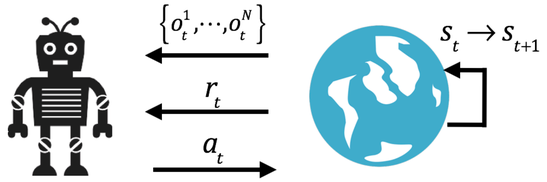
Multi-View Reinforcement Learning

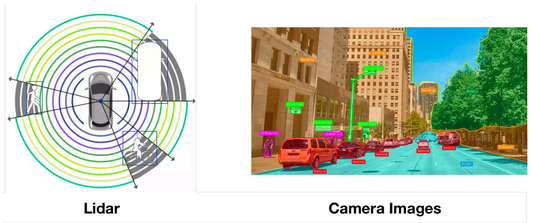
Minne Li*, Lisheng Wu*, Haitham Bou Ammar^, Jun Wang
University College London
*Equal contributions. ^Honorary Lecturer at University College London

Motivation

This work is concerned with multi-view reinforcement learning (MVRL), which allows for decision making when agents share common dynamics but adhere to different observation models.

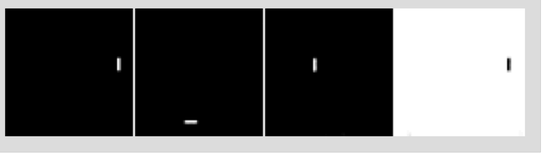


Examples



Lidar

Camera Images



Original Transposed H-swapped Inverse

In Atari Pong, balls and paddles always follow the same physical principles of persistence, continuity, cohesion and solidity, although they are observed through totally different frame of reference.

Multi-view MDP

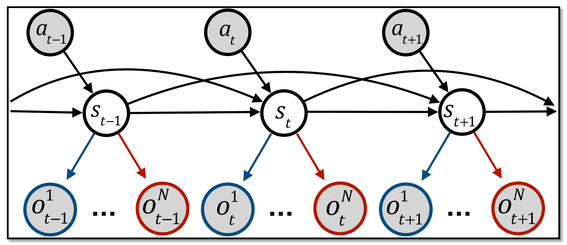
Given $\left\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{O}_1, \mathcal{P}_{obs}^1, \dots, \mathcal{O}_N, \mathcal{P}_{obs}^N \right\rangle$, a history of heterogeneous observations (and potentially actions) $\mathcal{H}_t = \left\{ \boldsymbol{o}_1^{i_1}, \dots, \boldsymbol{o}_1^{i_1} \right\}$, where $\boldsymbol{o}_k^{i_k} \sim \mathcal{P}_{obs}^{i_k}(.|\boldsymbol{s}_k)$ for $k \in \{1, \dots, T\}$ and $i_k \in \{1, \dots, N\}$, the goal is to find optimal action policy $\boldsymbol{\pi}^*$ to maximize the return:

$$\mathcal{G}_T\left(\boldsymbol{\tau}^{\mathcal{M}}\right) = \sum\nolimits_t \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t),$$

where $\boldsymbol{\tau}^{\mathcal{M}} = \left[\boldsymbol{s}_1, \boldsymbol{o}_1^{i_1}, \boldsymbol{a}_1, \dots, \boldsymbol{s}_T, \boldsymbol{o}_T^{i_T}\right]$,

by $\max_{\pi^{\mathcal{M}}} \mathbb{E}_{\tau^{\mathcal{M}}}\left[\mathcal{G}_T\left(\boldsymbol{\tau}^{\mathcal{M}}\right)\right]$, where $\boldsymbol{\tau}^{\mathcal{M}} \sim p_{\pi^{\mathcal{M}}}\left(\boldsymbol{\tau}^{\mathcal{M}}\right)$ and

$$p_{\pi^{\mathcal{M}}}\left(\boldsymbol{\tau}^{\mathcal{M}}\right) = \mathcal{P}_0(\boldsymbol{s}_1)\mathcal{P}_{\text{obs}}^{i_1}\left(\boldsymbol{o}_1^{i_1}|\boldsymbol{s}_1\right)\pi^{\mathcal{M}}\left(\boldsymbol{a}_1|\mathcal{H}_1\right) \prod\nolimits_{t=2}^{T} \mathcal{P}_{\text{obs}}^{i_t}\left(\boldsymbol{o}_t^{i_t}|\boldsymbol{s}_t\right)\mathcal{P}(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})\pi^{\mathcal{M}}\left(\boldsymbol{a}_{t-1}|\mathcal{H}_{t-1}\right)$$



Graphical model of multi-view learning.

Model-free Solutions (MV-MF)

Updating the policy parameters $\boldsymbol{\omega}$ as

$$\boldsymbol{\omega}^{k+1} = \boldsymbol{\omega}^k + \eta^k \frac{1}{M} \sum\nolimits_{j=1}^{M} \sum\nolimits_{t=1}^{T} \nabla_{\boldsymbol{\omega}} \log \pi^{\mathcal{M}}\left(\boldsymbol{a}_t^j | \mathcal{H}_t^j\right)\left(\mathcal{R}(\boldsymbol{s}_t^j, \boldsymbol{a}_t^j) - \mathcal{B}_{\phi}(\mathcal{H}_t^j)\right)$$

where $\mathcal{B}_{\phi}(\mathcal{H}_t)$ is the observation-based baseline.

Problems of MV-MF: sample-inefficient.

Model-based Solutions

By maximizing the **model evidence** of the observation outcome

$$p\left(\boldsymbol{o}_{1:T}^{\mathcal{MC}}\right) = \prod_{i_t=1}^{N} p\left(\boldsymbol{o}^{i_t\mathcal{C}}\right) = \prod_{i_t=1}^{N} \int_{\boldsymbol{s}_1} \cdots \int_{\boldsymbol{s}_T} p(\boldsymbol{o}_1^{i_t}, \boldsymbol{a}_1, \dots, \boldsymbol{o}_{T-1}^{i_t}, \boldsymbol{a}_{T-1}, \boldsymbol{o}_T^{i_t}, \underbrace{\boldsymbol{s}_1, \dots, \boldsymbol{s}_T}_{\text{latent variables}}) \underbrace{d\boldsymbol{s}_1 \dots d\boldsymbol{s}_T}_{\text{marginalization}},$$

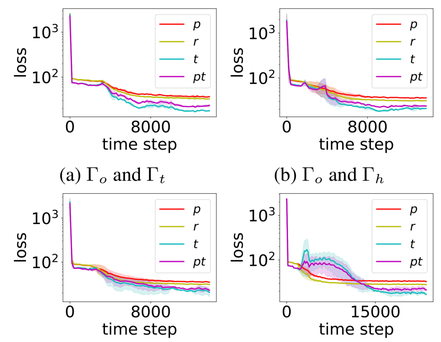
where $\boldsymbol{o}_{1:T}^{\mathcal{MC}}$ collects all multi-view observations and actions across time-steps.

The above can be derived as maximizing the **multi-view ELBO** (see paper for details)

$$\max_{\boldsymbol{\theta}_{\mathrm{m}}, \boldsymbol{\phi}} \sum\nolimits_{i_t=1}^{N} \sum\nolimits_{t=1}^{T} \left[\mathbb{E}_{q_{\boldsymbol{\phi}^{i_t}}(\boldsymbol{s}_t|\mathcal{H}_t)}\left[\log p_{\boldsymbol{\theta}^{i_t}}(\boldsymbol{o}_t^{i_t}|\boldsymbol{s}_t)\right] - \mathrm{KL}\left(q_{\boldsymbol{\phi}^{i_t}}(\boldsymbol{s}_t|\mathcal{H}_t) || p_{\boldsymbol{\theta}^{i_t}}(\boldsymbol{s}_t|\hat{\mathcal{H}}_t)\right) \right]$$

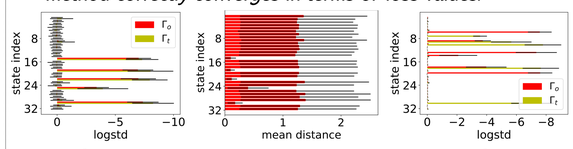
We can then have **model-based MVRL through cross-view policy transfer (MV-PT)**, which only require very few data from a specific view to train the multi-view model. This can then be used for action selection by: 1) inferring the corresponding latent state, and 2) feeding the latent state into the policy learned from another view with greater accessibility.

Modeling Results



(a) Γ_o and Γ_t (b) Γ_o and Γ_h (c) Γ_o and Γ_c (d) Γ_o and Γ_m

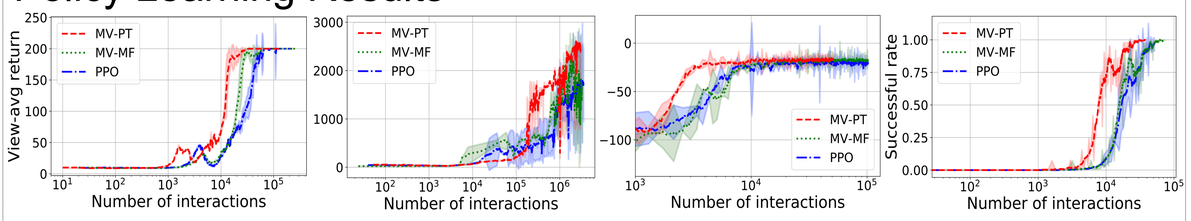
Training multi-view models on Atari Pong with four view variations (t, h, c, m). Legend: p: prediction loss; r: reconstruction loss; t: transformation loss; pt: predicted transformation loss. These results demonstrate that our method correctly converges in terms of loss values.



(a) $\log(\boldsymbol{\sigma}^2)$, multi-view (b) Distance of μ_o, μ_t (c) $\log(\boldsymbol{\sigma}^2)$, WM

Difference between inferred latent states from the original view and the transposed view. Results demonstrating that our method is capable of learning key elements -- a property essential for multi-view dynamics learning. These results also demonstrate that extracting such key-elements is challenging for world-models.

Policy Learning Results



(a) Cartpole Policy Learning (b) Hopper Policy Learning (c) RACECAR Policy Learning (d) Parking Policy Learning

Policy learning results demonstrating that our method outperforms others in terms of sample complexities. We demonstrate these results on 4 different dynamical systems. In all the above, our method can outperform state-of-the-art significantly.

Computer Science Department,
University College London

Contact
Minne Li: minne.li@cs.ucl.ac.uk
Lisheng Wu: lisheng.wu.17@ucl.ac.uk