META-WORLD CONDITIONAL NEURAL PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Meta-World Conditional Neural Process (MW-CNP), a conditional world model generator that leverages sample efficiency and scalability of Conditionbal Neural Processes Garnelo et al. (2018) architecture to allow an agent to sample from the generated world model. We intend to reduce agent's interaction with the target environment as much as possible. Thus, we designed a model-based meta-RL framework where the RL agent can be conditioned on significantly fewer samples collected from the target environment to imagine the unseen environment. We emphasize that the agent does not have access to the task parameters throughout training and testing.

1 INTRODUCTION

Diverse set of skills can be acquired through incremental learning on environments with emergent complexity. In line with this, world model generation has been an integral part of open-ended learning Wang et al. (2019). Although creating a variety of challenging simulation environments is a promising direction, it is bounded by the limitations of the physics engine and the predetermined set of environment parameters. To overcome this problem, we propose generating world models from agent's own experience.

Research on lifelong learning focus on learning incrementally and adapting rapidly to unknown tasks from incoming streams of data Parisi et al. (2019). Challenges associated with lifelong learning include catastrophic forgetting, negative transfer, and resource limitations. In addition, data to be processed after deployment and optimal model hyperparameters are not available a priori.

Meta reinforcement learning is an enduring area of interest that enables fast adaptation in test tasks. Recent works have integrated meta learning to offline RL Mitchell et al. (2020) and lifelong learning Finn et al. (2019); Nagabandi et al. (2018); Berseth et al. (2021). In open-ended learning, a policy with high representational capacity can be utilized for increasingly challenging environments. In this work we are interested in few shot learning in settings where the transition dynamics of the environment changes across tasks. More specifically, the agent is trained on a distribution of tasks with varying transition dynamics and expected to adapt to an unseen task with few samples from the target environment. Sampling in real world is expensive, hence reducing the number of samples used for fast adaptation is an important research direction for sim-to-real RL.

We propose Meta-World CNP, to generate world models from few samples collected from the target environment. Our contributions include (1) generating world model from varying number of observations, (2) aggregating the experience collected from the target environment with experience collected from the world model.

2 PRELIMINARIES

2.1 CONDITIONAL NEURAL PROCESSES

CNPs are proposed as a novel neural architecture by Garnelo et al. (2018) to predict the parameters of a probability distribution conditioned on a permutation invariant prior data and target input. CNPs attempt to represent a family of functions using Bayesian Inference and high representational capacity of neural networks. The architecture of CNPs consist of parameter sharing encoders and a decoder named as the query network. In particular, first a random number of input and true output

pairs $(x_{f^i}, y_{f^i}^{true})$ are sampled from the function $f^i \in F$. Each pair is encoded into a latent representation via the encoder networks. Then, these representations averaged along the dimensions to account for the invariance to permutation and number of inputs. The resulting representation is concatenated with the target input query and fed to the query network. The query network outputs the predicted mean and standard deviation for the queried input $(x_{f^i}^q)$.

2.2 No-Reward Meta Learning

NORML, proposed by Yang et al. (2019), is an extension of MAML-RL framework Finn et al. (2017) for settings where the environment dynamics change across tasks instead of the reward function. The goal of NORML is to utilize past experience to quickly adapt to tasks with unknown parameters from a few samples with missing reward signals. Provided that the change in dynamics can be represented in $(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$, NORML learns an pseudo-advantage function $A_\psi\left(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}\right)$. It is important to note that the aim of A_ψ is to guide the meta-policy adaptation instead of fitting to the advantage function. A_ψ is used to compute task specific parameters in the MAML inner loop from a set of state transitions of task i D_i^{train} that does not contain reward signal. The learned advantage function is optimized in the MAML outer loop using the reward information present in rollouts D_i^{test} obtained from updated task-specific policy parameters.

3 Our Proposal: Meta-World Conditional Neural Process (MW-CNP)

In few-shot learning, the goal is to quickly adapt to an unseen task using few labeled data in the target environment. Meta-World Conditional Neural Processes (MW-CNP) can generate world models using few samples from the target environment.

Problem Setup We denote the initial state distribution as $\rho(s_0) : \mathcal{S} \to \mathbb{R}$, state transition distribution of $task_i$ as $\rho_i(s_{t+1}|s_t, a_t) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and the reward function as $r_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. During NORML meta-training we store transitions for each $task_i$ as a set of observations $O_i = \{(s_t, a_t, s_{t+1})\}_{t=0}^{t-1} \subset S \times A \times S$ without the task parameter. After we obtain the meta policy and learn the advantage function, we train our MW-CNP model using the offline replay buffer of transitions.

MW-CNP Training The loss function of MW-CNP can be expressed as

$$\mathcal{L}(\theta_E, \theta_D) = -\log P\left(s_q'^{\,true} \mid f_{\theta_D}\left([s^q, a^q] \oplus \frac{1}{n}\sum_k^n g_{\theta_E}\left((s_k, a_k, s_k')\right)\right)\right)$$

where f_{θ_D}, g_{θ_E} are the decoder the encoder networks, $[s^q, a^q]$ is the target state action query, (s_k, a_k, s_k') are the randomly sampled transitions from the set of observations O_i.

MW-CNP Evaluation During evaluation, the agent is allowed to collect few samples from the unseen target environment. These samples are encoded into representations of fixed size by an encoder network with shared weights. A shared representation of the task is obtained using an averaging module. It is important to note that the shared representation can be obtained from randomly selected number of data points. This representation is used to predict the parameters of the distribution of next state when state action tuples are queried.

The learned advantage function uses (s, a, s') tuple as input and outputs an advantage estimation value. Hence, once experience from the generated world model is obtained, these experiences are fed to the learned advantage function in the form of (s, a, s'). Together with the estimated advantage values, and the experience from the real target environment the adapted policy parameters are computed for fast adaptation to the target task.

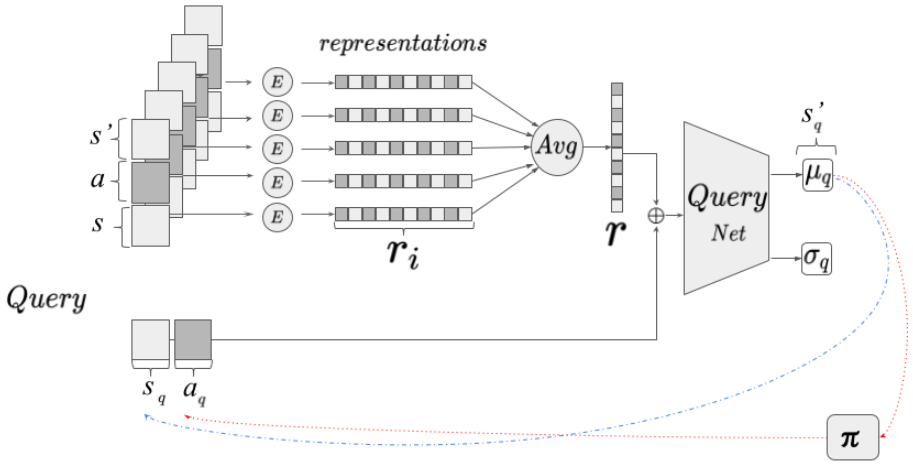

Figure 1: Structure of the Meta World-CNP (shapes are adapted from Seker et al. (2019)).

4 EXPERIMENTS

In this section, we analyze the performance of our method and compare it with NORML in 2D point agent. The goal of the point agent, which is initialized at [x=0,y=0], is to move to the position [x=0,y=1], where x,y are the positions on the 2D plane. The reward function, that is used in RL training, is the same across all tasks which is the negative Euclidean distance from the goal position. Different tasks are created by generating different artificial force fields that push the agent in different directions (ϕ) during the movement of the agent. The movement of the agent during RL roll-outs consist of 10 samples.

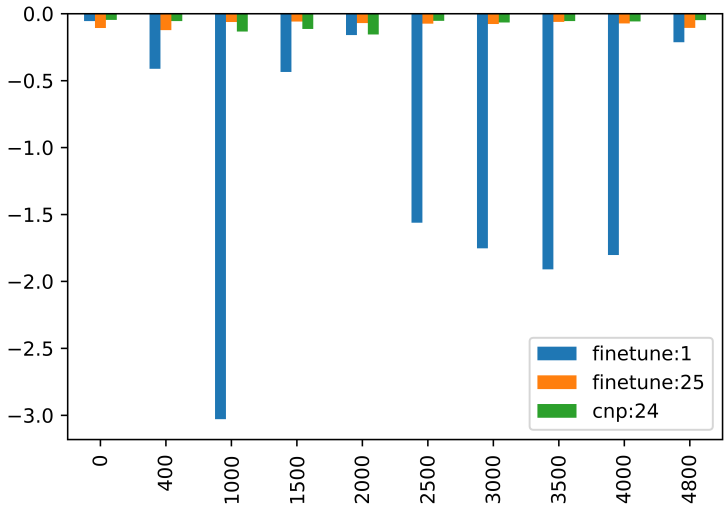

Figure 2: Expected reward (y-axis) of finetuning over generated data and ground truth data in target tasks. X-axis represents the target task indices.

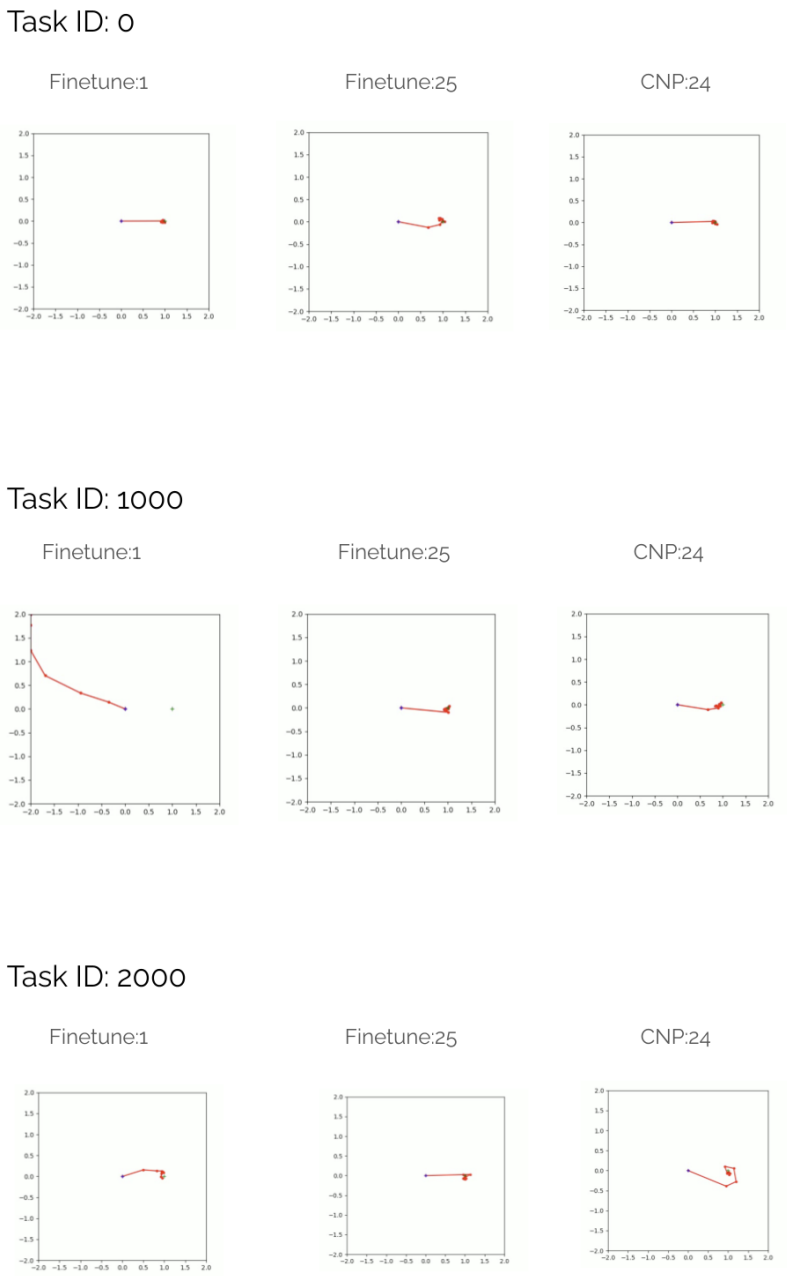

Figure 3: Illustrations of trajectories for Tasks 0, 1000, and 2000

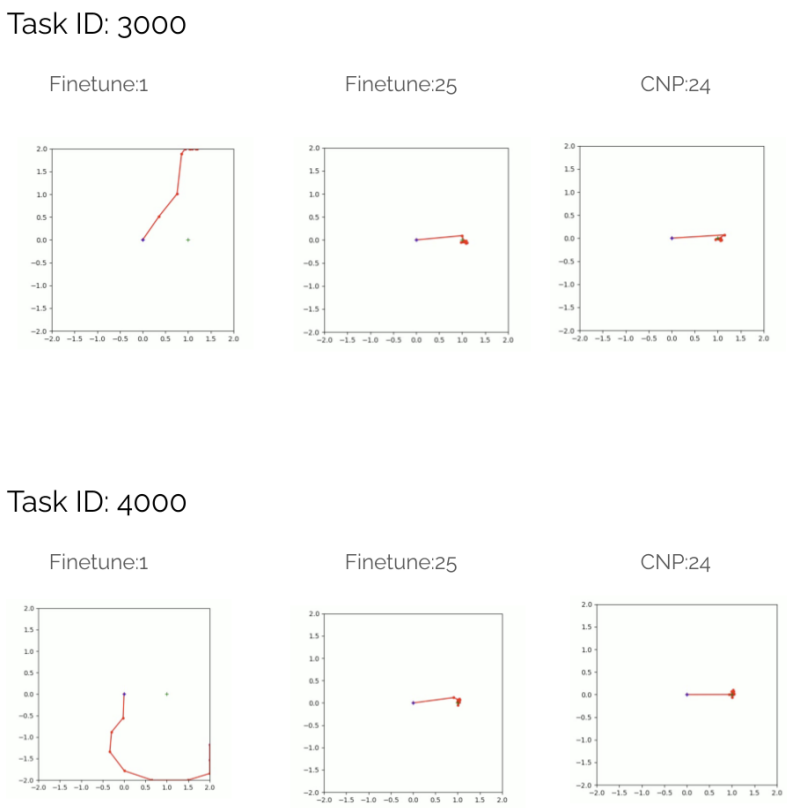

Figure 4: Illustrations of trajectories for Tasks 3000, and 4000

The agent is initially trained in 5000 tasks, i.e. in environment with 5000 different force field. Then, it is tested in a novel target task. The rollouts obtained from the experiment will be referred as actual rollouts whereas the rollouts obtained from the imagination (world model) of our system will be named as imagined rollouts. Faced with the target task and environment, our method MW-CNP used 1 actual rollout and 24 imagined rollouts for finetuning, whereas the baseline NORML method used 1 actual rollout in one condition and 25 actual rollouts in another condition. The average expected reward obtained in the target environment over 25 indepedent runs are shown in Fig. 2 with green bars. As shown, when MW-CNP and NORML used the same amount of actual rollouts (1 actual rollout) from the target environment, MW-CNP outpeforms NORML. Even when NORML used significantly higher number of actual rollouts than MW-CNP (25 to 1), their performances are similar as shown in the figure.

Figure 2 shows the average expected reward obtained in the target environment over tasks. 5000 tasks are defined over the $[-\pi, \pi]$ interval. We compare results for NORML finetuning over 1 rollout, NORML finetuning over 25 rollouts(ground truth) and finetuning over the augmentation 24 rollouts from the MW-CNP model and 1 rollout from the target environment over 10 tasks.

We condition the MW-CNP model on the same rollout used for 1 rollout-finetunig. The results show that the samples generated from the agent's imagination created by the MW-CNP generated sophisticated samples that can be used for finetuning the meta policy for fast domain adaptation. All in all, using only 1/25 the interaction with the target environment used in NORML, the agent can adapt better or comparable to the unseen task.

5 Discussion

We showed that meaningful experience can be collected using the MW-CNP framework that can guide the meta-policy adaptation.We compared expected return obtained from finetuning over generated data with MW-CNP and ground truth data. While our initial experiments are conducted in low-dimensional input space, CNPs were shown to perform well with high dimensional inputs like image data Seker et al. (2019). Therefore, we plan to extend our work and apply it to high-dimensional sensorimotor spaces such as to manipulator robots with rgb-d cameras.

References

Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, and Sergey Levine. Comps: Continual meta policy search. *CoRR*, abs/2112.04467, 2021. URL https://arxiv.org/abs/2112.04467.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1920–1930. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/finn19a.html.

Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. *CoRR*, abs/1807.01613, 2018. URL http://arxiv.org/abs/1807.01613.

Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline meta-reinforcement learning with advantage weighting. *CoRR*, abs/2008.06043, 2020. URL https://arxiv.org/abs/2008.06043.

Anusha Nagabandi, Chelsea Finn, and Sergey Levine. Deep online learning via meta-learning: Continual adaptation for model-based RL. *CoRR*, abs/1812.07671, 2018. URL http://arxiv.org/abs/1812.07671.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Muhammet Yunus Seker, Mert Imre, Justus H Piater, and Emre Ugur. Conditional neural movement primitives. In *Robotics: Science and Systems*, 2019.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley. Paired open-ended trailblazer (POET): endlessly generating increasingly complex and diverse learning environments and their solutions. *CoRR*, abs/1901.01753, 2019. URL http://arxiv.org/abs/1901.01753.

Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Jie Tan, and Chelsea Finn. Norml: No-reward meta learning. *arXiv preprint arXiv:1903.01063*, 2019.