KAIST AI602 Project Report: Compositional Meta-RL

Yoonyoung Cho [*1] **Jisu Han** [*1]

Abstract

In recent years, meta-reinforcement learning (meta-RL) methods have achieved impressive results in robot learning. Such methods enable agents to learn new tasks by making use of prior experience. However, current meta-RL methods cannot generalize on challenging out-of-distribution tasks. We hypothesize that to solve these challenging scenarios, meta-learning agents need to develop *compositional* reasoning; by doing so, the agent can quickly learn to perform novel tasks by reusing previously learned components. In this work, we propose to implement this intuition on a meta-learning framework. To this end, we take inspiration from the recent success in multi-task learning with modularization. To validate our method, we experiment with our model on a custom MetaWorld-ML4 benchmark, a challenging robot manipulation domain which necessitates out-of-distribution generalization. We confirmed our method is comparable with prior algorithms by tasks' success rate on our custom Meta-World benchmark experiments.

1. Introduction

For practical robots, we need algorithms that can learn how to quickly solve new tasks by utilizing prior experience. In fig. 1, the robot can quickly generalize to tasks such as *furniture assembly* by exploiting prior knowledge in other tasks such as *woodworking* or *box-packing*. Meta-learning is a potential solution that can *learn to learn* novel tasks by leveraging experiences on prior tasks. Specifically, in the context of reinforcement learning (RL), meta-RL has emerged as a promising resolution to the chronic overfitting problem of RL agents that cannot generalize to seemingly trivial distributional shifts.

[*]Equal contribution [1]KAIST Graduate School of AI, Seoul, Korea. Correspondence to: Yoonyoung Cho <yoonyoung.cho@kaist.ac.kr>, Jisu Han <jshan@kaist.ac.kr>.

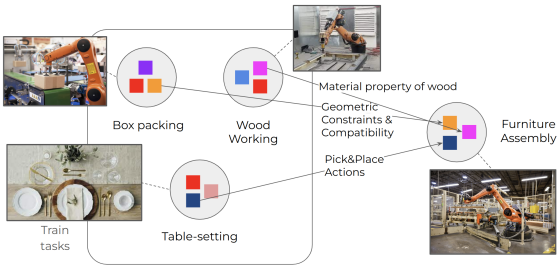


Figure 1. Illustration of compositional generalization in RL for robot manipulation; the robot learns contextual elements such as material properties, action affordances and constraints from prior tasks, which accelerate learning and bootstrap performance on novel tasks.

However, current meta-RL methods cannot generalize to diverse tasks – with only 40% success rate in Meta-World (Yu et al., 2020). We believe this is because these methods lack the ability to develop *compositional* reasoning of the task. Humans commonly solve complex tasks by decomposing them into easier sub-tasks and then combining the sub-task solutions. For example, opening a doorknob is composed of common sub-tasks such as reaching, grasping, and rotating objects. Once we solve such sub-tasks, we can intuitively open other doors that comprise similar actions. This type of compositional reasoning permits reuse of the sub-task solutions when tackling future tasks that share part of the underlying compositional structure (Mendez et al., 2022).

There are two main lines of prior works. In meta-learning, MAML (Finn et al., 2018) learns initial parameters at first, and then quickly converges by gradient-descent updates. FAMLE (Kaushik et al., 2020) also shares the same intuition as MAML, solely learned by multiple initial parameters. Since the parameters are disjointly learned per each training task, these *gradient-based* meta-learning approaches are architecturally limited from sharing parameters, which eventually prohibits compositional generalization.

In contrast, on *context-based* meta-learning approaches, PEARL (Rakelly et al., 2019) learns the probabilistic encoder to infer context variables. However, it fails to generalize to novel tasks (Yu et al., 2020). We believe this is due to their lack of compositional reasoning. While PEARL

may learn arbitrary context structures, it is well-known that compositional architectures are unlikely to emerge by chance (Liska et al., 2018). Therefore, we believe we must provide architectural bias to PEARL in order to enforce compositionality.

To facilitate compositional learning, we take inspiration from the recent success in multi-task learning with modularization (Purushwalkam et al., 2019; Yang et al., 2020). These works leverage the modularization of a network to learn common, reusable parts across tasks. Especially, (Yang et al., 2020) show that such a policy can learn multiple tasks efficiently with the modularized network and one-hot encoded task ID. However, these works are not directly applicable to meta-RL since they require *a priori* information about each task, given as task ID or language descriptions.

On the other hand, PEARL (Rakelly et al., 2019) can autonomously learn and recognize task context based on interaction history. In light of this, we propose to meta-learn task contexts with PEARL (Rakelly et al., 2019), while adopting a modularized context-conditioned policy architecture from SM (Yang et al., 2020).

2. Problem Formulation

We formulate our problem as a *goal-conditioned meta-RL* problem. Concretely, given the set of meta-training tasks $\mathcal{T}_{1:N}$ from the task distribution $p(\mathcal{T})$, we seek to generalize to a novel task $\mathcal{T}'$ from the same task distribution. Each task is formulated as a goal-conditioned MDP, which is defined as $\mathcal{T} = \big(S, A, G, P(s_{t+1}|s_t, a_t), R(s_t, a_t; g), \gamma\big)$.

Here, S is the state space, A is the action space, G is the goal space, $P(s_{t+1}|s_t, a_t)$ is the transition distribution, $R(s_t, a_t; g)$ is the reward function, and γ is the discount factor. To make the problem more practical, instead of assuming access to the explicit goal $g \in G$, we only provide $l \in L$ as the text description about the goal, such as "push the red button on the table". During meta-training process, given a set of training tasks sampled from $p(\mathcal{T})$, the agent learns a policy that adapts to the task by quickly recognizing the *context* c. Context c is a set of elements which contain tasks and domain-level information, expressed as $c = \{\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_N\}$. To infer the current context, the agent leverages the history of past transitions $H_t = (o_1, a_1, r_1, \cdots, o_t, a_t, r_t)$ to condition the latent context inference module $\phi(c_t|c_{t-1}, H_{1:t}, l)$. At test-time, we first seek to recognize the current context c_t, and then quickly adapt the policy $\pi(\cdot)$ to solve the new task to maximize the expected return.

3. Related Works

Context-based meta-RL Prior works have applied context variables to meta-RL domain. This approach is called *context-based* meta-RL, since these methods adapt to new tasks by aggregating experience into a latent representation on which the policy is conditioned. PEARL (Rakelly et al., 2019) infers context variables during meta-training and rapidly adapts on meta-test time by updating the context variables via data samples collected through exploration. PEARL significantly outperforms the performance from prior meta-RL methods (i.e. MAML (Finn et al., 2018) and RL^2 (Duan et al., 2016)) on MuJoCo simulator domain. However, on Meta-World benchmark (Yu et al., 2020), prior meta-RL methods outperform PEARL by more than 2 times. Since Meta-World benchmark is evaluated on task distributions that are sufficiently broad to enable generalization to new behaviors, PEARL cannot achieve sufficient generalization on diverse domains. We expect our method will be able to generalize well by leveraging compositional reasoning.

Policy Modularization Recently, soft modularization (Yang et al., 2020) has achieved remarkable success in a multi-task learning setup in which a single policy must generalize to multiple different tasks. The key ingredient behind their success was *modularization*, where the base policy consists of multiple inter-connected modules, reconfigured by the routing network that dynamically determines their connectivity based on the given task.

This work achieves *compositional generalization* by "infinite use of finite means" (Chomsky, 1965); or, in other words, re-using modular components of the policy across diverse tasks. By splitting the policy into modules, they avoid the problematic cross-task interference (Yang et al., 2020) which is harmful to the agent's training process; however, by routing the shared modules, they also enable the transfer of shared modules across multiple tasks. By the combination of these two strategies, the policy can learn shared, reusable elements of the task. However, this work is not directly applicable to novel tasks as in the Meta-RL setup, since the routing network requires hard-coded information of the current task ID.

4. Our Approach

PEARL (Rakelly et al., 2019) can autonomously learn task contexts from interaction data, yet lack the ability to compositionally generalize. Soft modularization (Yang et al., 2020) learns a compositional policy and a routing network that can generalize well across multiple tasks, yet they require hard-coded information about the task *a priori*, such as task ID.

We observe that these two works are complementary: by leveraging the meta-learning capability of PEARL (Rakelly

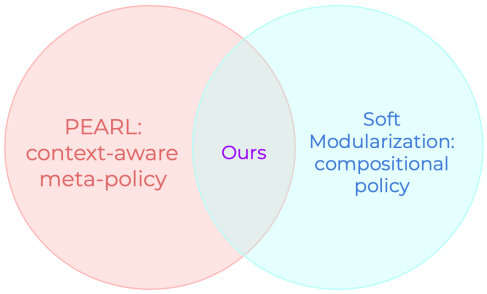


Figure 2. Conceptual illustration of our approach. By leveraging PEARL (Rakelly et al., 2019) for context recognition and modularized policy from soft-modularization (Yang et al., 2020), we seek to achieve a policy that compositionally generalizes to novel tasks.

et al., 2019) while also leveraging the compositionality of SM (Yang et al., 2020), we can achieve a *compositional meta-policy* that can generalize to novel tasks. Intuitively, our strategy is to first recognize the context, and then adaptively reconfigure the policy modules to respond to the current task.

Concretely, we utilize PEARL (Rakelly et al., 2019) to learn a set of compositional contexts to condition Soft Modularization (Yang et al., 2020). We train a context inference network from PEARL (Rakelly et al., 2019) jointly with the policy. The routing network in Soft Modularization (Yang et al., 2020) is conditioned on the context to reconfigure the base policy in response to the current task. This way, we can train a context-conditioned policy that can generalize to novel contexts which are compositions of prior elements.

Training. The overall meta-training pipeline is shown in alg. 1, adapted from PEARL (Rakelly et al., 2019). As with PEARL, we iterate between the data-collection process (L2-12) and the training steps (L13-25) for each of $i = 1 \dots T$ training tasks, where the policy is conditioned on the goal-aware task-specific context z_i derived from the goal embedding g_i and the interaction history H^i. To realize this, we modify the pipeline to (1) include the goal embedding g_i, l_i. In practice, we utilize the CLIP (Radford et al., 2021) embeddings to compute the latent representation of the goal for the context encoder. Moreover, we (2) adopt the routing network that re-configures the modularized base policy with $p_{1:N}$ conditioned on the current context z for the N modules of the base policy.

Architecture. Our (CMRL) model architecture is shown in fig. 3. We utilize the context network formulated as a 2-layer MLP with 128 hidden channels, where given the interaction history h_t and the goal embedding g, the context z is sampled at each step from $z_t \sim p(z_t|h_t; g)$ where p is product of multivariate gaussians $p = \prod_{j=1}^{N} \mathcal{N}(\mu_j, \sigma_j)$ for each gaussian $\mathcal{N}(\mu_j, \sigma_j)$ computed for each frame j in

Algorithm 1 CMRL Meta-training Loop

Require: Batch of training tasks $\{\tau_i\}_{i=1}^{T}$ from $p(\tau)$, learning rates $\alpha_1, \alpha_2, \alpha_3$
Initialize replay buffer $\mathcal{B}^i = \emptyset$ for each training task
while not done **do**
  for each τ_i **do**
    Interaction history $H^i = \emptyset$
    for $k = 1, \dots, K$ **do**
      Query goal embedding g_i from text l_i
      Sample $z \sim q_\phi(z|H^i, g_i)$
      Get Routing weights $p_{1:N} \sim \phi_R(z)$
      Gather data from $\pi_\theta(a|s, z, w)$ and add to $\mathcal{B}^i$
      Update $H^i = \{(s_j, a_j, s'_j, r_j)\}_{j:1\dots N} \sim \mathcal{B}^i$
    end for
  end for
  for step in training steps **do**
    for each τ_i **do**
      Sample interaction history $H^i \sim \mathcal{B}^i$ and rollout $b^i \sim \mathcal{B}^i$
      Query goal embedding g_i from text l_i
      Sample $z \sim q_\phi(z|H^i, g_i{}')$
      $\mathcal{L}^i_{actor} = \mathcal{L}_{actor}(b^i, z)$
      $\mathcal{L}^i_{critic} = \mathcal{L}_{critic}(b^i, z)$
      $\mathcal{L}^i_{KL} = \beta D_{KL}(q(z|H^i)||r(z))$
    end for
    $\phi \leftarrow \phi - \alpha_1 \nabla_\phi \sum_i \left(\mathcal{L}^i_{critic} + \mathcal{L}^i_{KL}\right)$
    $\theta_\pi \leftarrow \theta_\pi - \alpha_2 \nabla_\theta \sum_i \mathcal{L}^i_{actor}$
    $\theta_Q \leftarrow \theta_Q - \alpha_3 \nabla_\theta \sum_i \mathcal{L}^i_{critic}$
  end for
end while

$h \sim B^i$. The routing network is a stack of $4\times$ modules as in Soft-Modularization (Yang et al., 2020) where the context z_t acts as the gating variable for the inputs to the routing networks. The module weights $p_{1:N}$ are multiplied by the bottleneck connections between the actor and critic networks. The whole network is trained end-to-end based in an offline-RL fashion with the SAC (Haarnoja et al., 2018) loss.

5. Experiments

To validate our compositional model, we experiment with an adaptation of the Meta-world Benchmark (Yu et al., 2020). For this experiment, we come up with a custom ML4 dataset.

Meta-Learning 4 (ML4): Few-shot adaptation to new test task with 3 meta-training tasks. With the objective to test generalization to new task, we hold out 1 test task and meta-train policies 3 tasks. We randomize object and goal positions and intentionally select training tasks with compositional similarity to the test task. Table 1 shows a

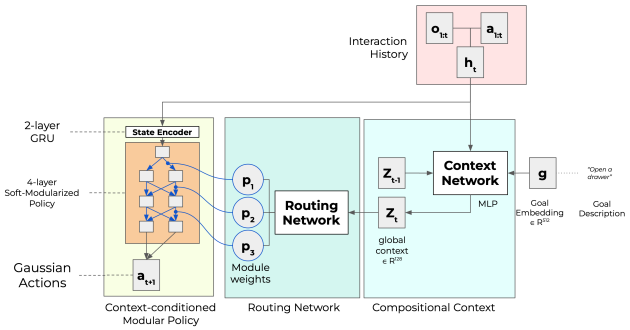


Figure 3. Our model architecture; given the interaction history and the goal embedding, we sample the context pertaining to the current task distribution. The routing network, conditioned on the current context, dynamically re-configures the modularized policy to output the actions parameterized as multivariate gaussians.

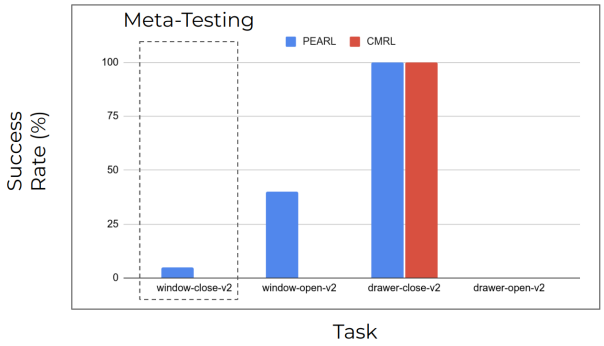
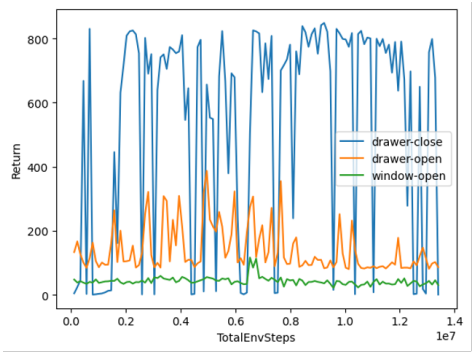


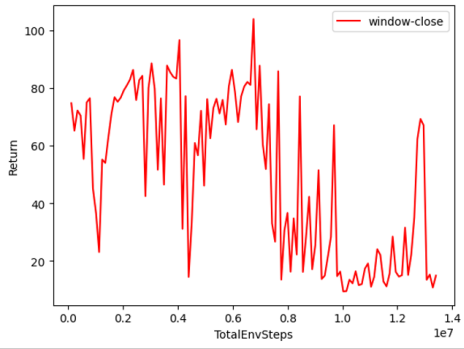
Figure 4. Comparison of success rates on the ML4 benchmark. Our model performs on comparable with baseline on CLOSE-DRAWER, but under-performs in others, including the meta-testing task.

list of meta-train/meta-test tasks and a description of each task: by meta-training on opening and closing drawers, we expect the agent to recognize the meaning of opening and closing; by operating the window in OPEN-WINDOW task, we expect the agent to learn the affordances of the window. From these prior experiences, we anticipate that the agent would generalize to the CLOSE-WINDOW task.

Results. We compare our method to PEARL (Rakelly et al., 2019): an off-policy actor critic meta-RL algorithm. We show each task's success rate results on the custom ML4 in fig. 4 and the demonstration example in fig. 6. We observe that our method is comparable with PEARL by tasks' success rate. Fig. 5 shows the return value of each episode during meta-training time and meta-test time. We confirmed CLOSE-DRAWER task takes a high return compared to other meta-training tasks. We believe this happens since the CLOSE-DRAWER task is so easy to achieve, the policy overfits that task.



(a) Episode return, for meta-training tasks



(b) Episode return, for meta-testing tasks

Figure 5. Return value of meta-training tasks and meta-testing task.

6. Conclusion

We presented CMRL, a compositional approach for context-based meta-RL. Currently, our algorithm does not perform as well as the baselines. We believe that this is due to overfitting to a single task; since the policy collapses to only solve a single task, we seek to introduce additional modules for enforcing compositionality and improved context recognition across multiple tasks. The proposed modification is shown in fig. 7

References

Chomsky, N. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl ²: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Finn, C., Xu, K., and Levine, S. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.

	Task	Goal description
Meta-train	CLOSE-DRAWER	Push and close a drawer. Randomize the drawer positions.
	OPEN-DRAWER	Open a drawer. Randomize drawer positions.
	OPEN-WINDOW	Push and open a window. Randomize window positions.
Meta-test	CLOSE-WINDOW	Push and close a window. Randomize window positions.

Table 1. A list of meta-train/meta-test tasks and a description of each task.

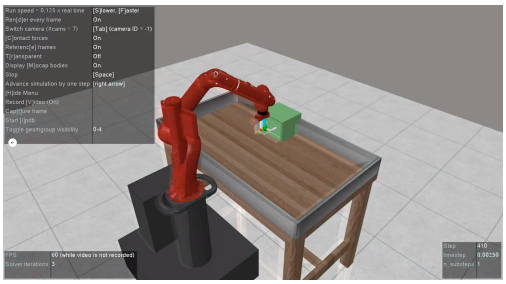


Figure 6. Demonstration of the robot successfully performing CLOSE-DRAWER task.

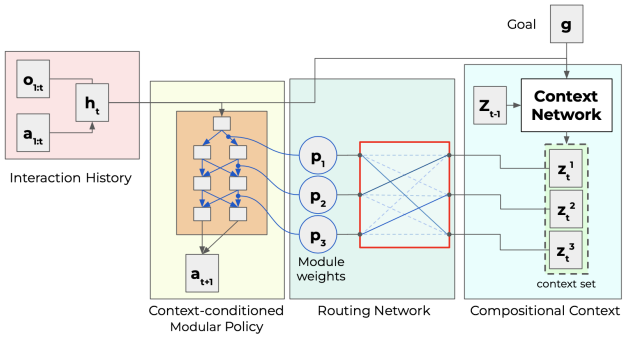


Figure 7. Proposed next iteration of our algorithm.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Kaushik, R., Anne, T., and Mouret, J.-B. Fast online adaptation in robotics through meta-learning embeddings of simulated priors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

Liska, A., Kruszewski, G., and Baroni, M. Memorize or generalize? searching for a compositional rnn in a haystack. *ArXiv*, abs/1802.06467, 2018.

Mendez, J. A., van Seijen, H., and Eaton, E. Modular lifelong reinforcement learning via neural composition. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5XmLzdslFNN.

Purushwalkam, S., Nickel, M., Gupta, A., and Ranzato, M. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3593–3602, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rakelly, K., Zhou, A., Finn, C., Levine, S., and Quillen, D. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pp. 5331–5340. PMLR, 2019.

Yang, R., Xu, H., WU, Y., and Wang, X. Multi-task reinforcement learning with soft modularization. In *Advances in Neural Information Processing Systems*, 2020.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.