*Code available at: https://github.com/duanwenbo/aaai22

Playing with Alchemy: A Benchmark and Evaluation for Meta-RL*

Wenbo Duan, Xiaoyang Wang
Department of Electrical and Electronic Engineering

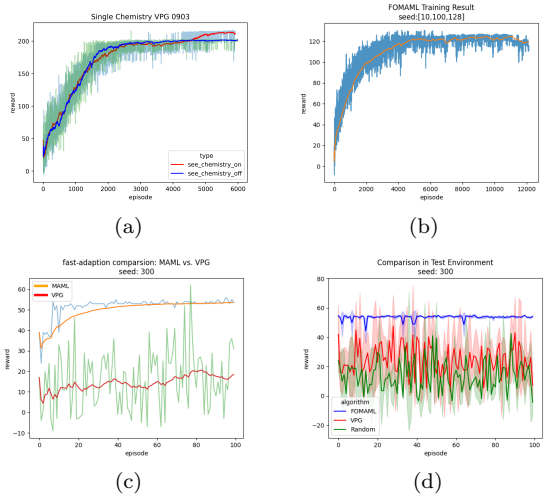
I. INTROCUTION

Meta reinforcement learning (meta-RL) has played an important role in the fast-adaptation of RL models. There haven't been many works on 'Alchemy', which is one of the few benchmarking platforms for meta-RL. Having implemented different RL algorithms on it, we concluded Alchemy as a challenging environment with complex latent structures, requiring high computational expense for evaluating meta-RL algorithms. By leveraging observations from studying the Alchemy, we design a customized environment with more flexible tasks. Three suggestions were proposed as ways to improve the effectiveness of the gradient-based meta-RL in subsequent experiments, with respect to robustness, computational expense and task distribution difficulties.

II. KEY RESULTS

Observations on the Alchemy Environment: Throughout trials of Vanilla Policy Gradient (VPG) and Proximal Policy Optimization (PPO) on a fixed 'chemistry' in Alchemy, we verified the effectiveness of the environment towards different types of reinforcement learning algorithms. The increasing and converging trends in Fig 1.a indicate the latent structure were being learnt by the agent.

However, meta-learning on this environment has been challenging. Regardless that Fig 1.b demonstrates the generalization ability of Model-Agnostic Meta-Learning (MAML) algorithm [1] in this environment, the final scores are far lower compared with the Oracle in [2]. According to our experiment, $1e5$ training episodes averagely takes about 26 hours, but an ideal training in the paper usually required the number of training steps in millions, which could take 100 times longer with our current computing resources.

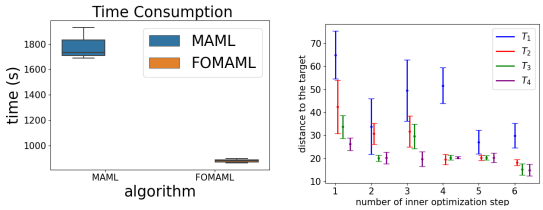


(a) (b)

(c) (d)

Figura 1. **a**: The single task performance validate the effectiveness of the algorithm and platform. **b**: The MAML result achieved a converge result but the final score is low. **c**,**d**: Two downside figures shows the effectiveness of the meta-learning compared with random, pre-trained algorithms in fine-tuning and testing scenarios.

Through our experiments, we conclude that despite the Alchemy achieves structural richness and structural transparency, the combination of 167,424 possible chemistries in the task distribution lead to a significantly difficult environment structure for the meta-RL agent to explore. Hence, benchmarking the performance of the algorithm in a non-industrial environment is not recommended.



(a) (b)

Figura 2. **a:** Time consumption of FOMAML is 40 % less than MAML. **b:** Average distance and standard deviation during training result. Increasing inner-optimization steps greatly increases the stability of training.

A Flexible Customized Environment: Inspired by issues on the Alchemy platform, we created a customized navigation environment which is inherited from the OpenAI gym library, designed with obstacles as well as the flexibility to switch between the obstacles distributions and task distributions. These flexibility made the environment a convenient tool to do fast verification and algorithms research.

Three Findings to Improve Training Performance: Based on the initial experiment result, we analyzed the factors that affect the performance of gradient-based meta reinforcement learning. Three aspects were found by exploring the customized environment. Multi-step inner optimization and the first-order approximation are verified with positive impact to some extend, while the variation of task distributions is found as a strong negative factor. As illustrated in Fig 2.a and 2.b, we concluded three useful suggestions for a more stable training:

A: Increasing the inner-optimization steps in gradient-based meta-RL algorithms is a way to balance between quick-adaptation and stable training, and hence, achieving a more robust model.

B: First Order MAML (FOMAML) can achieve a similar result with 40 % less time consumption than MAML.

C: Designing a difficulty classifier in task distribution is expected to lead to a better performance result.

[1] C. Finn, P. Abbeel, and S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *ICML* (2017) pp. 1126–1135.

[2] J. X. Wang, M. King, N. Porcel, *et al.*, Alchemy: A structured task distribution for meta-reinforcement learning, arXiv preprint arXiv:2102.02926 (2021).