



北京大學
PEKING UNIVERSITY

Robust Task Representations for Offline Meta-Reinforcement Learning via Contrastive Learning

ICML 2022

Haoqi Yuan, Zongqing Lu

School of Computer Science, Peking University

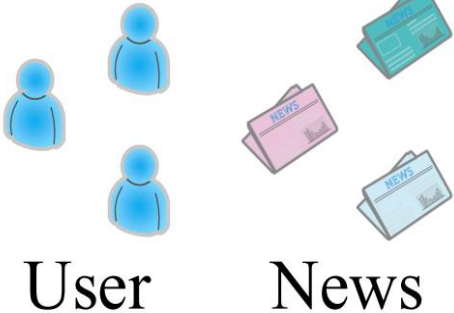
Deep Reinforcement Learning



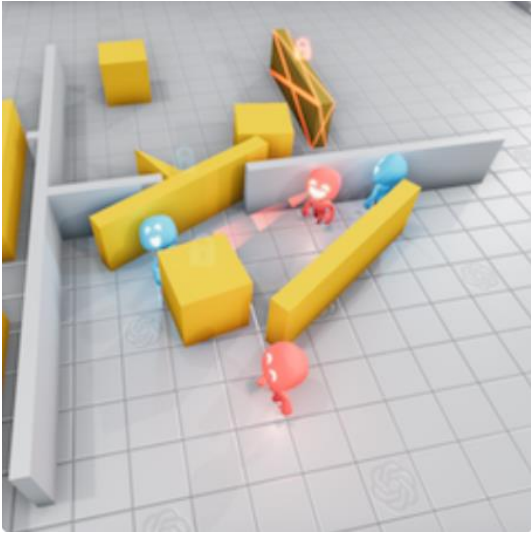
[Julian et al., 2020]



[Finn et al., 2017]



[Zheng et al., 2018]



[Baker et al., 2020]

Deep Reinforcement Learning

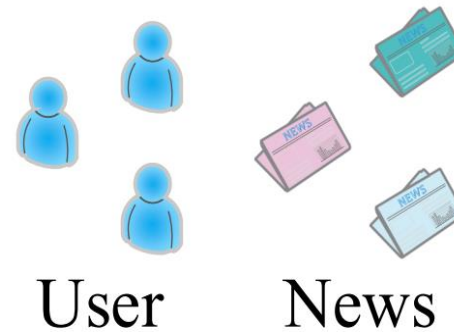
Data efficiency & Generalization



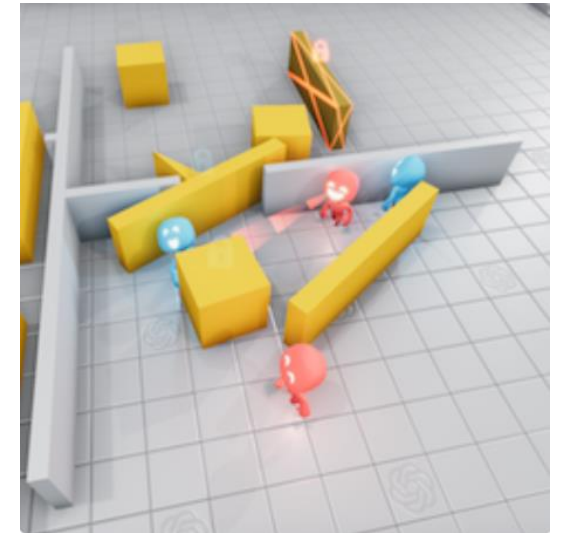
[Julian et al., 2020]



[Finn et al., 2017]



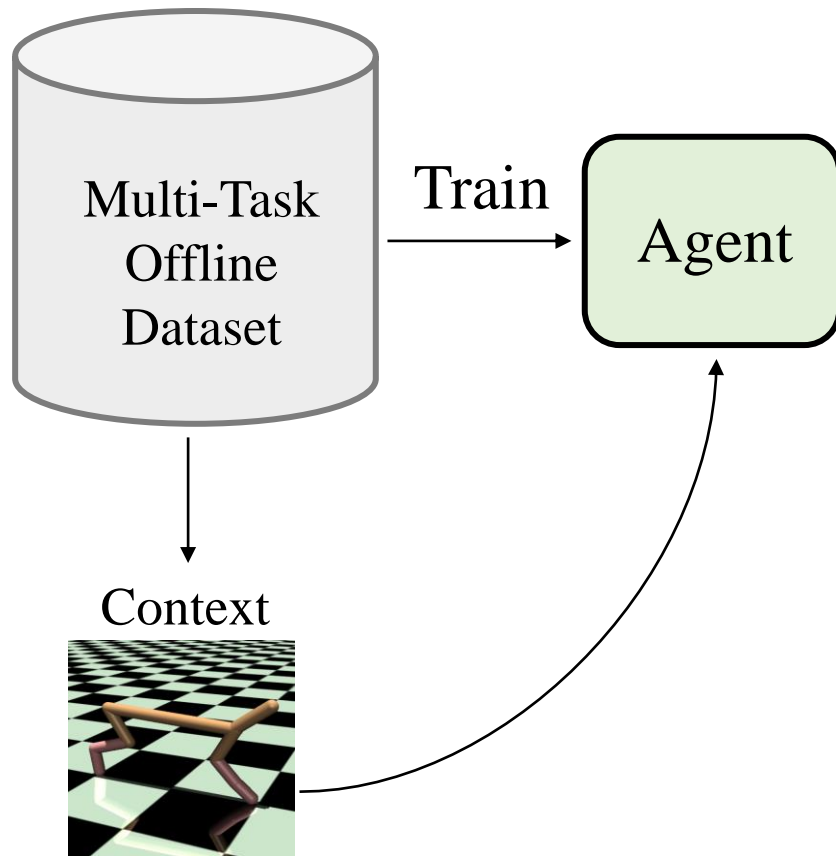
[Zheng et al., 2018]



[Baker et al., 2020]

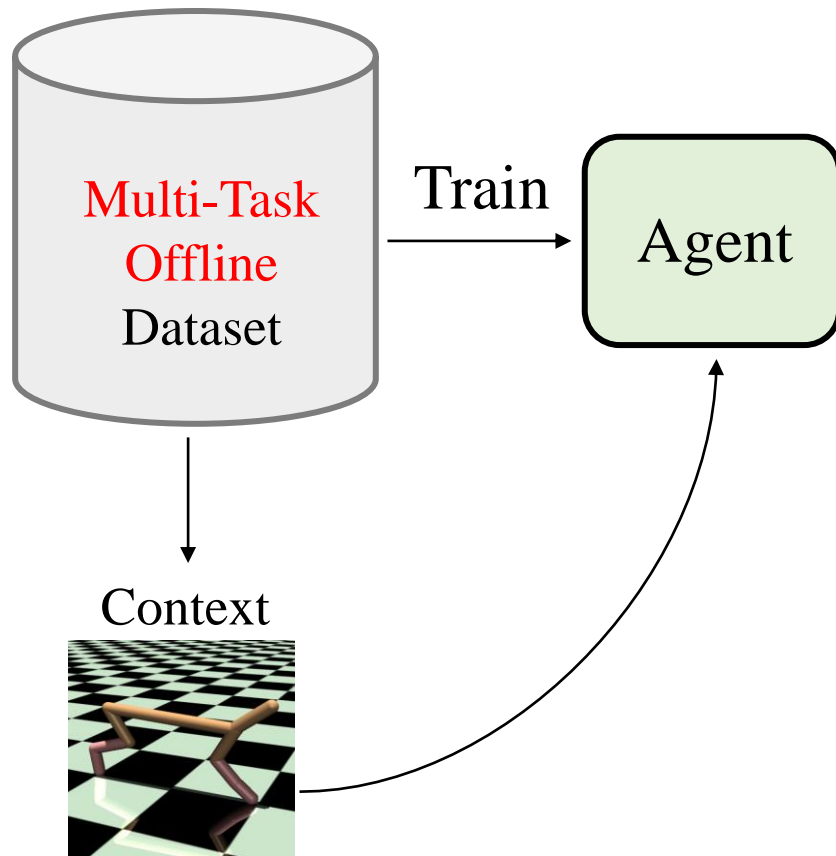
Offline Meta-RL

Meta-Training



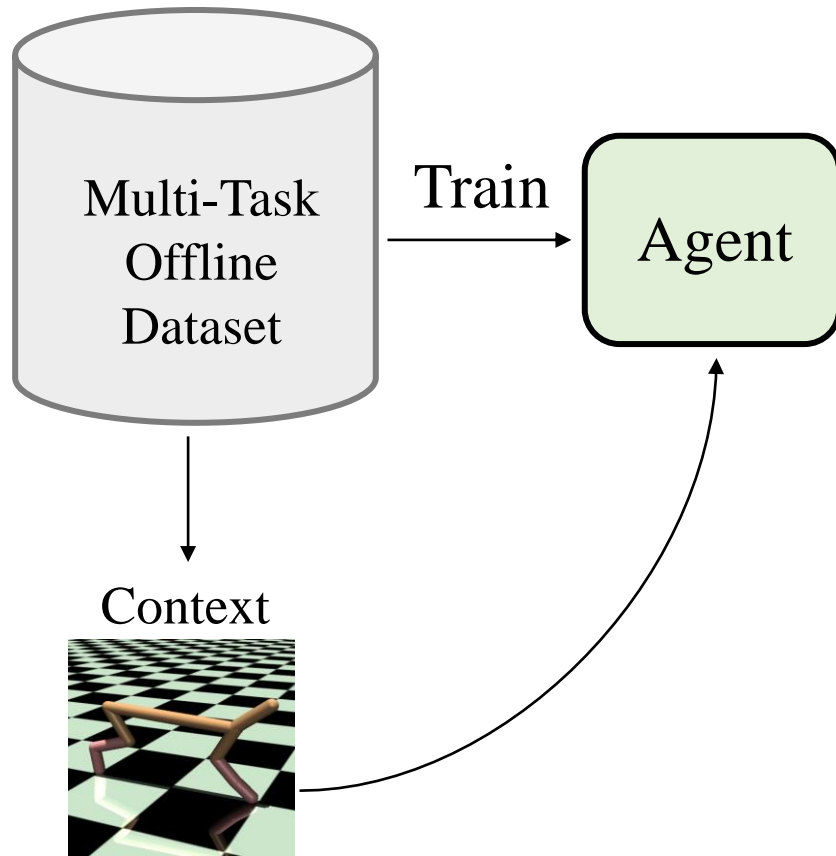
Offline Meta-RL

Meta-Training

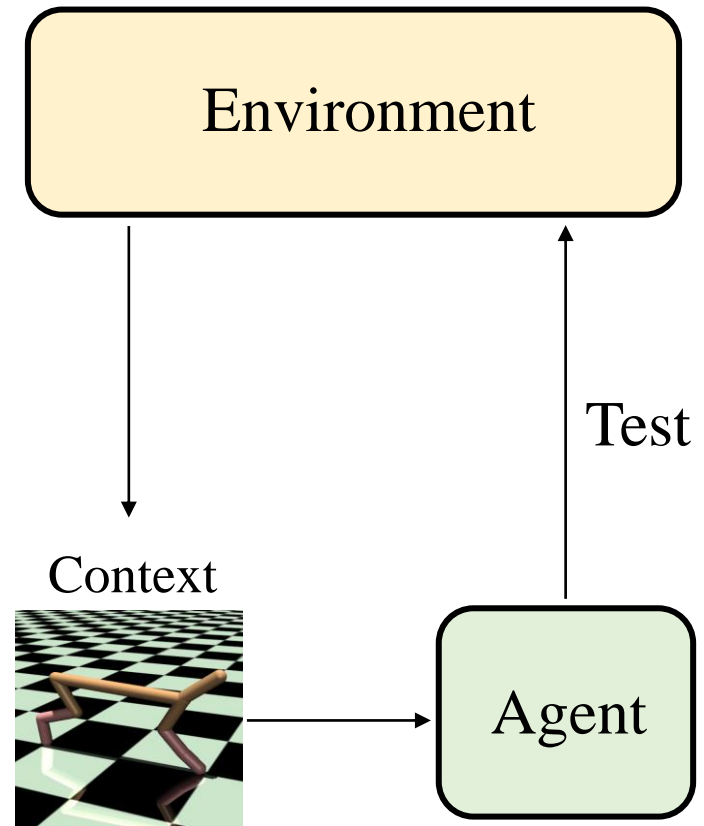


Offline Meta-RL

Meta-Training

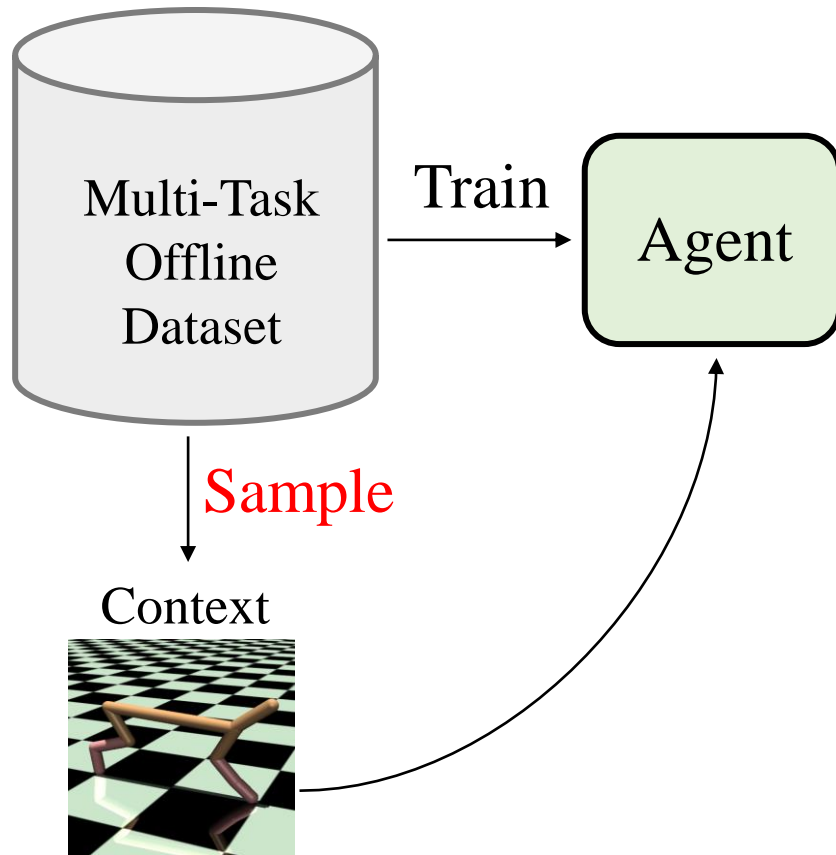


Meta-Test

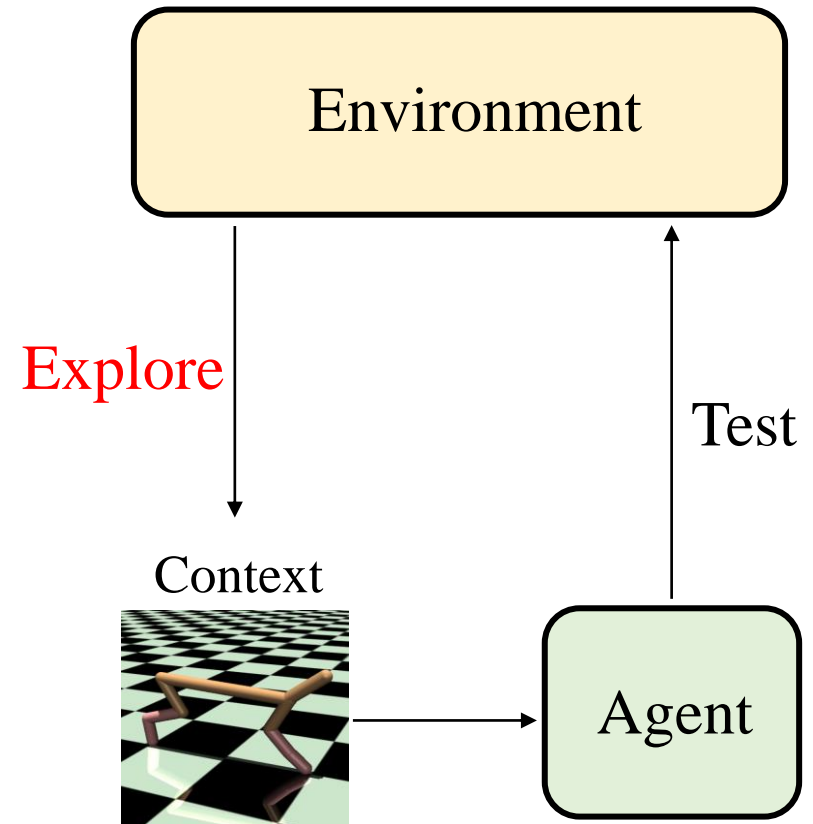


Offline Meta-RL

Meta-Training



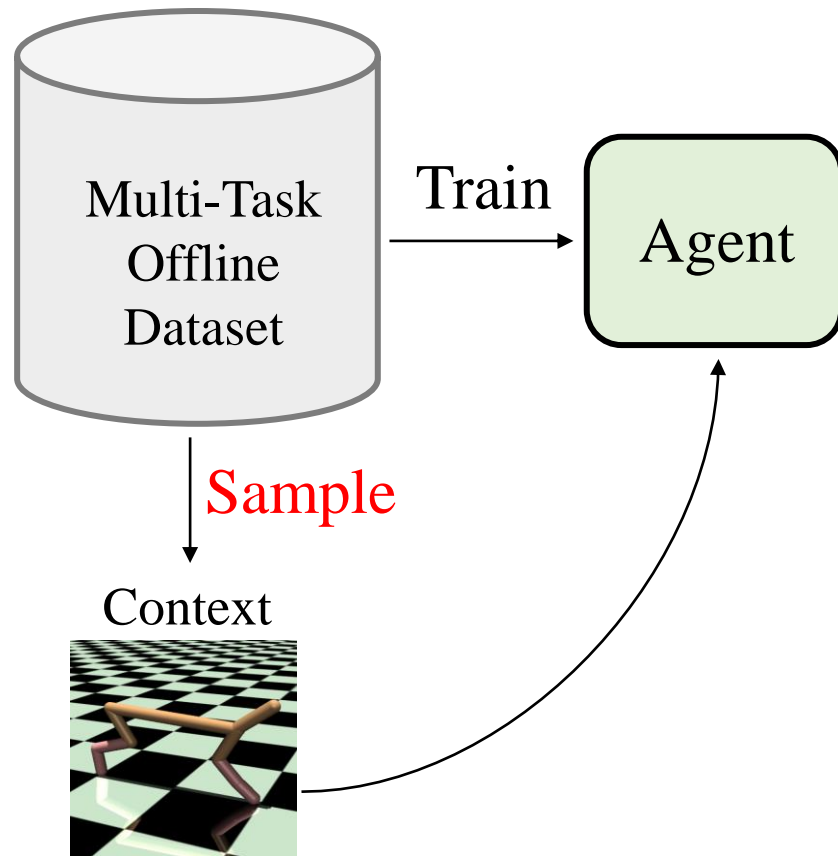
Meta-Test



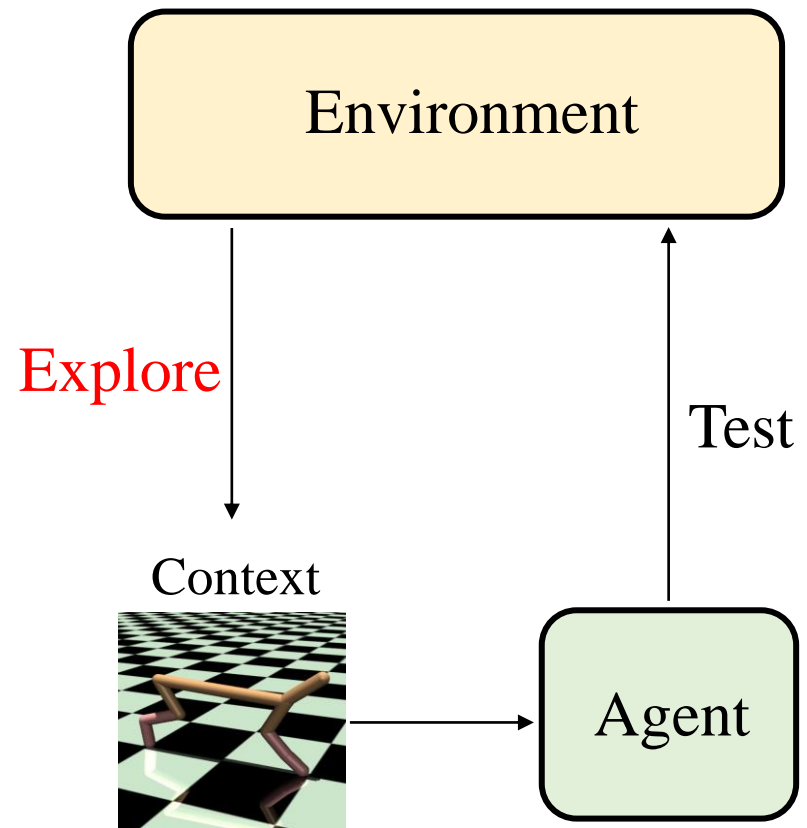
Offline Meta-RL

Distribution Shift

Meta-Training



Meta-Test



Offline Meta-RL

- Li et al. FOCAL: Efficient Fully Offline Meta-Reinforcement Learning Via Distance Metric Learning and Behavior Regularization. ICLR, 2021.
- Li et al. Provably Improved Context-Based Offline Meta-RL with Attention and Contrastive Learning. arXiv:2102.10774, 2021.
- Lin et al. Model-Based Offline Meta-Reinforcement Learning with Regularization. ICLR, 2022.
- Mitchell et al. Offline Meta-Reinforcement Learning with Advantage Weighting. ICML, 2021.
- Li et al. Multi-Task Batch Reinforcement Learning with Metric Learning. ICLR, 2020.
- Dorfman et al. Offline Meta Learning of Exploration. arXiv:2008.02598, 2020.
- Pong et al. Offline Meta-Reinforcement Learning with Online Self-Supervision. arXiv:2107.03974, 2021.

Offline Meta-RL

- Li et al. FOCAL: Efficient Fully Offline Meta-Reinforcement Learning Via Distance Metric Learning and Behavior Regularization. ICLR, 2021.
- Li et al. Provably Improved Context-Based Offline Meta-RL with Attention and Contrastive Learning. arXiv:2102.10774, 2021.
- Lin et al. Model-Based Offline Meta-Reinforcement Learning with Regularization. ICLR, 2022.
- Mitchell et al. Offline Meta-Reinforcement Learning with Advantage Weighting. ICML, 2021.
- Li et al. Mu
- Dorfman et al. Offline Meta Learning of Exploration. arXiv:2008.02598, 2020.
- Pong et al. Offline Meta-Reinforcement Learning with Online Self-Supervision. arXiv:2107.03974, 2021.

Context-based methods

Hard to generalize to out-of-distribution context in test.

Offline Meta-RL

- Li et al. FOCAL: Efficient Fully Offline Meta-Reinforcement Learning Via Distance Metric Learning and Behavior Regularization. ICLR, 2021.
- Li et al. Provably Improved Context-Based Offline Meta-RL with Attention and Contrastive Learning. arXiv:2102.10774, 2021.

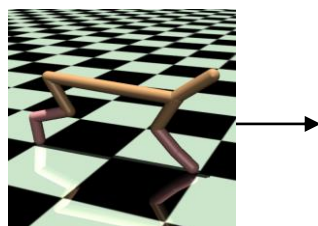
Require additional supervision or exploration.

- Lin et al. Model-Based Offline Meta-Reinforcement Learning with Regularization. ICLR, 2022.
- Mitchell et al. Offline Meta-Reinforcement Learning with Advantage Weighting. ICML, 2021.

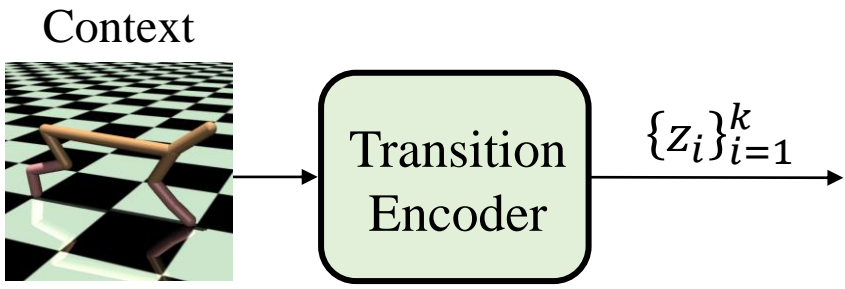
- Li et al. Multi-Task Batch Reinforcement Learning with Metric Learning. ICLR, 2020.
- Dorfman et al. Offline Meta Learning of Exploration. arXiv:2008.02598, 2020.
- Pong et al. Offline Meta-Reinforcement Learning with Online Self-Supervision. arXiv:2107.03974, 2021.

CORRO - Framework

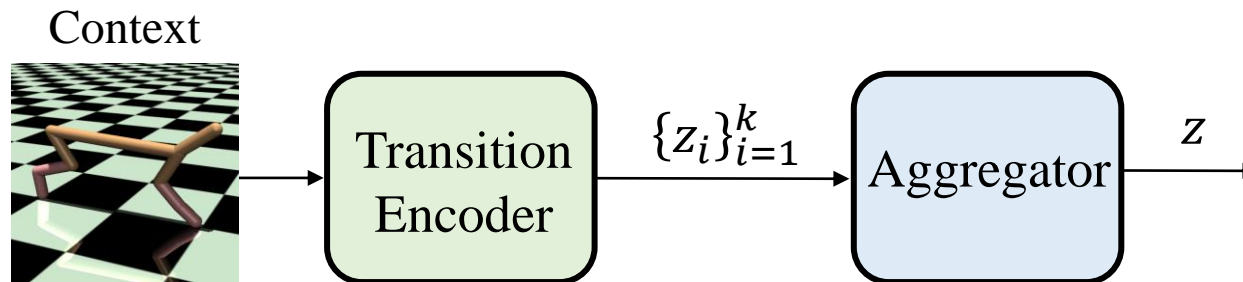
Context



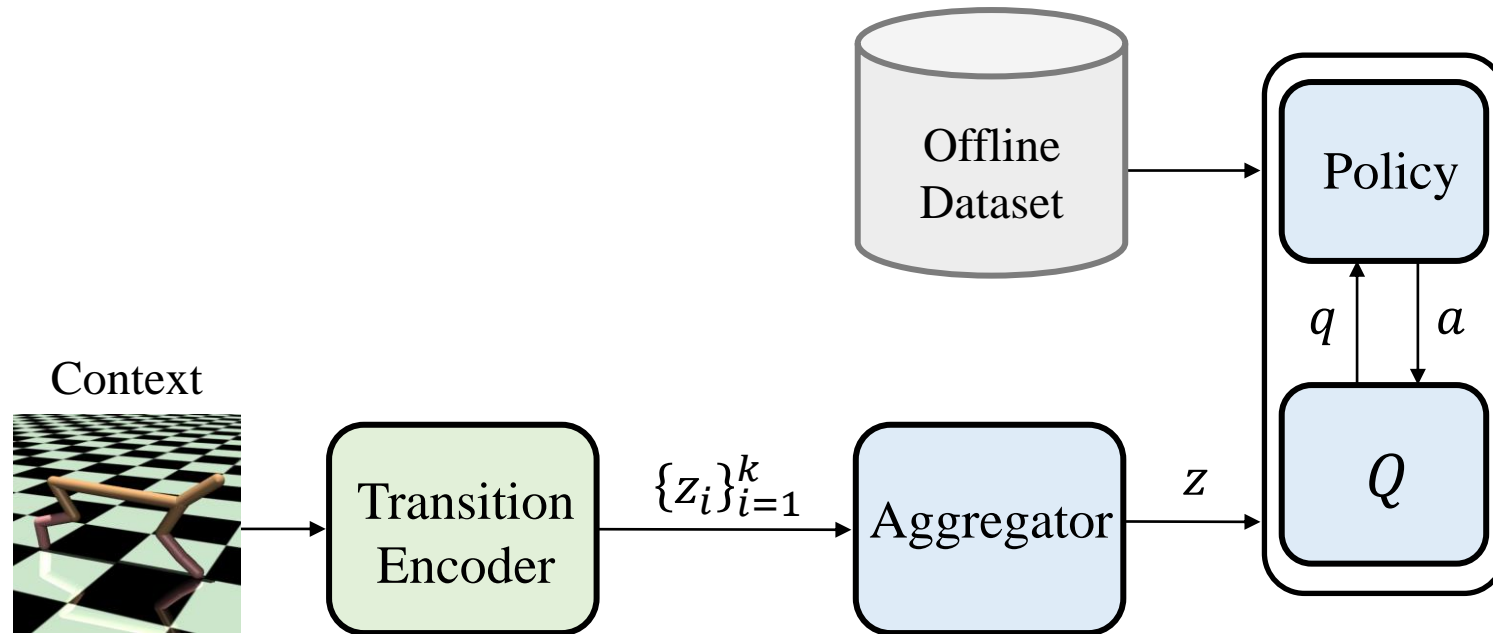
CORRO - Framework



CORRO - Framework

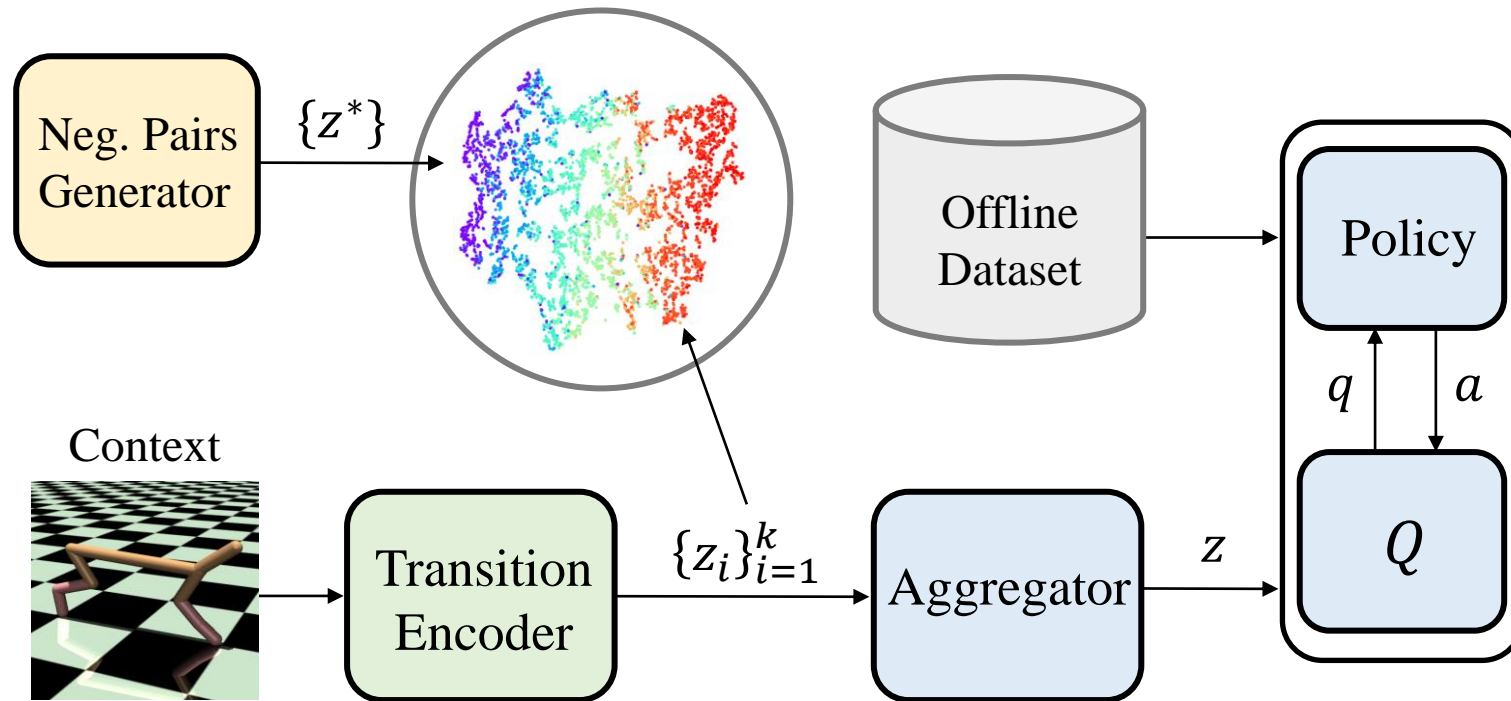


CORRO - Framework



Offline Meta-RL

CORRO - Framework



Contrastive Learning

Offline Meta-RL

CORRO - Task Representation Learning

- Mutual information maximization between the representation and the task:

$$\max I(z; M) = \mathbb{E}_{z, M} \left[\log \frac{p(M|z)}{p(M)} \right]$$

CORRO - Task Representation Learning

- Mutual information maximization between the representation and the task:

$$\max I(z; M) = \mathbb{E}_{z, M} \left[\log \frac{p(M|z)}{p(M)} \right]$$

- Optimization with contrastive learning:

$$\max_{\theta_1} \sum_{\substack{M_i \in \mathcal{M} \\ x, x' \in X_i}} \left[\log \left(\frac{\exp(S(z, z'))}{\sum_{M^* \in \mathcal{M}} \exp(S(z, z^*))} \right) \right]$$

CORRO - Task Representation Learning

- Mutual information maximization between the representation and the task:

$$\max I(z; M) = \mathbb{E}_{z, M} \left[\log \frac{p(M|z)}{p(M)} \right]$$

- Optimization with contrastive learning:

$$\max_{\theta_1} \sum_{\substack{M_i \in \mathcal{M} \\ x, x' \in X_i}} \left[\log \left(\frac{\exp(S(z, z'))}{\sum_{M^* \in \mathcal{M}} \exp(S(z, z^*))} \right) \right]$$

positive pair

negative pairs

- S : cosine similarity.
- z, z' : representations of two samples from task M_i .
- z^* : the representation of a sample ‘generated’ by MDP M^* . z^* and z share the same (s_t, a_t) pair.

CORRO - Negative Pairs Generation

- We should approximately generate the negative samples drawn from the distribution $P_{M^* \sim P(M)}(r, s' | s, a)$.

CORRO - Negative Pairs Generation

- We should approximately generate the negative samples drawn from the distribution $P_{M^* \sim P(M)}(r, s' | s, a)$.
- **Generative Modeling:**
 - Approximate the data distribution $p(r, s' | s, a, z)$.
 - Train a conditional VAE on the datasets.

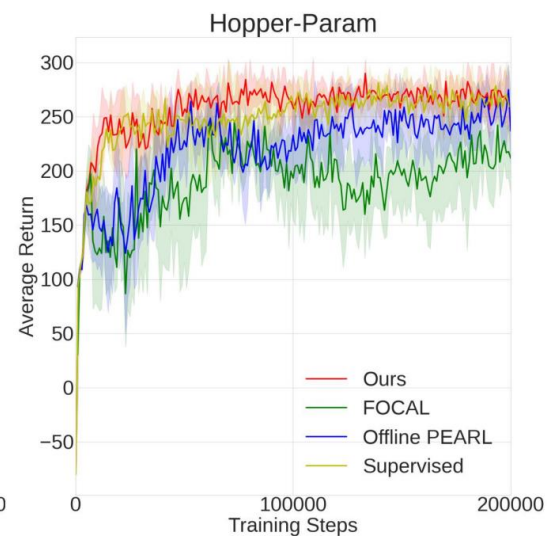
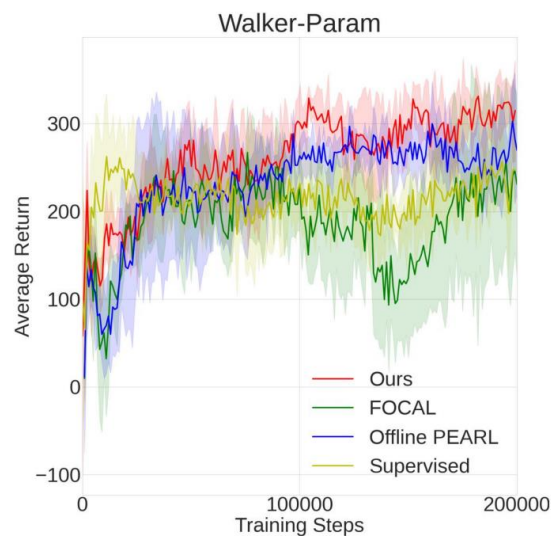
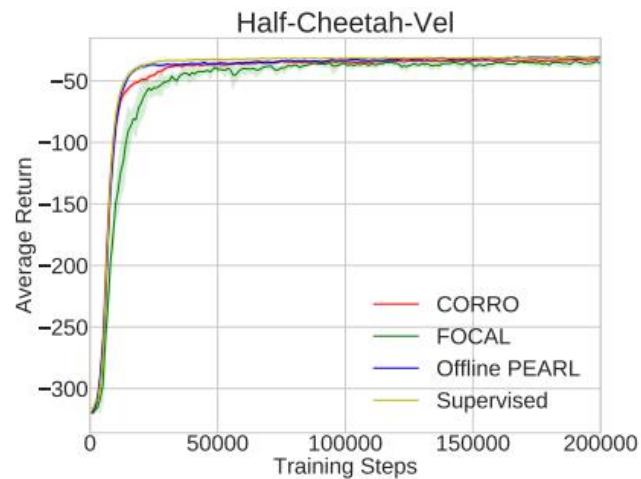
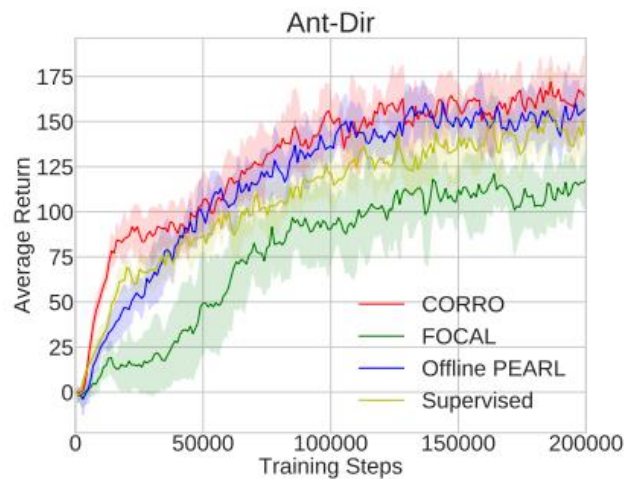
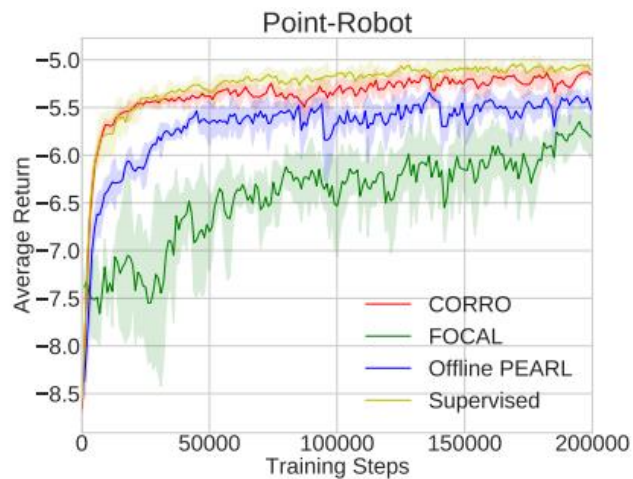
CORRO - Negative Pairs Generation

- We should approximately generate the negative samples drawn from the distribution $P_{M^* \sim P(M)}(r, s' | s, a)$.
- **Generative Modeling:**
 - Approximate the data distribution $p(r, s' | s, a, z)$.
 - Train a conditional VAE on the datasets.
- **Reward Randomization:**
 - Add random perturbation to the reward, to imitate the sample from other tasks.
 - Tasks should not differ in transition functions.

Experiments

- Meta-test performance in diverse task distributions and offline datasets.
- The robustness of task representations.
- How to choose the strategy of negative pairs generation.

Experiments - Tasks Adaptation



Experiments - Robust Task Inference

Environment	Supervised.		Offline PEARL		FOCAL		CORRO	
	<i>IID</i>	<i>OOD</i>	<i>IID</i>	<i>OOD</i>	<i>IID</i>	<i>OOD</i>	<i>IID</i>	<i>OOD</i>
Point-Robot	-4.89 \pm 0.10	-5.84 \pm 0.14	-5.4 \pm 0.17	-6.74 \pm 0.19	-6.06 \pm 0.42	-7.34 \pm 0.20	-5.19 \pm 0.05	-6.39 \pm 0.05
Ant-Dir	136 \pm 17.6	131.7 \pm 11.4	155.4 \pm 24.4	141.5 \pm 11.3	109.8 \pm 12.8	53.5 \pm 16.4	156.8 \pm 35.2	154.7 \pm 25.8
Half-Cheetah-Vel	-31.6 \pm 0.7	-32.1 \pm 0.9	-31.2 \pm 0.5	-242.7 \pm 6.0	-38.0 \pm 4.0	-204.1 \pm 9.5	-33.7 \pm 1.1	-89.7 \pm 7.4
Walker-Param	232.7 \pm 29.2	221.2 \pm 43.4	259.1 \pm 48.2	254.7 \pm 35.8	225.4 \pm 56.4	193.3 \pm 151.5	301.5 \pm 37.9	284.0 \pm 19.3
Hopper-Param	269.2 \pm 20.3	251.9 \pm 28.8	244.0 \pm 18.5	236.6 \pm 18.5	195.6 \pm 62.3	199.7 \pm 51.9	267.6 \pm 25.6	268.0 \pm 13.8

Out-of-distribution context.

Experiments - Robust Task Inference

Environment	Supervised.	Offline PEARL	FOCAL	CORRO
Point-Robot	-5.32 ± 0.20	-7.06 ± 0.99	-8.64 ± 0.26	-5.59 ± 0.57
Ant-Dir	149.8 ± 20.5	148.4 ± 35.3	89.8 ± 8.7	163.0 ± 35.8
Half-Cheetah-Vel	-37.6 ± 0.8	-35.4 ± 1.8	-41.6 ± 3.3	-42.9 ± 0.7
Walker-Param	221.7 ± 91.1	276.6 ± 37.7	245.6 ± 67.8	300.5 ± 34.2
Hopper-Param	253.5 ± 21.2	245.9 ± 18.9	203.6 ± 46.6	273.3 ± 3.9

Randomly explored context.

Experiments - Negative Pairs Generation

	Method	Contrastive Loss	<i>IID</i> Return	<i>OOD</i> Return
Half-Cheetah-Vel	Generative	0.07	-33.7 ± 1.1	-89.7 ± 7.4
	Randomize	0.83	-34.3 ± 1.5	-84.5 ± 1.3
	Relabeling	0.04	-40.8 ± 1.5	-245.3 ± 12.9
	None	1.20	-34.1 ± 2.4	-97.6 ± 3.1
Point Robot	Generative	2.83	-9.41 ± 0.42	-9.42 ± 0.42
	Randomize	0.54	-5.19 ± 0.05	-6.39 ± 0.05
	Relabeling	0.04	-9.22 ± 0.24	-9.27 ± 0.22
	None	1.46	-5.24 ± 0.27	-6.52 ± 0.08

Our code is available at
<https://github.com/PKU-AI-Edge/CORRO>

Thank you!