Reward Translation via Reward Machine in Semi-Alignable MDPs

Yun Hua*[1], Haosheng Chen*[2], Wenhao Li[3], Bo Jin[3], Baoxiang Wang[4], Hongyuan Zha[4], Xiangfeng Wang[2]

[1] *Shanghai Jiao Tong University*
[2] *East China Normal University*
[3] *Tongji University*
[4] *Chinese University of Hong Kong Shenzhen*

ICML
International Conference On Machine Learning

Abstract

Addressing reward design complexities in deep reinforcement learning is facilitated by knowledge transfer across different domains. To this end, we define *reward translation* to describe the cross-domain reward transfer problem. However, current methods struggle with non-pairable and non-time-alignable incompatible MDPs. This paper presents an adaptable reward translation framework *neural reward translation* featuring *semi-alignable MDPs*, which allows efficient reward translation under relaxed constraints while handling the intricacies of incompatible MDPs. Given the inherent difficulty of directly mapping semi-alignable MDPs and transferring rewards, we introduce an indirect mapping method through reward machines, created using limited human input or LLM-based automated learning. Graph-matching techniques establish links between reward machines from distinct environments, thus enabling cross-domain reward translation within semi-alignable MDP settings. This broadens the applicability of DRL across multiple domains. Experiments substantiate our approach's effectiveness in tasks under environments with semi-alignable MDPs.

Contributions

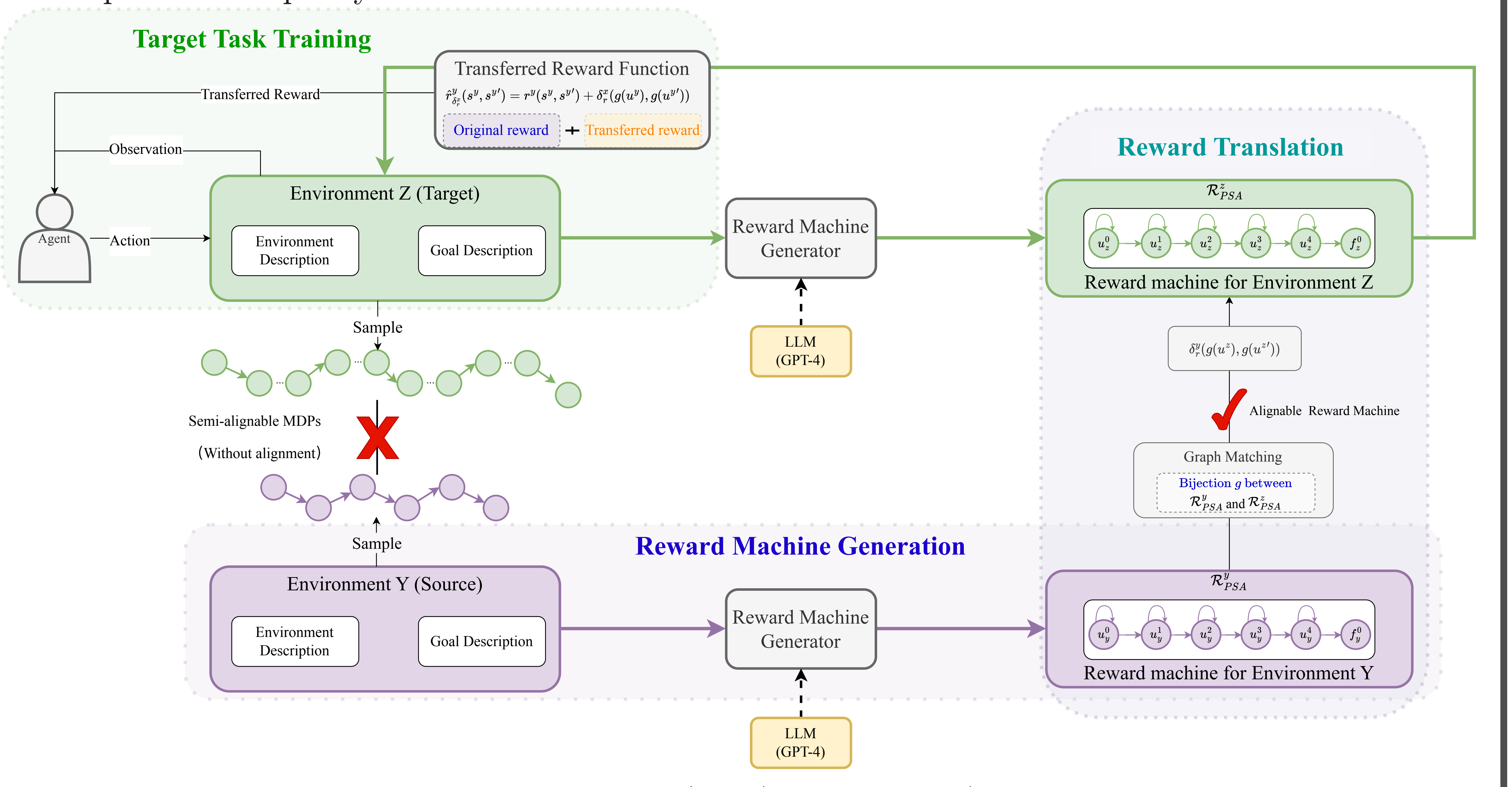
- The introduction of semi-alignable Markov decision processes, providing a crucial theoretical foundation for facilitating reward translation in cross-domain reinforcement learning and extending beyond alignable MDPs.
- The development of a novel framework called neural reward translation, designed to address the reward translation problem within semi-alignable MDPs by building upon the foundation provided by reward machines.
- The proposal of several semi-alignable environments, showcasing the effectiveness of the Neural Reward Translation approach in handling reward translation tasks where environments operate under semi-aligned MDPs.

Conclusion

In conclusion, this paper introduced the concept of *semi-alignable MDPs* alongside the *Neural Reward Translation* framework to facilitate *reward translation* in reinforcement learning to reduce reward design complexities. NRT employs reward machines to address reward translation challenges within semi-alignable MDPs and features an innovative large language model-based generator for the automatic generation of reward machines. Our method significantly enhancing training efficiency across various environments. Although challenges persist in constructing appropriate reward machines and deciphering relationships in complex tasks, future research endeavors will continue to explore the vast potential of semi-alignable MDPs and work towards broadening NRT's applicability in a diverse range of situations and domains.

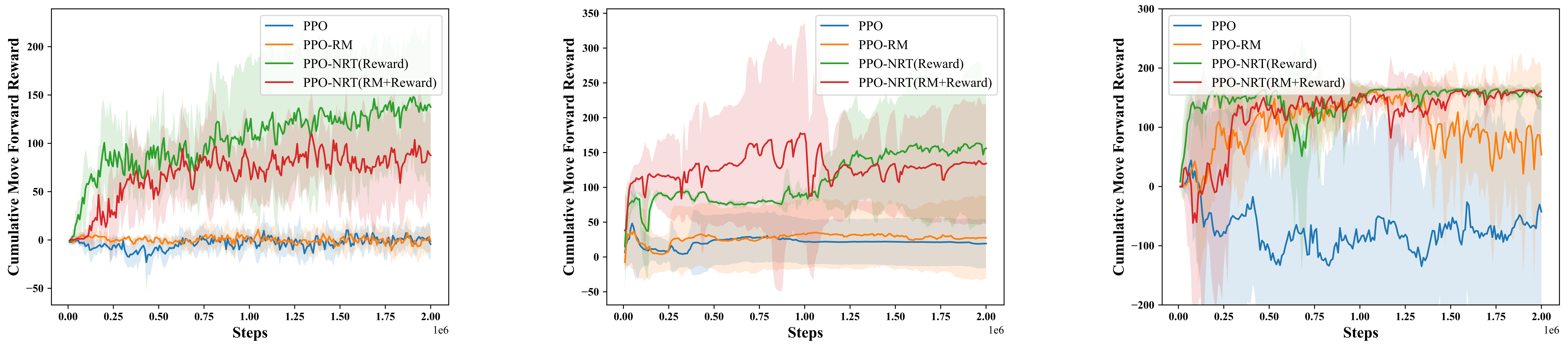
Overview of NRT

- **Reward Machine Generation** The primary role of this component is to construct reward machines by incorporating domain knowledge. Traditional approaches often rely on handcrafted design or learning from demonstration techniques. In this paper, an LLM-based reward machine generator is employed to automatically generate these reward machines. The LLM is given the environment, observation, action, and target descriptions, which provide details about the task and domain, as well as the definition of a reward machine. We utilize a chain of thought (COT) approach when designing the training prompt for the LLM.
- **Reward Translation** This component enables cross-domain reward translation in semi-alignable MDPs. NRT first constructs reward machines for both tasks, with the source task's RM incorporating dense rewards derived from its optimal value function. Through graph matching, NRT aligns these reward machines and transfers rewards based on two defined inter-machine relationships.
- **Target Task Training** This component guides the target task using transferred rewards from the source domain. During training, the agent receives observations from the target environment which are processed by the reward machine to determine states, then computes transferred rewards from these observation-state pairs, enabling standard RL algorithms to optimize the policy for cumulative reward maximization.



An overview of the Neural Reward Translation (NRT) framework: 1) Reward machine generation uses a generator to construct RMs based on task and environment descriptions. 2) Reward translation aligns source and target RMs using graph matching, facilitating reward transfer. 3) Target task training leverages transferred rewards for efficient learning.

Experiment Results (Part)



The learning curves for mujoco experiment(Ant, Hopper and Halfcheetah). To intuitively precept the learning process of the agent, we use the original reward provided by OpenAI-Gym which consists forward reward and control reward to show the learning curve.