Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning

Zhaohan Daniel Guo [*1] Bernardo Avila Pires [*1] Bilal Piot [1] Jean-Bastien Grill [2] Florent Altché [2]
Rémi Munos [2] Mohammad Gheshlaghi Azar [1]

Iara Santelices

CS 885

March 9, 2022

Outline

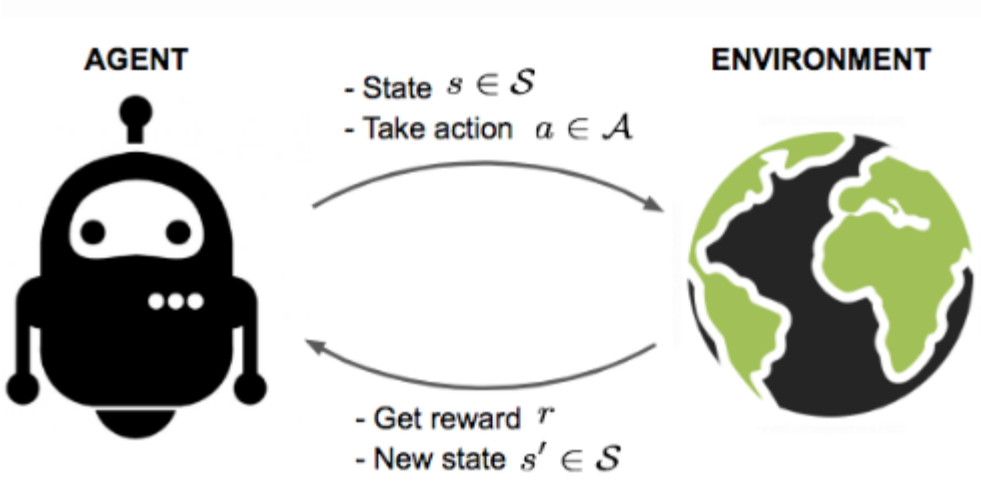
- Introduction
- Background
- Proposed Solution
- Empirical Evaluation
- Conclusion

Introduction

- **Topic:** representation learning for multi-task reinforcement learning(RL) in partially observable environments
- **Problem tackled:** Current approaches to multi-task RL in partially observable environments require high levels of accuracy which is difficult to achieve
- **Solution Proposed:** Predicting future latent observations to improve RL performance

Vocabulary

- **Reinforcement Learning(RL):** Agent takes in observations take actions to maximize its reward
- **Multi-task RL:** RL where the agent must complete many tasks at the same time. Example: autonomous driving
  - Task 1: detect pedestrians
  - Task 2: detect other vehicles
  - Task 3: detect signs
  - etc.



Background: Vocabulary

- **Representation Learning:** when a system learns the underlying features of raw data required for classifiers, predictors or other algorithms. Related: auto-encoder
- **Latent features:** not directly observable features of raw data

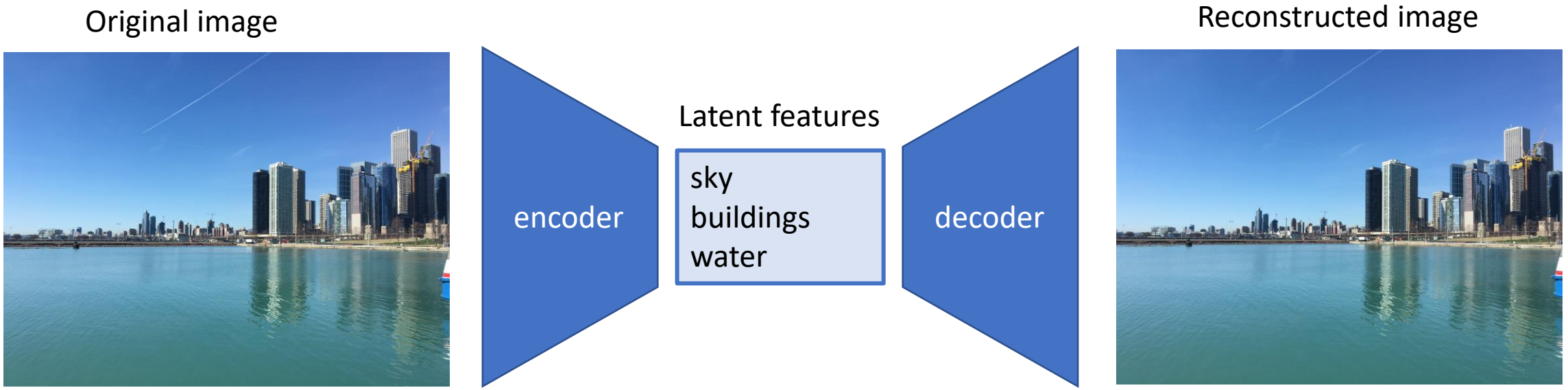


Figure 1. Autoencoder architecture

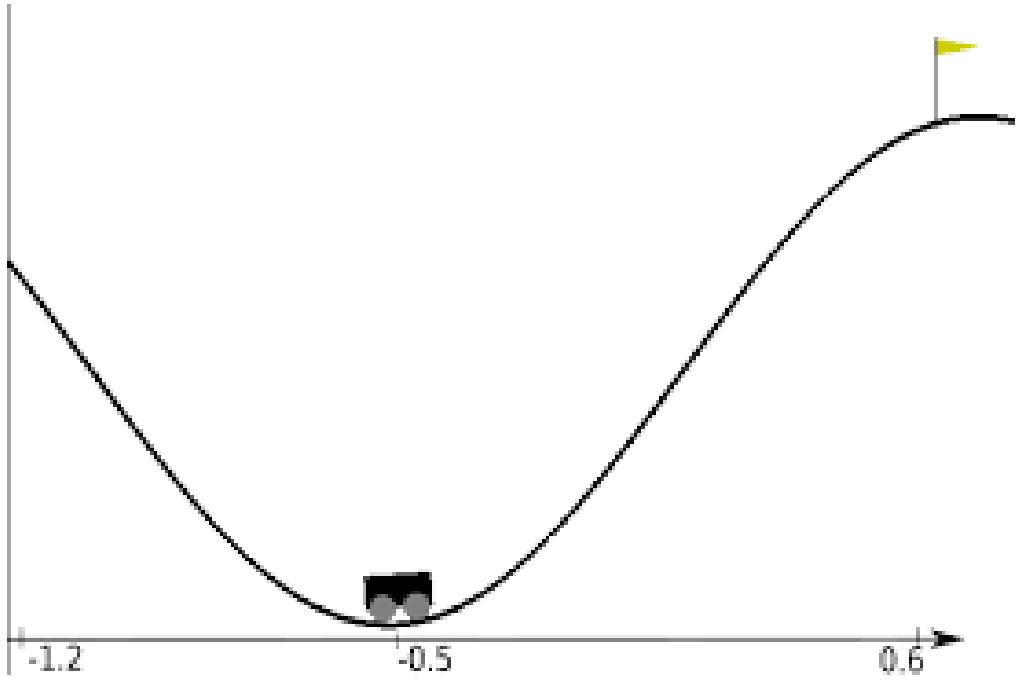
Background: Vocabulary

- **Partially observable environment:** where the agent cannot observe its full state information
- **Bootstrapping:** Updating an estimate with the value of another estimate.

$$Q(s,a) := Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

Background: Motivation

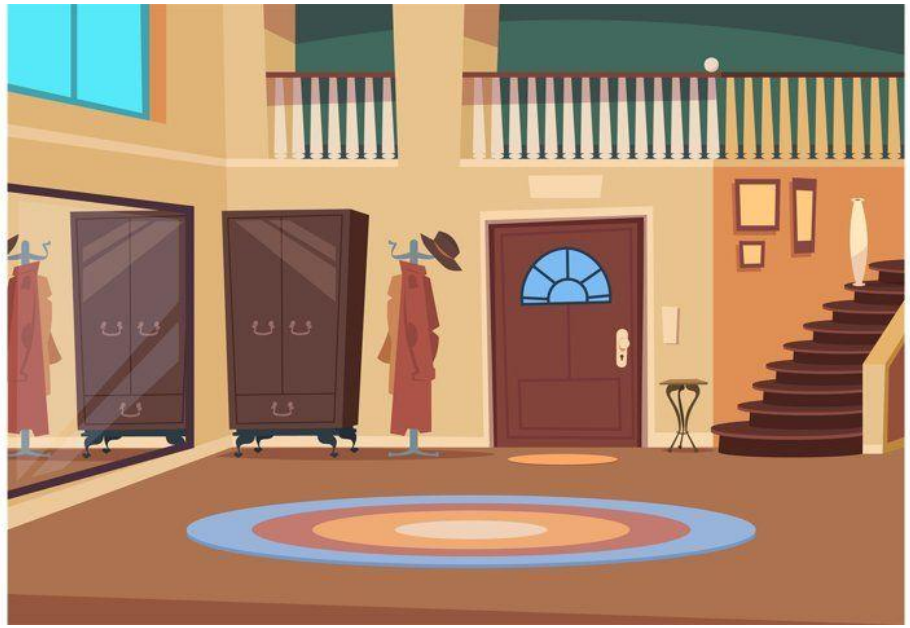
- Deep RL in complex multi-task and partially observable environments is an ongoing area of research.
- Traditional RL: reward is the learning signal
- RL with Representation learning: observations are the learning signal



Mountain car environment

Background: Motivation

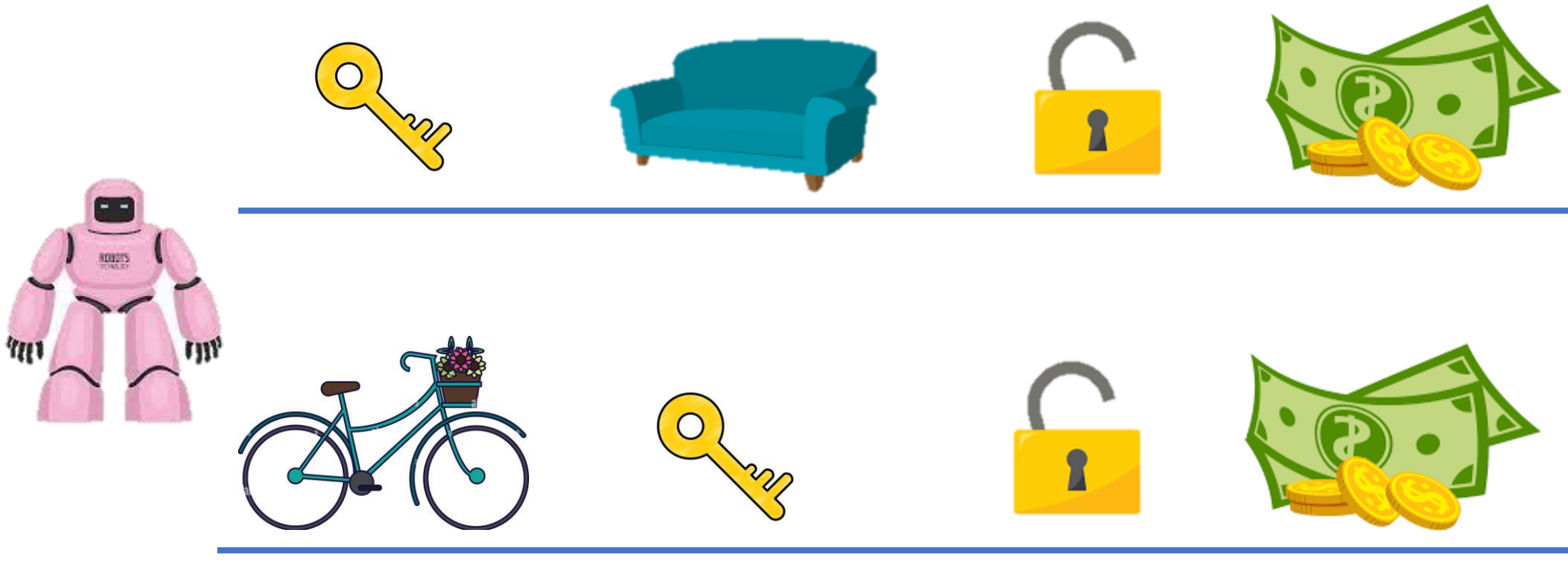
- **Current approach**: Representation learning with RL focuses on predicting future observations
- **Problem**: requires high level of accuracy; difficult in complex environments
- **New Approach**: predict future *latent* observations



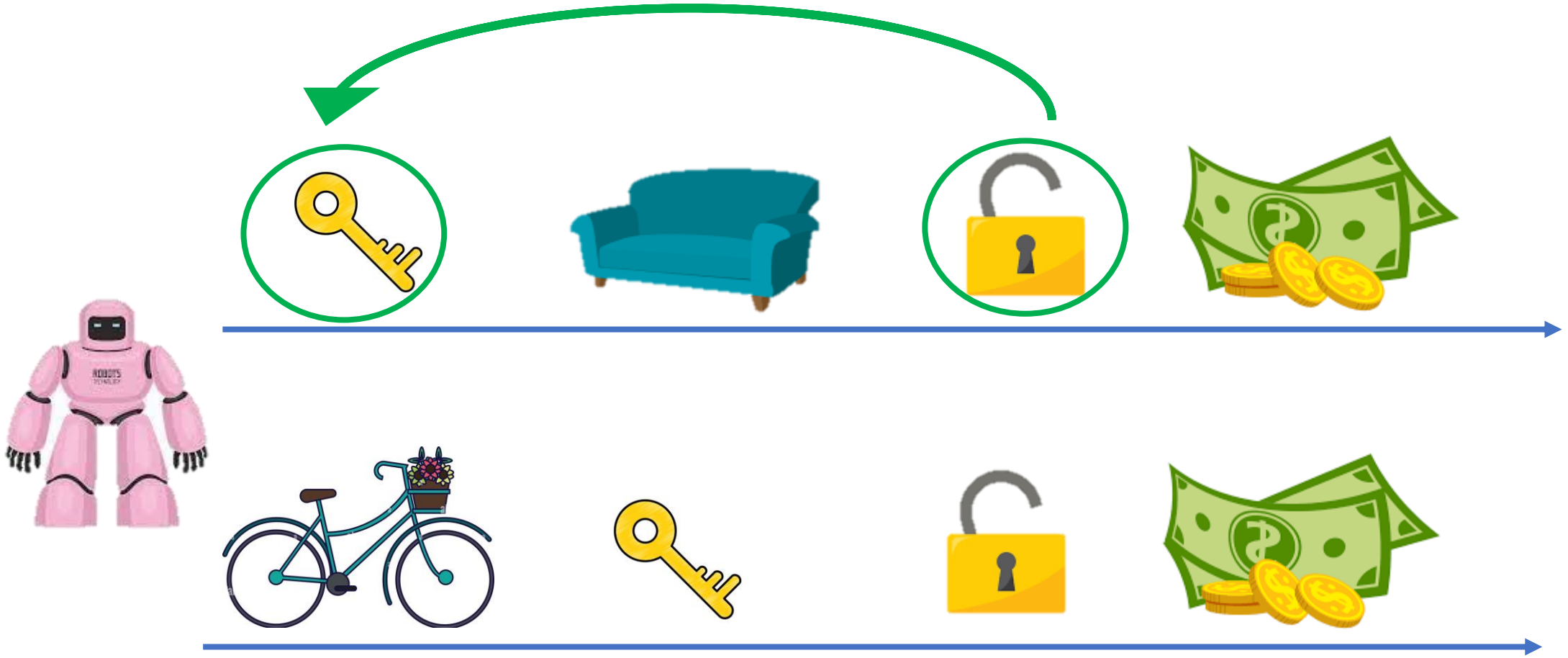
Solution: Prediction of Bootstrapped Latents

PBL (pebble)

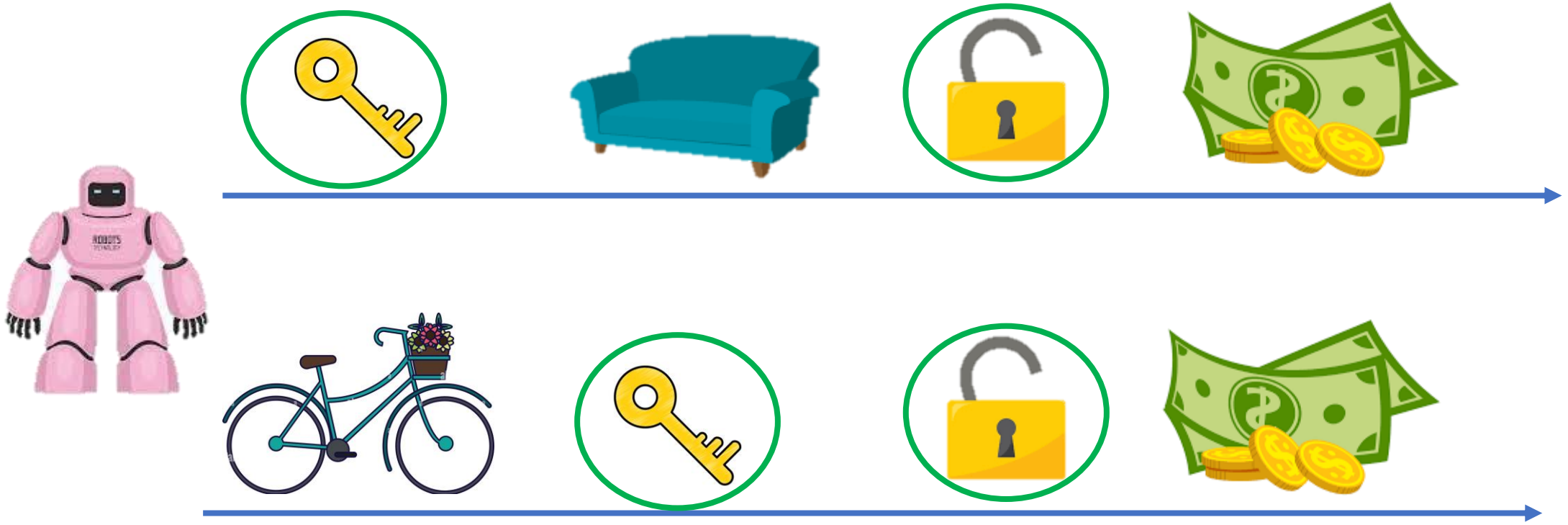
Background: Simple Example



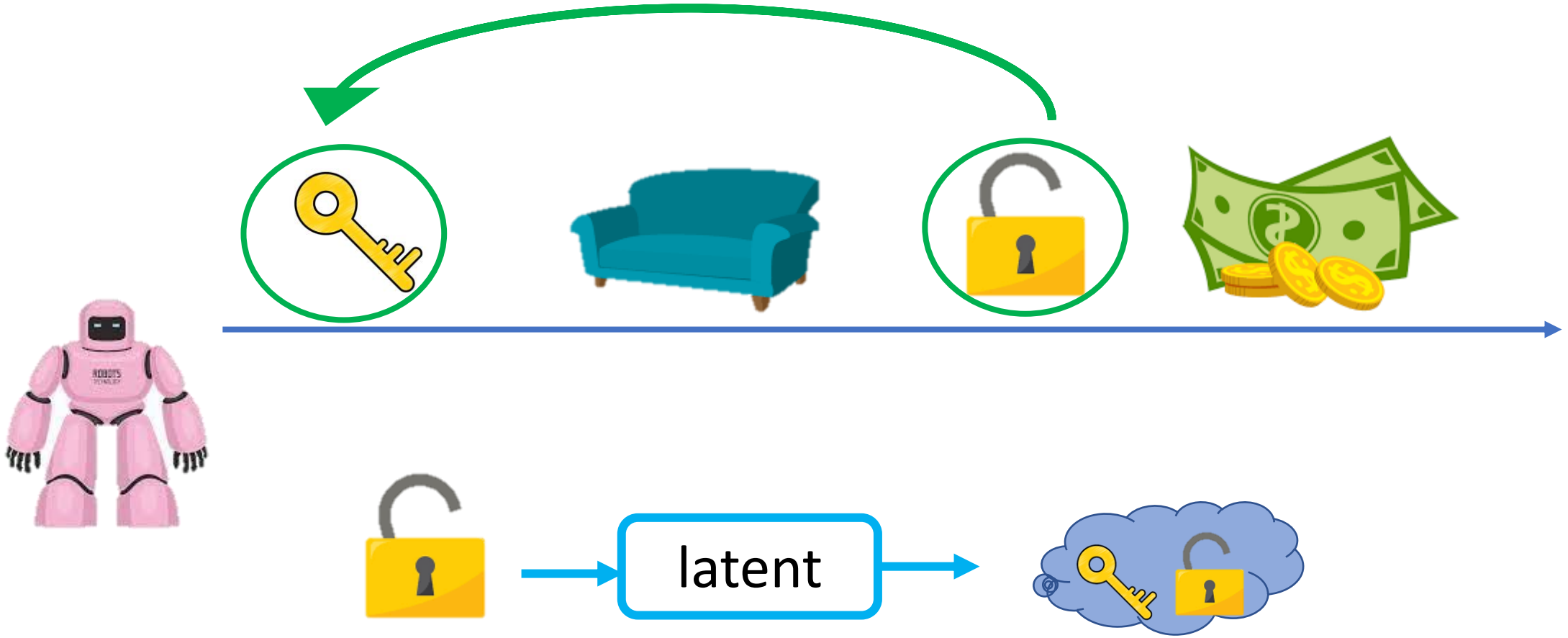
Background: Simple Example



Background: Simple Example

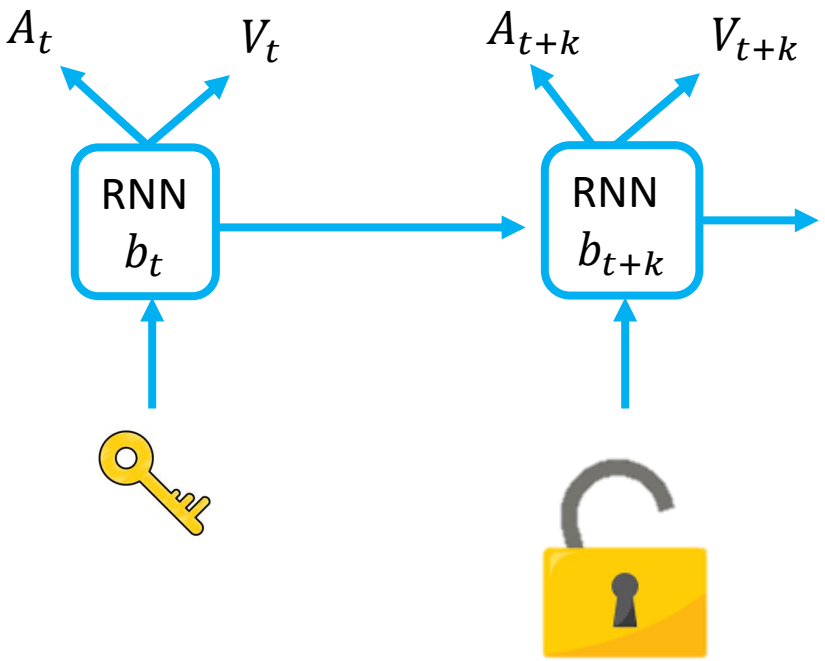


Background: Simple Example



Latent observation:

Content: Predictions of Bootstrapped Latents: PBL



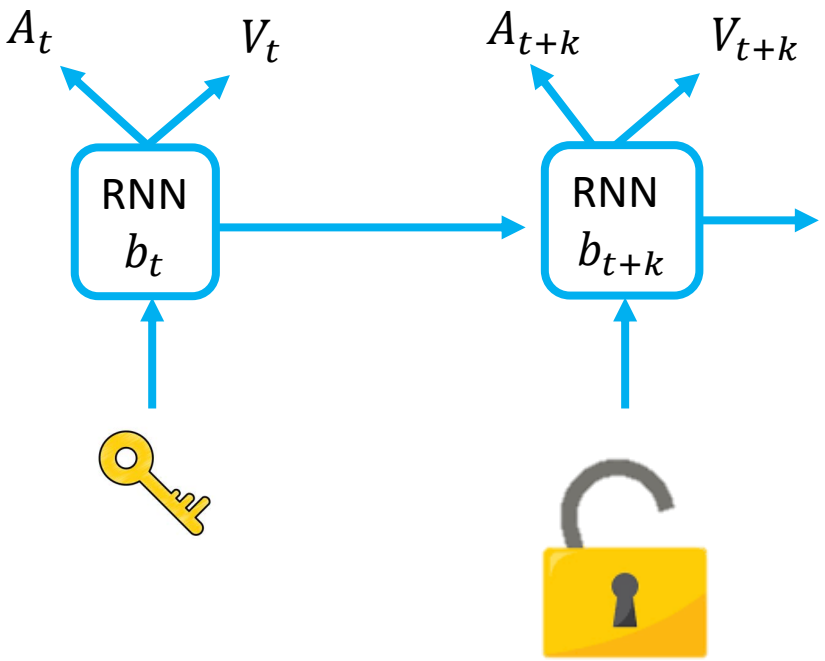
High Level:

Input: observation

Output: value(V_t) and action (A_t)

Content: Predictions of Bootstrapped Latents: PBL in more detail

B_t: agent state/compressed full history
$B_{t,k}$: compressed partial history



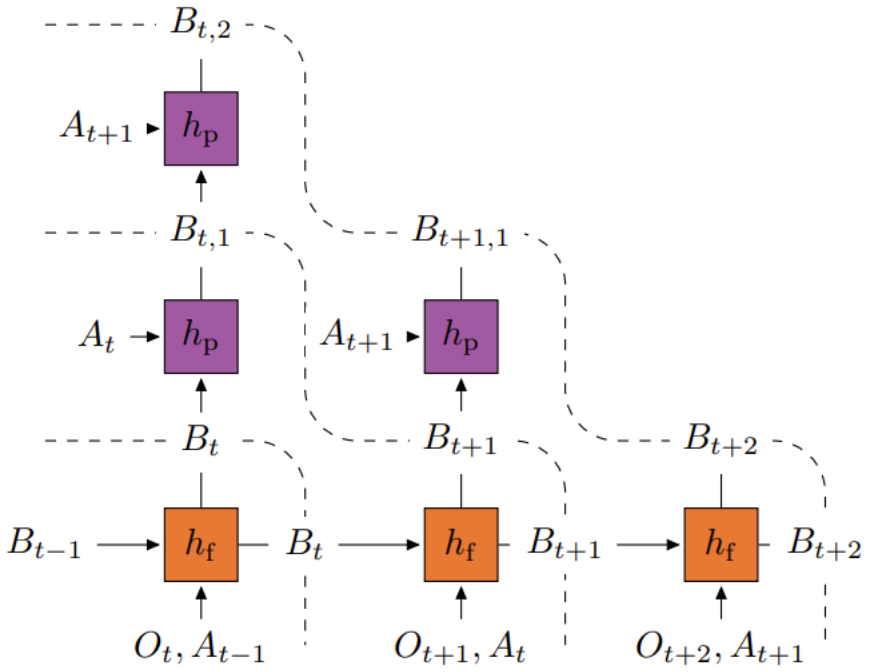
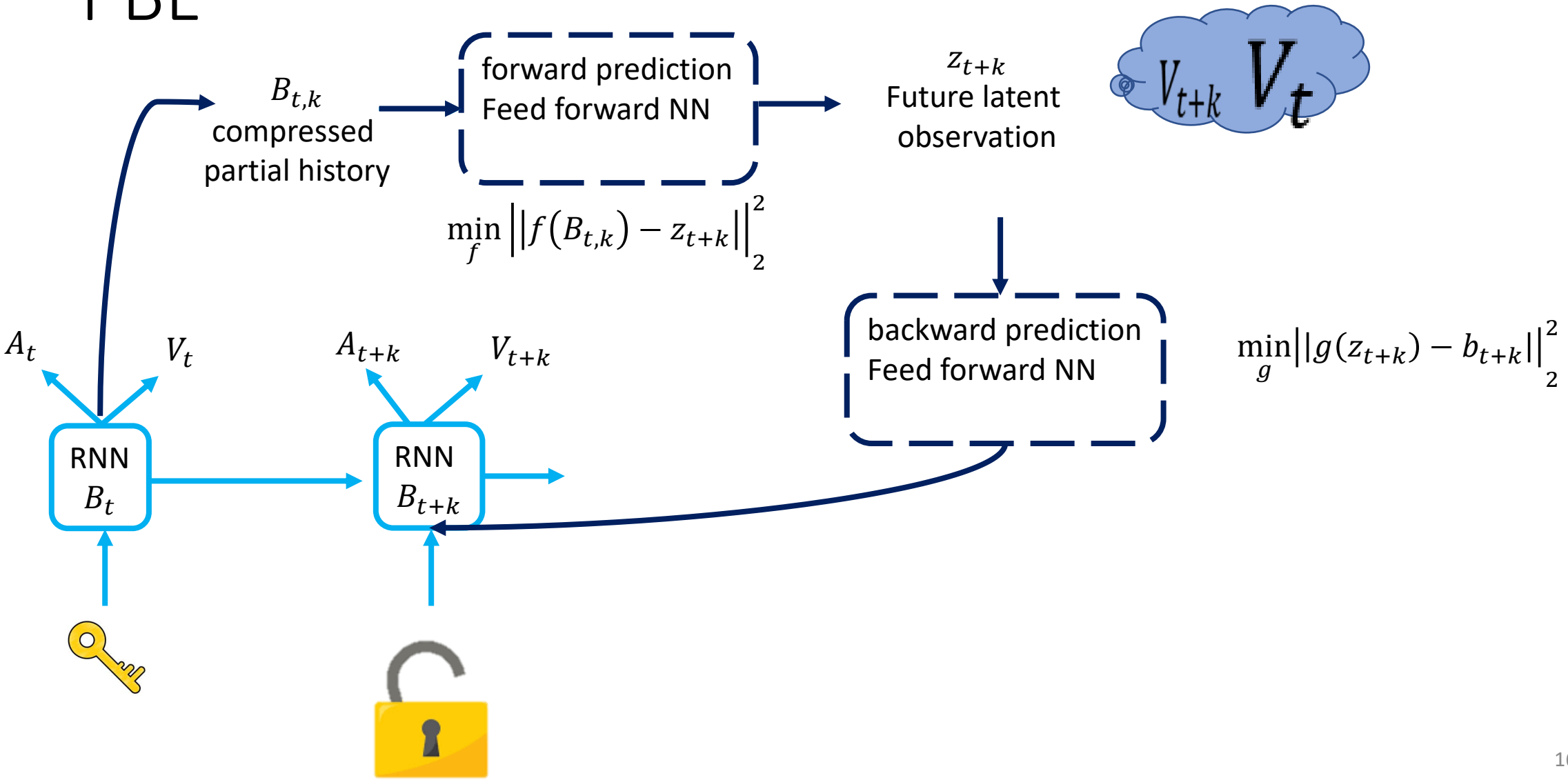
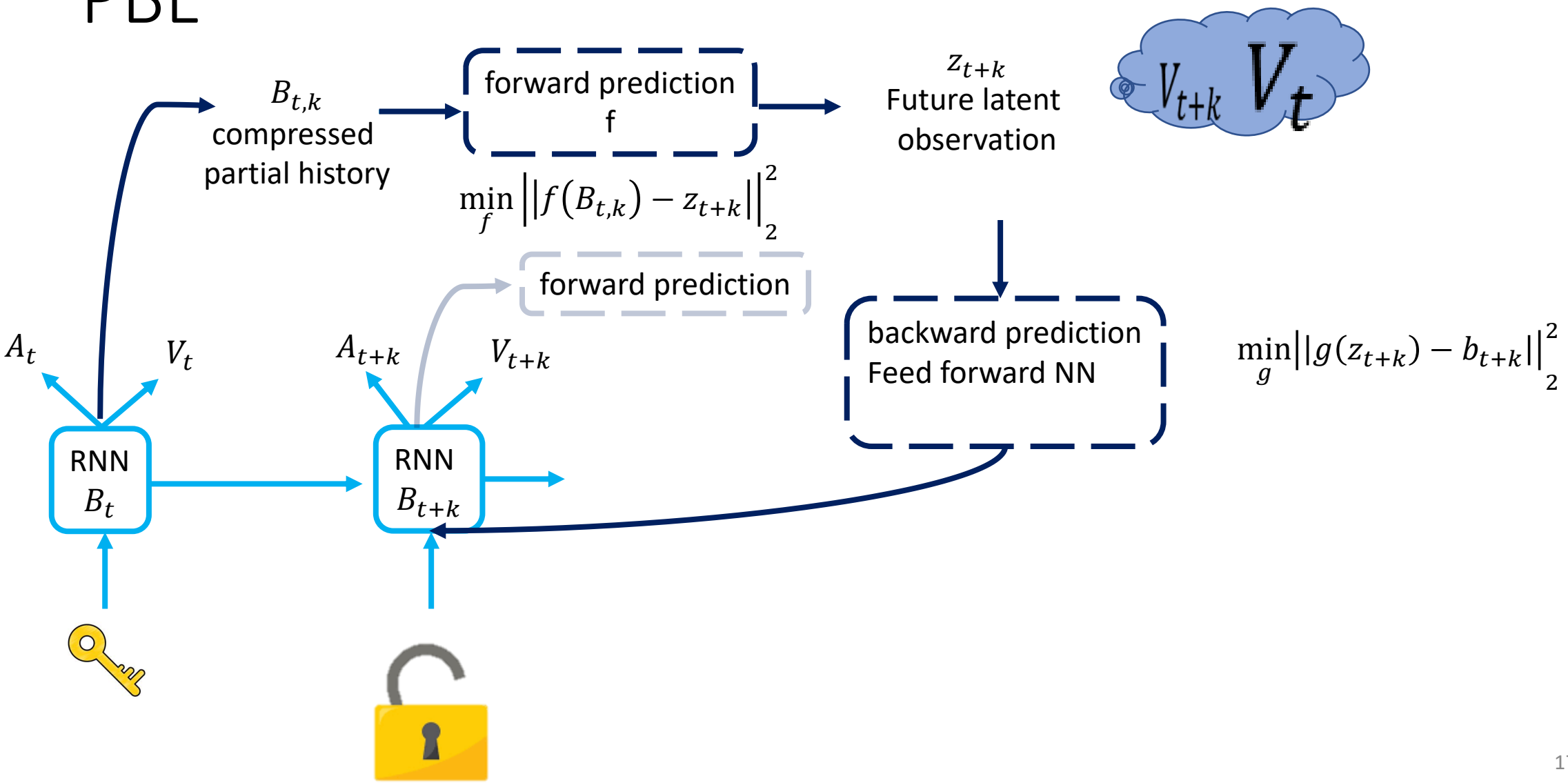


Figure 1. Recurrent architecture for compressing partial histories. Networks used for processing observations and actions have been omitted, and dashed lines connect histories and partial histories

Content: Predictions of Bootstrapped Latents: PBL



Content: Predictions of Bootstrapped Latents: PBL



Content: Training PseudoCode

Bootstrap Latent-Predictive Representations

Algorithm 1 Training Step Pseudocode for PBL

Require: Minibatch of trajectories $B = \{O_t^{(i)}, A_t^{(i)}, R_t^{(i)}\}$, RNN h_p, RNN h_f, MLPs g, g', f, future prediction horizon k, RLLoss (reinforcement learning loss)

Encode observation $Z_t^{(i)} \doteq f(O_t^{(i)})$

Let $B_0^{(i)} \doteq \mathbf{0}$ and $B_t^{(i)} \doteq h_\mathrm{f}(B_{t-1}^{(i)}, O_t^{(i)}, A_{t-1}^{(i)}) \doteq B_{t,0}^{(i)}$ and $B_{t,k}^{(i)} \doteq h_\mathrm{p}(B_{t,k-1}^{(i)}, A_{t+k-1}^{(i)})$

$\mathrm{Forward}(B_t^{(i)}) \doteq \frac{1}{k}\sum_{j=1}^{k} \|g(B_{t,j}^{(i)}) - \mathrm{StopGradient}(Z_{t+j}^{(i)})\|_2^2$

$\mathrm{Reverse}(Z_t^{(i)}) \doteq \|g'(Z_t^{(i)}) - \mathrm{StopGradient}(B_t^{(i)})\|_2^2$

Take gradient step of $\min \frac{1}{|B|}\sum_{i,t}\Big(\mathrm{Forward}(B_t^{(i)}) + \mathrm{Reverse}(Z_t^{(i)}) + \mathrm{RLLoss}(B_t^{(i)}, R_t^{(i)})\Big)$

Because gradients are stopped on the target, algorithm does not collapse to trivial solution

Content: Related Work

- Deep MDP + CRAR [2] [3]:
  - Algorithm learns transition model in latent space
  - Depends on a reward function(Deep MDP)
  - Depends on entropy maximization (CRAR)
- Grill et al 2020 [4]:
  - Self-supervised image representation learning

Content: Related Work

- Pixel Control [5]:
  - Q-learning
  - Current state-of-the-art for DMLab 30
- Simcore DRAW [6]:
  - VAE based representation learning for single-task RL
- Contrastive Predictive Coding (CPC) [7]:
  - Predict future latent representations using auto-regressive models

Content: Advantages and Disadvantages

- Advantages:
  - Z_t is a latent embedding that can combine different observation modalities: images and text for example
  - PBL can encode dynamical and structural dependencies between tasks
- Disadvantage:
  - Complicated architecture that requires two additional networks
    - Often difficult to understand what the algorithm is learning
  - Predicting every future latent from one time step ahead to the horizon is computationally expensive
    - Subsampling can improve computation time with a minimal loss to performance

Empirical Evaluation

Empirical Evaluation

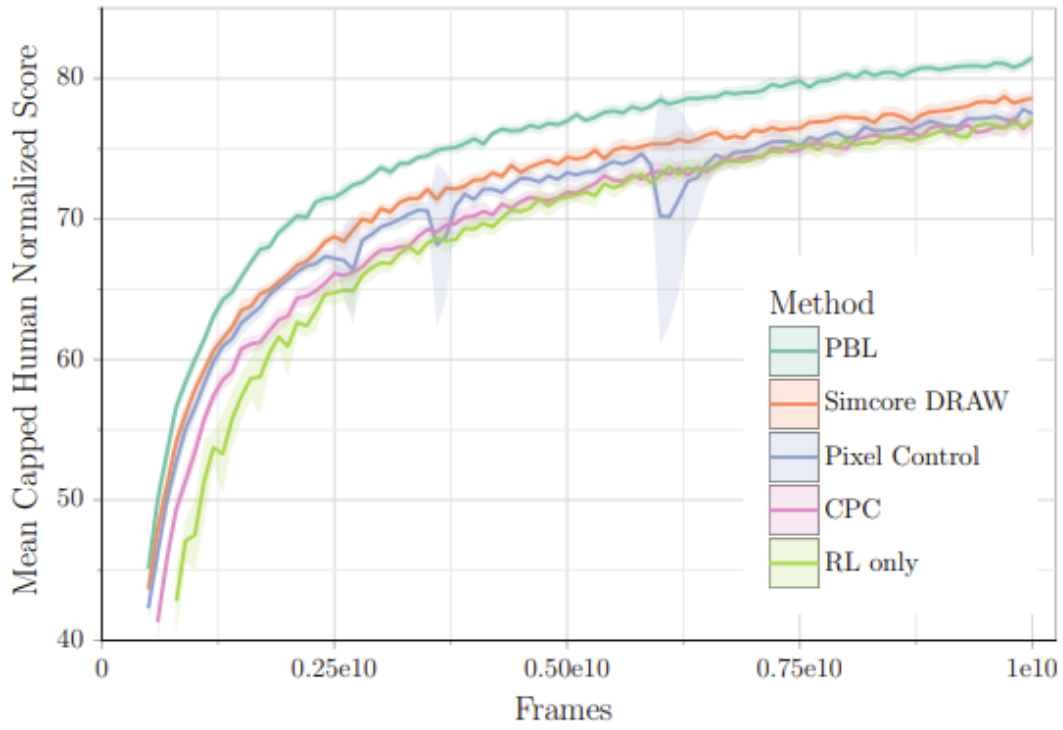
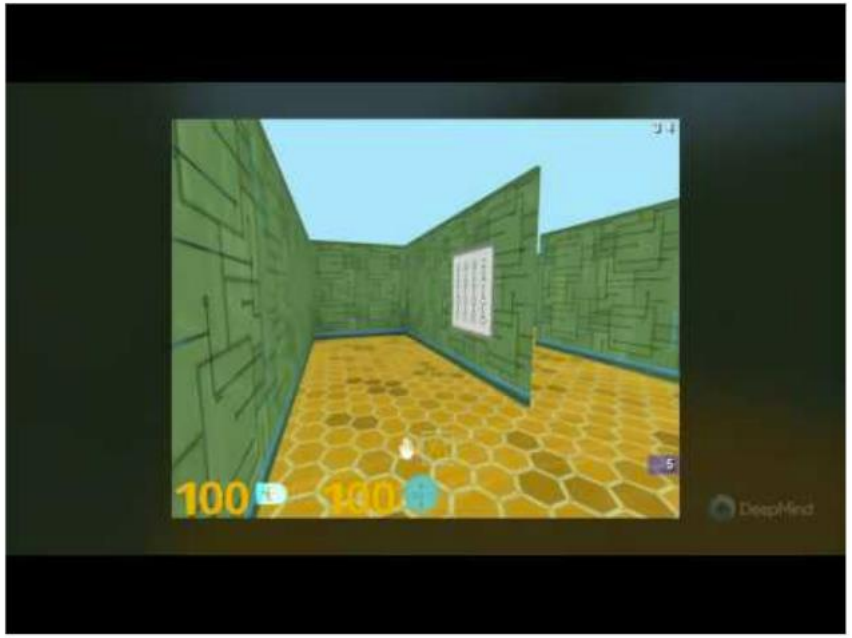
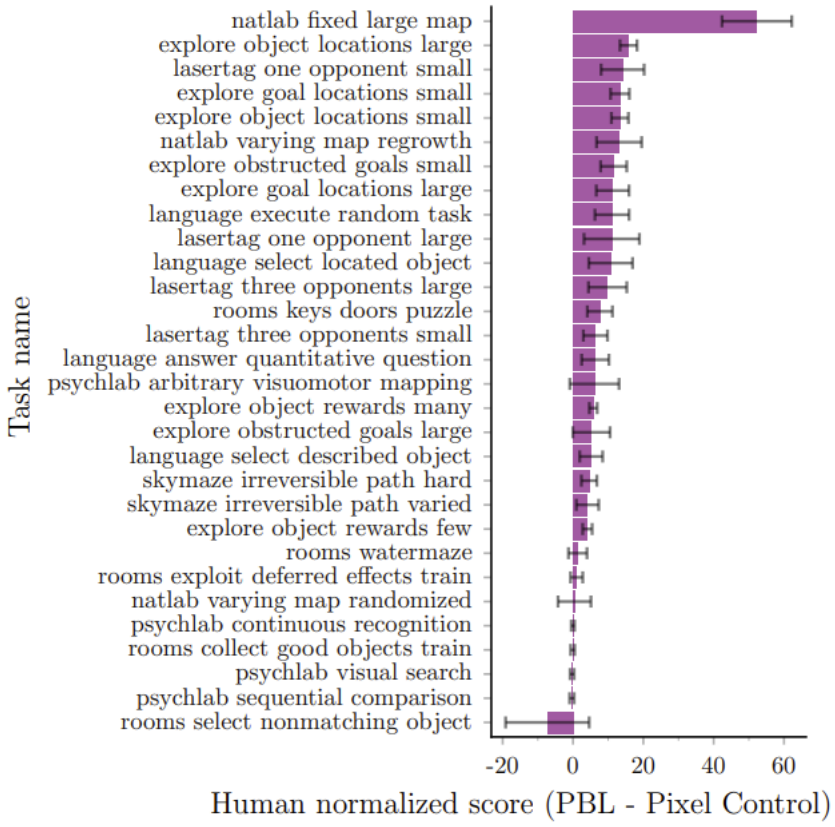


Figure 3. Mean capped human normalized score for compared methods.



Results of PBL and other representation learning methods in DMLabs-30 environments

Empirical Evaluation



Comparing PBL to pixel control for individual tasks in DML-30

Empirical Evaluation

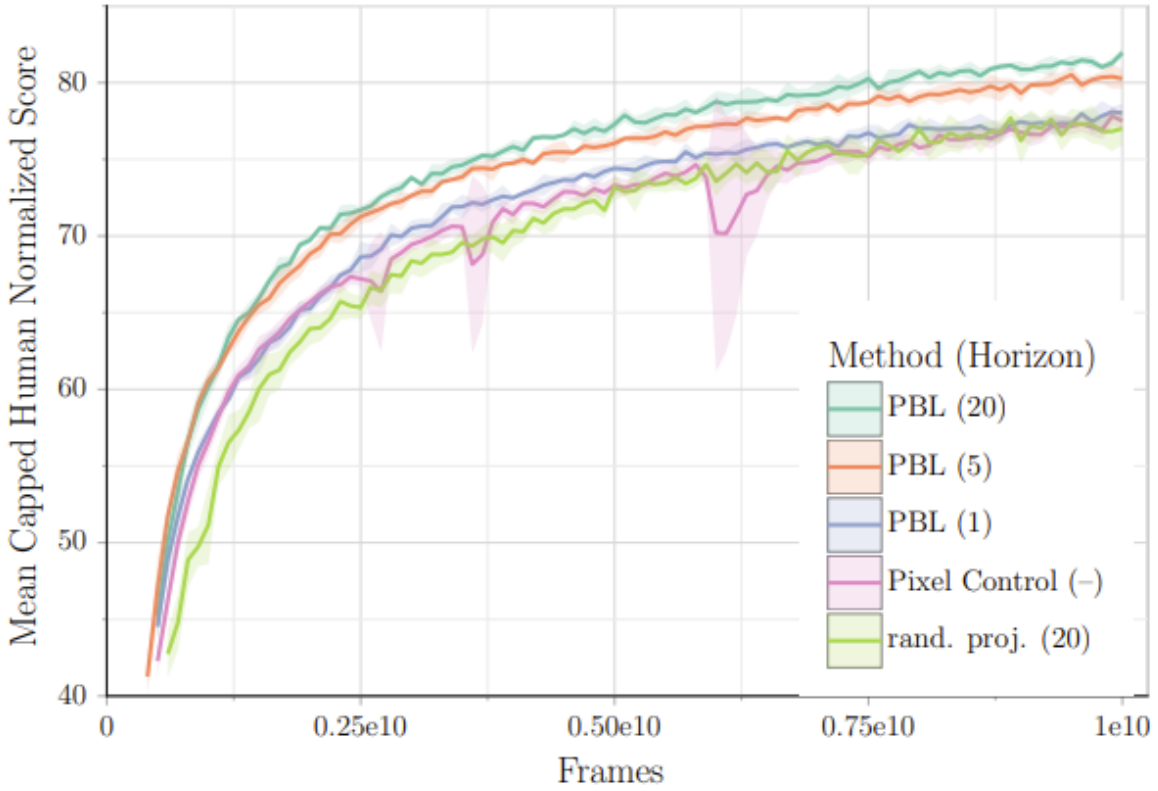


Figure 5. PBL Performance Across Forward Prediction Horizon, compared to pixel control and random projection.

Empirical Evaluation

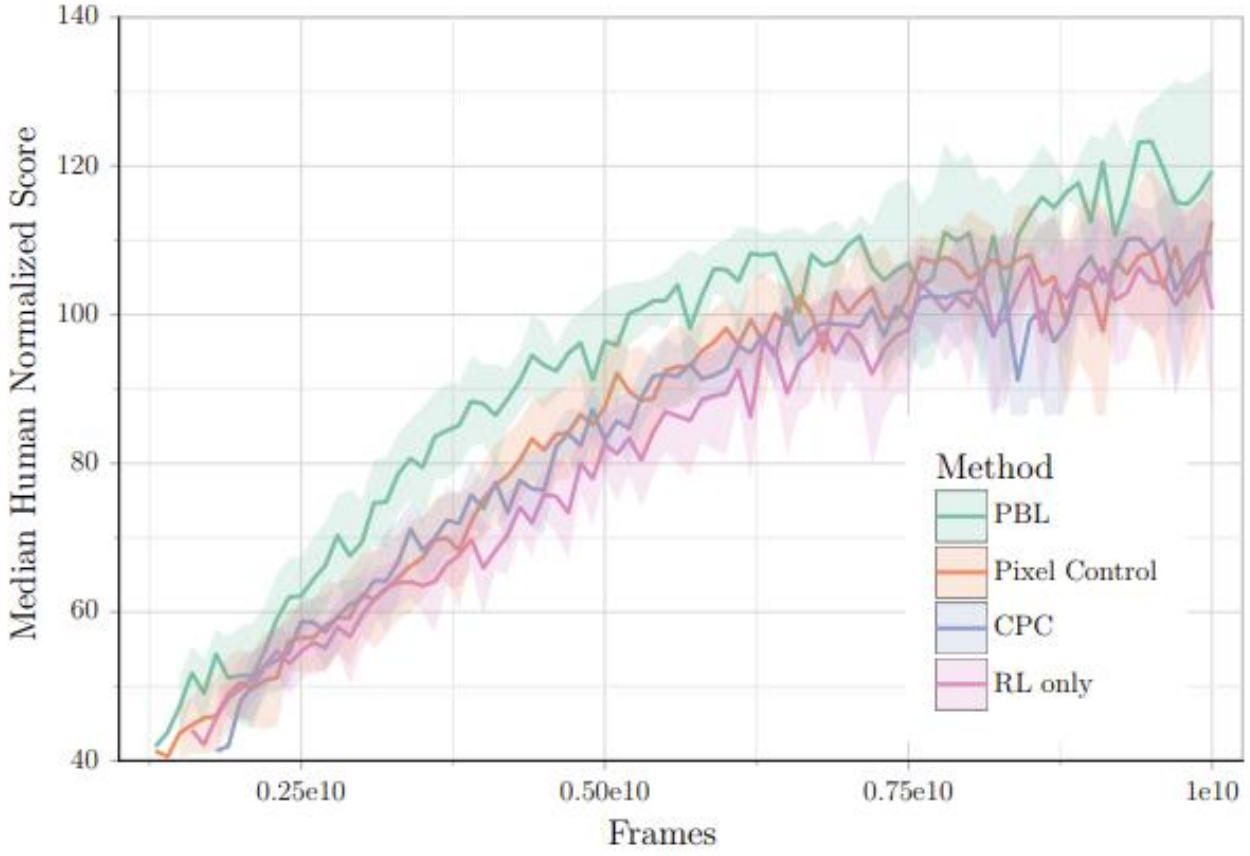


Figure 7. Median human normalized score for compared methods on Atari57.

Empirical Evaluation: Glass Box

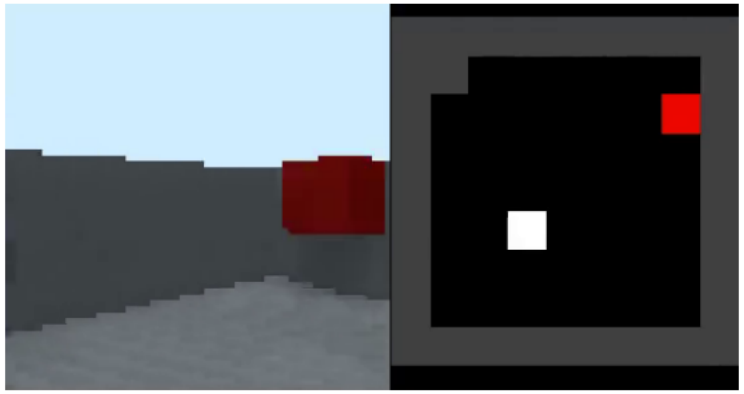


Figure 8. 3D room example: Agent's first-person view (left) and top-down grid-view indicating the object position (right).

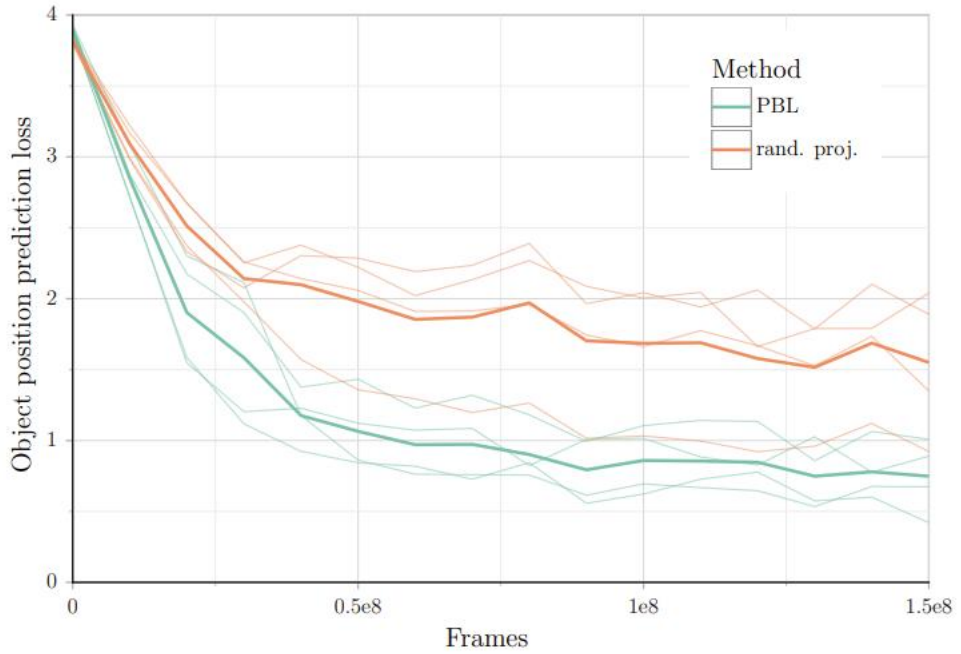


Figure 9. Object position prediction loss for random projection vs PBL. Light lines denote different independent runs.

Conclusion

- **Contribution:** Present a novel method for training latent embeddings for representation learning.
- **Take home message:** by learning meaningful future latent observations RL agents can improve performance

Conclusion: Future Work

- Authors Present:
  - Evaluating pebble in other machine learning domains
  - Transfer learning

- Other Ideas:
  - Implementing PBL with real-world partially observable and multi-task environments (autonomous driving)
  - Evaluating pebble's raw performance (not human normalized score)

References

[1] Z. D. Guo *et al.*, “Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning,” 2020.

[2] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare, “DeepMDP: Learning Continuous Latent Space Models for Representation Learning,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 3802–3826, Jun. 2019, doi: 10.48550/arxiv.1906.02736.

[3] V. François-Lavet, Y. Bengio, D. Precup, and J. Pineau, “Combined Reinforcement Learning via Abstract Representations,” *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 3582–3589, Sep. 2018, doi: 10.48550/arxiv.1809.04506.

[4] J. B. Grill *et al.*, “Bootstrap your own latent: A new approach to self-supervised Learning,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, Jun. 2020, doi: 10.48550/arxiv.2006.07733.

[5] M. Jaderberg, V. Mnih, T. Czarnecki, Wojciech Marian Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, “Reinforcement Learning with Unsupervised Auxiliary Tasks | OpenReview,” 2017, Accessed: Mar. 04, 2022. [Online]. Available: https://openreview.net/forum?id=SJ6yPD5xg.

[6] K. Gregor, D. J. Rezende, F. Besse, Y. Wu, H. Merzic, and A. van den Oord, “Shaping Belief States with Generative Environment Models for RL,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[7] A. van den Oord DeepMind, Y. Li DeepMind, and O. Vinyals DeepMind, “Representation Learning with Contrastive Predictive Coding,” Jul. 2018, doi: 10.48550/arxiv.1807.03748.