framework for spatial crowdsourcing. CAFE introduces a credit assignment mechanism that distributes rewards based on workers' contributions and spatiotemporal constraints, coupled with bi-level meta-optimization to jointly optimize credit assignment and RL policy. To handle nonstationary spatial task distributions, CAFE employs an adaptive fine-tuning procedure that efficiently adjusts credit assignment parameters while preserving collaborative knowledge. Experiments on two real-world datasets validate the effectiveness of our framework, demonstrating superior performance in terms of task completion and equitable reward redistribution.

1. Introduction

To illustrate CSC, consider the example shown in Fig.1, which involves four workers (A-D) and two tasks (X, Y). The initial configuration assigns workers A and B to tasks X and Y respectively, while workers C and D remain unallocated. The scenario is characterized by worker heterogeneity: worker C is exclusively qualified for task Y, while worker D possesses the versatility to perform either task. Task completion times directly impact payment structures, and the assignment of worker D to either task X or Y yields different compensation outcomes. This study aims to optimize the platform's revenue through strategic task allocation. Notably, as the scale of workers and tasks expands in real-world applications, the computational complexity of CSC increases exponentially.



→ Assigned
⇢ Optional

Unmatched Workers

Worker order	Optional tasks
Worker C	Task Y
Worker D	Task X, Task Y

Matched Workers

Worker order	Assigned tasks
Worker A	Task X
Worker B	Task Y

Fig. 1. An example of collaborative spatial crowdsourcing.

2. Problem Definition

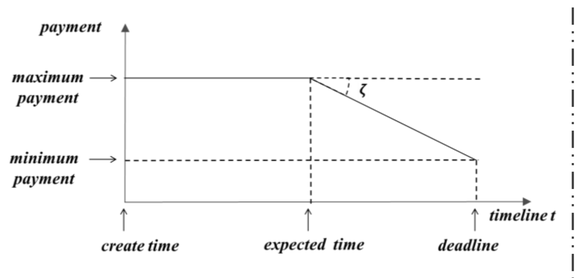


Fig. 2. Payment setting of task.

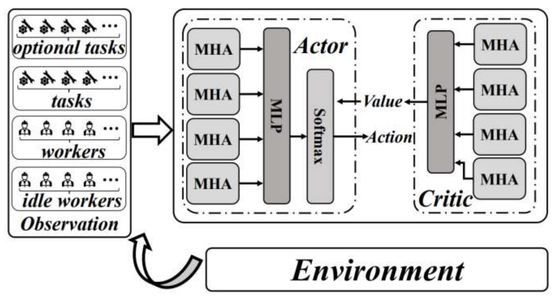


Fig. 3. Actor-Critic network structure.

Given a stream of tasks T and a set of works W , CSC problem aims to find the optimal assignment M to maximize the global revenue P .

$$P = max \sum_{\tau \in \mathrm{T}} \rho^{\tau}$$

3. Method

Reward Redistribution

To comprehensively evaluate how individual workers' actions affect overall revenue, we propose a reward redistribution mechanism that quantifies each worker's contribution.

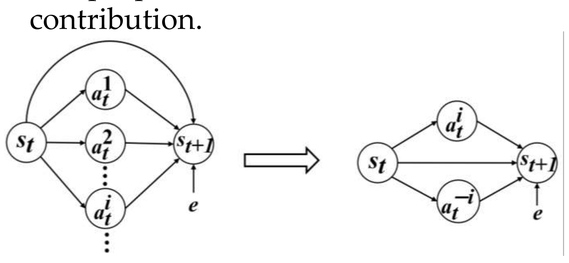


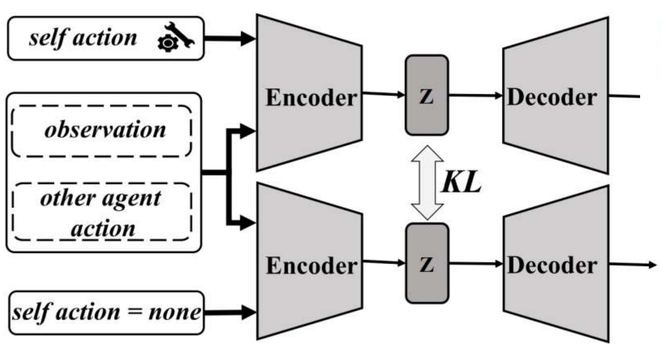
Fig. 4. Structural causal model of MARL. Meanwhile, e represents the external disturbance affecting state transitions. For notational convenience, we denote by a_t^{-i} the joint actions of all agents except agent i at time step t.

To comprehensively evaluate how individual workers' actions affect overall revenue, we propose a reward redistribution mechanism that quantifies each worker's contribution. Within this model, directional arrows represent inherent causal dependencies between components. To quantify how each agent's action a_t^i causally impacts the subsequent state s_{t+1}, we employ conditional mutual information (CMI):

$$CMI_t^i := I(s_{t+1}; a_t^i | s_t, a_t^{-i}) \approx I(o_{t+1}; a_t^i | o_t, a_{t,o}^{-i})$$

To comprehensively evaluate how individual workers' actions affect overall revenue, we propose a reward redistribution mechanism that quantifies each worker's contribution. Within this model, directional arrows represent inherent causal dependencies between components. To quantify how each agent's action a_t^i causally impacts the subsequent state s_{t+1}, we employ conditional mutual information (CMI):

$$CMI_t^i \approx I(z_\rho; a_t^i | o_t, a_{t,o}^{-i}) \approx D_{KL}[\, q(z_\rho | o_t, a_t^i, a_{t,o}^{-i}) || q(z_\rho^{-i} | o_t, a_{t,o}^{-i})].$$



Encoder: $q(z_\rho | o_t, a_t^i, a_t^{-i})$,

Decoder: $p(\sum_{\tau \in \mathcal{T}} \rho | z_\rho)$.

Fig. 5. The reward for agent's action a_t^i is constructed using a Variational Autoencoder (VAE). The encoder architecture mirrors the agent's critic structure, while the decoder is implemented as a fully connected neural network.

$$J = -D_{KL}[q\,(z|o_t, a_t^i) || p(z)] + E_{z \sim q(Z|\cdot)}[log\; p(\sum_{\tau \in \mathrm{T}} (\rho | z_\rho)]$$

To mitigate this misalignment, we introduce a regularization term into the reward function:

$$r^i(\phi) = \phi_1 \cdot CMI_t^i + \phi_2$$

Implicitly Learning

To optimize the hyperparameters in the reward function, we propose a novel bi-level optimization method based on implicit gradient tune the reward function's hyperparameters. This approach enables the concurrent optimization of hyperparameters during the reinforcement learning process.

$$\text{inner-level:} \quad \theta^* = \arg\max_{\theta} J(\theta),$$

$$\text{outer-level:} \quad \phi^* := \arg\max_{\phi} [J(\theta, \phi) - \mathcal{L}_\Lambda]$$

which is expressed as:

we use $\mathcal{L}_\Lambda$ to regulate the long-term reward,

$$\mathcal{L}_\Lambda(\phi) = \mathrm{E}_{\lambda \sim \Lambda} \left[\left(\sum_{t=1}^{T} r(\phi) - \eta \cdot \sum_{t=1}^{T} \rho \right)^2 \right]$$

We define $F(\phi)$ as the outer-level loss function, and its gradient can calculate by implicit gradient:

$$\frac{\mathrm{d}F}{\mathrm{d}\phi} \approx \frac{\partial J}{\partial \phi} - \frac{\partial J}{\partial \theta^l} \cdot \left(\frac{\partial^2 J}{\partial \theta_i^l \partial \theta_j^l} \right)^{-1} \cdot \frac{\partial^2 J}{\partial \theta^l \partial \phi} - \frac{\mathrm{d}\mathcal{L}_\Lambda}{\mathrm{d}\phi}$$

Swift Parameter Refinement

Firstly, we use the Taylor series expansion to the second-order term

$$\mathcal{L}_\Lambda(\phi) \approx \mathcal{L}_\Lambda(\hat{\phi}) + \left(\phi - \hat{\phi}\right)^T \cdot \left[\frac{\mathrm{d}\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi}\bigg|_{\phi=\hat{\phi}} \right] + \left(\phi - \hat{\phi}\right)^T \cdot \left[\frac{\mathrm{d}^2\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi^2}\bigg|_{\phi=\hat{\phi}} \right] \cdot \left(\phi - \hat{\phi}\right)$$

$$\frac{\mathrm{d}\mathcal{L}_\Lambda}{\mathrm{d}\phi} \approx \left[\frac{\mathrm{d}\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi}\bigg|_{\phi=\hat{\phi}} \right] + \left[\frac{\mathrm{d}^2\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi^2}\bigg|_{\phi=\hat{\phi}} \right] \cdot \left(\phi - \hat{\phi}\right)$$

Consider the concept of implicit gradients for optimizing the network parameters.

$$\frac{\mathrm{d}\mathcal{L}_\Lambda(\phi^*)}{\mathrm{d}\phi} = 0$$

$$\left[\frac{\mathrm{d}\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi}\bigg|_{\phi=\hat{\phi}} \right] + \left[\frac{\mathrm{d}^2\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi^2}\bigg|_{\phi=\hat{\phi}} \right] \cdot \left(\phi^* - \hat{\phi}\right) = 0$$

$$\phi^* = \hat{\phi} - \left[\frac{\mathrm{d}\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi}\bigg|_{\phi=\hat{\phi}} \right] \cdot \left[\frac{\mathrm{d}^2\mathcal{L}_\Lambda(\phi)}{\mathrm{d}\phi^2}\bigg|_{\phi=\hat{\phi}} \right]^{-1}$$

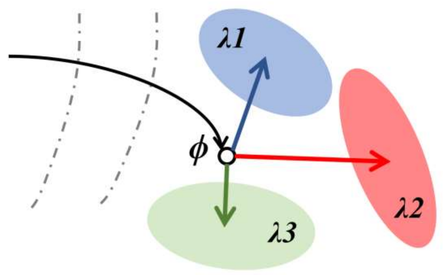
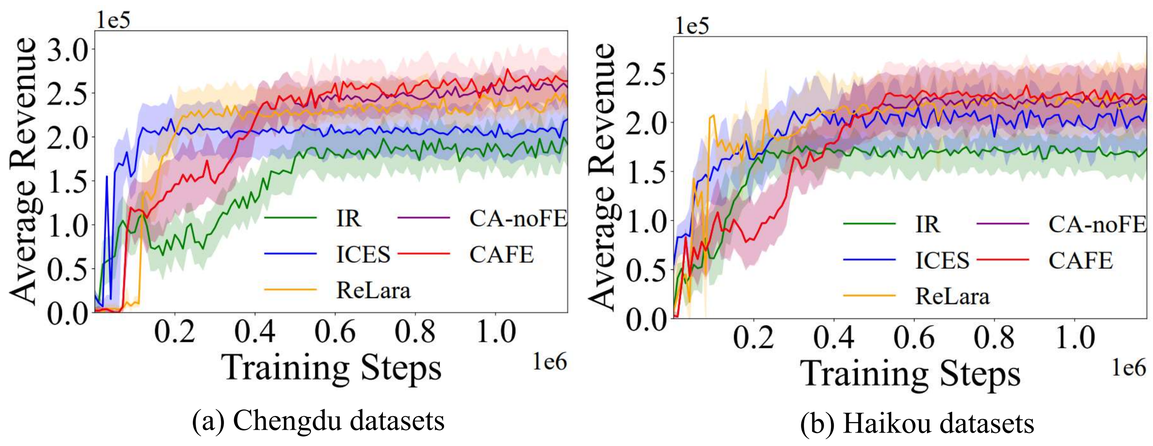


Fig. 6. Fine-tuning the parameters to adapt to changes in crowdsourcing tasks quickly.

4. Experiments

Comparison of Reward Redistribution



(a) Chengdu datasets

(b) Haikou datasets

Fig. 7. Comparison of reward redistribution approaches in model training.

Revenue performance across varying numbers of tasks

Dataset	Methods	$N_\tau = 50$	$N_\tau = 100$	$N_\tau = 150$	$N_\tau = 200$	$N_\tau = 250$	$N_\tau = 300$	$N_\tau = 350$
Chengdu	Greedy	6.53 ±0.16	12.14 ±0.17	17.77 ±0.37	21.85 ±0.15	26.07 ±0.39	29.30 ±0.28	32.94 ±0.81
	IPPO	6.47 ±0.22	12.42 ±0.26	17.25 ±0.42	23.74 ±0.80	27.01 ±0.36	34.40 ±1.16	40.09 ±1.25
	PAU	6.57 ±0.22	11.65 ±0.20	18.36 ±0.88	24.18 ±0.79	26.94 ±0.81	33.57 ±0.92	39.88 ±1.86
	CA-noFE	6.59 ±0.18	12.52 ±0.31	18.10 ±0.55	24.40 ±0.80	27.43 ±0.73	34.55 ±1.06	41.14 ±1.53
	CAFE	**6.69 ±0.17**	**12.67 ±0.22**	**18.48 ±0.80**	**24.66 ±0.71**	**27.63 ±0.74**	**35.37 ±0.98**	**41.55 ±1.42**
Haikou	Greedy	6.42 ±0.14	11.78 ±0.40	17.03 ±0.22	21.52 ±0.39	26.16 ±0.50	29.18 ±0.39	31.25 ±0.42
	IPPO	6.17 ±0.18	12.23 ±0.43	17.30 ±0.32	22.84 ±0.32	27.19 ±0.24	32.48 ±0.88	35.03 ±1.38
	PAU	6.68 ±0.15	11.88 ±0.19	18.25 ±0.79	23.28 ±0.90	25.95 ±0.90	32.87 ±1.01	35.48 ±0.88
	CA-noFE	6.64 ±0.15	12.14 ±0.31	**18.52 ±0.59**	23.28 ±0.90	27.55 ±0.38	33.21 ±0.57	36.09 ±1.50
	CAFE	**6.71 ±0.15**	**12.43 ±0.35**	18.51 ±0.47	**23.40 ±0.87**	**28.37 ±0.45**	**34.15 ±0.69**	**36.99 ±0.97**

Revenue performance across varying numbers of workers

Dataset	Methods	$N_w = 50$	$N_w = 100$	$N_w = 150$	$N_w = 200$	$N_w = 250$	$N_w = 300$	$N_w = 350$
Chengdu	Greedy	7.63 ±0.42	15.67 ±0.51	19.81 ±0.38	21.85 ±0.15	22.88 ±0.37	23.76 ±0.36	24.22 ±0.28
	IPPO	10.45 ±0.34	20.82 ±1.14	23.48 ±1.01	23.74 ±0.80	22.79 ±0.31	24.42 ±0.20	23.71 ±0.42
	PAU	11.48 ±0.49	21.54 ±1.05	23.96 ±0.90	24.18 ±0.79	24.73 ±0.85	24.17 ±0.62	24.94 ±0.55
	CA-noFE	11.94 ±0.62	21.82 ±0.89	24.24 ±1.04	24.40 ±0.80	25.29 ±0.93	24.90 ±0.51	25.63 ±0.60
	CAFE	**12.17 ±0.49**	**21.85 ±1.02**	**24.60 ±0.93**	**24.66 ±0.71**	**25.66 ±0.86**	**25.69 ±0.58**	**26.56 ±0.59**
Haikou	Greedy	7.71 ±0.30	15.89 ±0.44	19.75 ±0.45	21.52 ±0.39	23.10 ±0.28	23.65 ±0.26	23.98 ±0.38
	IPPO	15.92 ±1.90	21.80 ±0.60	22.16 ±1.24	22.84 ±0.32	24.12 ±0.25	26.68 ±0.33	27.63 ±0.28
	PAU	16.38 ±1.73	22.58 ±0.38	23.14 ±1.02	23.28 ±0.90	23.98 ±0.91	26.89 ±0.60	**30.31 ±1.21**
	CA-noFE	16.60 ±1.83	22.87 ±0.45	23.21 ±0.79	23.28 ±0.90	24.51 ±0.80	27.26 ±0.49	29.87 ±0.55
	CAFE	**16.62 ±1.76**	**23.32 ±0.33**	**23.60 ±0.87**	**23.40 ±0.87**	**24.75 ±0.91**	**27.65 ±0.45**	30.02 ±0.54

* Please contact *ieyfli@zzu.edu.cn* for further information.