Pre-training with Augmentations for Efficient Transfer in Model-Based Reinforcement Learning

Bernardo Esteves[1,2](✉), Miguel Vasco[1,2], and Francisco S. Melo[1,2]

[1] INESC-ID, Lisbon, Portugal
[2] Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
bernardo.esteves@tecnico.ulisboa.pt

Abstract. This work explores pre-training as a strategy to allow reinforcement learning (RL) algorithms to efficiently adapt to new (albeit similar) tasks. We argue for introducing variability during the pre-training phase, in the form of *augmentations* to the observations of the agent, to improve the sample efficiency of the fine-tuning stage. We categorize such variability in the form of perceptual, dynamic and semantic augmentations, which can be easily employed in standard pre-training methods. We perform extensive evaluations of our proposed augmentation scheme in model-based algorithms, across multiple scenarios of increasing complexity. The results consistently show that our augmentation scheme significantly improves the efficiency of the fine-tuning to novel tasks, outperforming other state-of-the-art pre-training approaches.

Keywords: Reinforcement learning · Transfer learning · Representation learning

1 Introduction

Reinforcement learning (RL) approaches have been successfully applied to complex scenarios like games [18,23], robotics [16] and control [17]. In spite of these sounding success stories, RL methods are known for being “data-hungry”: they require millions of interaction steps between the learning agent and the environment, which makes the deployment of RL-based systems extremely expensive and difficult in real-world scenarios, where such intense levels of interaction are prohibitive. As an example, Rainbow [12] required over 34, 000 GPU hours (over 1, 400 days) to train, not considering hyper-parameter tuning [20]. Additionally, a RL system trained for a particular task often fails to generalize to other, similar tasks [11]. Such behavior stands in contrast to the human learning process:

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with ref. UIDB/50021/2020 and the project RELEvaNT, ref. PTDC/CCI-COM/5060/2021. MV acknowledges the FCT PhD grant SFRH/BD/139362/2018.

N. Moniz et al. (Eds.): EPIA 2023, LNAI 14115, pp. 133–145, 2023.
https://doi.org/10.1007/978-3-031-49008-8_11

humans efficiently reuse knowledge of similar tasks (such as motion primitives and environmental physics) to efficiently learn to perform novel tasks [13]. This stark difference motivates the need for knowledge transfer approaches that may help to address the sample-efficiency of RL algorithms.

According to Laskin et al. [15], two families of approaches have been proposed in literature to address sample complexity of RL methods: (i) introducing auxiliary tasks, usually relying on data augmentation approaches, that seek to build general-purpose representations for the perceptual observations of the agent that facilitate the learning of control policies [15,21]; and (ii) learning environment models that allow the generation of artificial samples that can be used for learning, thus requiring less interactions with the actual environment [9,22]. This paper builds on the benefits of these two lines of research and addresses the question: "*how does pre-training using different augmentations impact the data efficiency of fine-tuning model-based RL in novel downstream tasks*?"

We focus on the problem of pre-training model-based RL agents and contribute with an in-depth categorization of *transferable features* across similar tasks. In particular, we discuss transfer between tasks that share *perceptual*, *dynamic* and *semantic* features. Driven by our discussion, we contribute a novel pre-training scheme for model-based RL that exploits such transferable features, which we name *Multiple-Augmented Pre-training Scheme* (MAPS). During the pre-training phase, MAPS introduces multiple variations on the observations of the agents, obtained from the current task or similar tasks, forcing the learning of more general-purpose representations and thus improving the efficiency of a subsequent fine-tuning phase in novel downstream tasks. The introduction of such variability in data has already been explored in contexts such as computer vision [5,8] and natural language processing [4,7].

We evaluate MAPS against different pre-training approaches in scenarios of increasing complexity, considering a state-of-the-art model-based RL framework (namely, DreamerV2 [9]). We perform an ablation study on the Mini-Grid environment that highlights how changes in the perceptual and dynamical conditions affect the transfer of information in model-based RL to similar tasks. Furthermore, in a more complex Mini-Grid scenario, we highlight the role of further introducing semantic variability during the pre-training phase, showing that MAPS outperforms other standard pre-training schemes. Finally, in an Atari environment, we highlight the scalability of MAPS to more complex scenarios, and show how pre-training with MAPS significantly improves the fine-tuning performance. In summary, the contributions of this work are threefold.

- We contribute a categorization of *transferable features* for the pre-training of model-based RL agents;
- We introduce Multiple-Augmented Pre-training Scheme (MAPS) that exploits such features to introduce variability during the pre-training phase;
- We evaluate MAPS against different pre-training approaches in scenarios of increasing complexity, showing how our approach allows agents to efficiently fine-tune to novel downstream tasks.

2 Related Work

Transferring knowledge to new tasks is often related to the field of Transfer Learning, which seek to bring learning improvements by relaxing assumption that the data used for learning between old and new tasks must be independent and identically distributed [25]. *Pre-training* is considered the predominant approach to perform experience transfer: we train a model on an initial task, also known as *pre-training task*, and then adapt the model on a new *downstream task*, by using the previously learned weights, via *fine-tuning* [3]. Pre-training has been successfully applied to a variety of fields beyond RL. For example, in computer vision, self-supervised representation learning approaches have seen significant developments, both in contrastive [5,10] and predictive methods [8]. In this work, inspired by these pre-training approaches in computer vision, we explore self-supervised augmentations in the field of model-based RL.

During the pre-training phase, it is common to train the model on large amounts of general data, and is common to use other learning objectives that are only used for pre-training. For example, in SimCLR [5], the authors present a new contrastive method to pre-train a large model with a large unlabeled dataset with 1.2 million images, that can then be fine-tuned with a small labeled dataset. However, in our work we do not have access to a huge dataset with millions of highly diverse trajectories and millions of diverse games easily available, thus we try to focus the pre-training on a small set of more similar tasks to attempt to extract information from these to the desired downstream task. In RL settings, both CURL [15] and ATC [24] propose contrastive auxiliary objectives for learning general representations of the agent's environments. However, they consider only model-free agents and employ only perceptual augmentations. In this work, we consider how perceptual, dynamical and semantic augmentations improve the transfer of model-based RL agents. In SGI [21] the authors propose to employ multiple auxiliary tasks to pre-train and the fine-tune an agent on the same task, and shown negative results on transferring representations between Atari games on a small data regime. Contrary to our work, they focus only on model-free methods, and only use random crops and intense jittering (both perceptual augmentations). In RAD [14] the authors explore ten different types of data augmentations, and show how using augmentations while learning the same task it helps improve the data-efficiency and generalization of RL methods. Compared with our work, RAD uses only perceptual augmentations and focus on model-free single task learning.

3 Method

In this work, we address the problem of adapting RL agents to novel downstream tasks. In particular, we consider a two-stage transfer approach: we initially *pre-train* agents on a given task T_p and subsequently transfer the agents to a novel downstream task T_d, where we *fine-tune* the agents to the novel task.

One of the challenges of the transfer process resides in the difference between the information provided to the agent in T_p and in T_d. During the pre-training

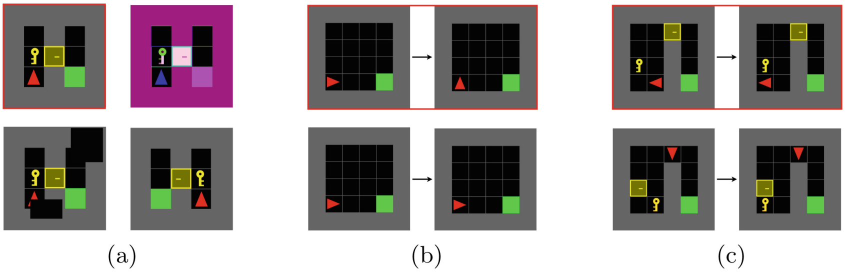


Fig. 1. Our proposed augmentation scheme for efficient adaptation: **a** *perceptual* augmentations exploit global transformations of the original observations (in red), such as color inversion, cropping and flipping; **b** *dynamical* augmentations exploit counterfactual transformations of original transitions in the environment (in red), such as randomly introducing "NoOp" actions; **c** *semantic* augmentations exploit conceptual transformations over the original observations (in red), such as changing the sprites of the player and objects

phase, the agent experiences a set of observations $O_p \in \mathcal{O}$, with $\mathcal{O}$ the set of all possible observations in the space of all possible tasks. From such observations, and auxiliary reward signals provided by the environment, the agent learns to perform the pre-training task T_p. However, during the adaptation phase, the agent reuses its experience to learn the downstream task T_d from a set of observations $O_d \in \mathcal{O}$, potentially disjoint from O_p, along with a new reward signal.

However, in many tasks there are intrinsic similarities that, if exploited, could improve the transfer procedure. For example, despite the differences in the observations in each scenario, the games "Space Invaders" and "Pepsi Invaders", depicted in Fig. 3c, d respectively, share some features between them; both share similar core semantics and dynamics of a shooting up game.

To exploit the potential intrinsic similarities between the pre-training and downstream tasks, we propose to introduce *augmentations* during the pre-training phase: we expand the set of pretraining observations $O_p^\star \supseteq O_p \in \mathcal{O}$ through augmentation functions $A(o)$ to allow the efficient adaptation to downstream tasks. In Sect. 3.1, we propose a categorization of augmentation functions to exploit *perceptual*, *dynamical* and *semantic* similarities between T_p and T_d. Additionally, in Sect. 3.2 we show how our augmentations can be easily introduced into standard pretraining schemes, with minimal computational overhead, an approach we denote by *Multiple Augmented Pre-training Scheme* (MAPS).

3.1 Augmentation Scheme

We now focus our attention on the nature of the augmentation functions $A(o)$ to improve the efficiency of the fine-tuning process on unknown, novel tasks T_d. As shown in Fig. 1, we propose three different categories of augmentations: *perceptual*, *dynamical* and *semantic*.

Perceptual Augmentations One of the significant ways observations can change from the T_p to T_d concerns features of the perceptions of the agents, such as color, orientation and size. We propose to expand the set of observations O_p to introduce such variability by considering *perceptual* augmentations.

As shown in Fig. 1a, perceptual augmentations correspond to global transformations on the observations of the agents. These augmentations introduce variability in general features of the observations, having no impact on the underlying task and dynamics of the environment. Moreover, perceptual augmentations are agnostic to the semantics in the perception itself (such as the players and enemies). Examples of such augmentations include color inversion for the whole observation, or random cropping and mirroring across different axes.

The use of augmentations has been explored by several self-supervision methods, such as SimClr [5] and CURL [15], that learn transferable representations by employing visual-based augmentations on image data. In this work, we introduce two more categories of augmentations to the observations of the agents.

Dynamical Augmentations Another potential change in the sequence of observations experienced by the agents from the T_p to T_d concerns the dynamics of the environment, i.e., how the environment changes as a function of the actions of the agent. We propose to expand the standard set of observations O_p in order to introduce such variability by considering *dynamical* augmentations.

As shown in Fig. 1b, dynamical augmentations correspond to changes on the observations of the agent, due to transformations on its actions. Contrary to perceptual augmentations, dynamical augmentations can only be perceived across multiple time-steps, having no impact in the general features nor in the semantics of the observation. Examples of such augmentations are operations of randomly employing "NoOp" actions or swapping the actions of the agent.

Semantic Augmentations Finally, observations from T_p and T_d can also differ regarding local, higher-level features of the observations, such as the sprites of the agent and the enemies. We propose to expand the set of observations O_p in order to introduce such variability by considering *semantic* augmentations.

Semantic augmentations correspond to local transformations on the observations of the agent. Much like dynamical augmentations, semantic augmentations often can only be perceived across multiple time-steps (see Fig. 1c) through specific visual modifications to game elements such as the player, or surrounding elements important to solve the task. Contrary to perceptual augmentations, these augmentations require prior knowledge over the semantics of the observations. As such knowledge is often difficult to obtain and manipulate in complex scenarios, we propose to use similar tasks to T_d, such as video games from the same type or genre, as a way to provide meaningful semantic augmentations.

3.2 Pre-training with Augmentations

Motivated by recent approaches in self-supervised visual learning [5,15], we argue that by pre-training an agent on the augmented set of observations $O_p^\star$, we force it to learn features that are more general, and thus able to transfer to the downstream task T_d more efficiently, during the fine-tuning stage.

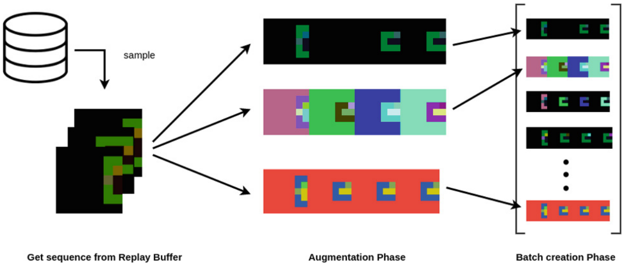
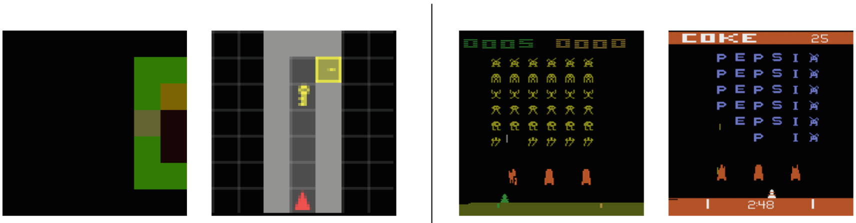


Fig. 2. The *Multiple-Augmented Pre-training Scheme* (MAPS) for efficient transfer of RL agents to novel similar tasks: initially, we obtain a sequence of observations that is used to train the agent; subsequently, we augment each specific sequence with a user-defined transformation; finally, we stack the multiple augmented sequences into a single training batch.



(a) MicroGrid. (b) MacroGrid (c) Space Invaders (d) Pepsi Invaders

Fig. 3. The environments employed in the evaluation of MAPS.

We denote our simple pre-training scheme with augmentations as *Multiple Augmented Pre-training Scheme* (MAPS). In MAPS, as shown in Fig. 2 each training sequence (either from the replay buffer or from the environment) is augmented with a random set of perceptual, dynamical and semantic augmentations. An augmentation can be applied per time-step or across multiple time-steps (such as throughout the episode). We then concatenate the diverse augmented sequences into a single batch, to be used in the training of the RL controller.

Despite the simplicity of the approach, we show in Sect. 4 that the joint pre-training approach of MAPS is able to outperform other transfer approaches in terms of sample-efficiency of the fine-tuning stage. By learning with the help of augmentations task, MAPS is able to generalize across a larger number of representations, thus being able to more easily adapt to new games.

4 Evaluation

We evaluate MAPS against other standard pre-training schemes in scenarios of increasing complexity, showing how our approach allows pre-trained model-based RL agents to efficiently transfer to novel, similar tasks.

4.1 Experimental Setup

To fully exploit the perceptual, dynamical and semantic variability within MAPS, we consider two different grid-based scenarios in our evaluation:

- *MicroGrid*: A smaller 5 × 5 grid world based on MiniGrid (Fig. 3a);
- *MacroGrid*: A larger 8 × 8 grid world based on MiniGrid, where the visual observations of the agents are upscaled to 64 × 64 pixels (Fig. 3b).

Both scenarios allow for fine control over the elements of the environment (such as colors, shapes and grid sizes), facilitating the creation of the necessary perceptual and dynamical augmentations for MAPS. In addition, the higher-resolution MacroGrid scenario allows to exploit semantic variability by changing the object sprites present in the environment. In both scenarios, we consider the *DoorKey* navigation task, which requires that the agent obtains a key to unlock the door that allows it to reach the goal. We instantiate the following classes of augmentations for MAPS in the grid-based scenarios:

- *Perceptual* (P): static color changes, color changes on every step, spatial visual changes;
- *Dynamical* (D): modifications that do not change optimal policy (random NoOp action), modifications that change the optimal policy (swap actions);
- *Semantic* (S): image occlusions (blinking), swap object sprites positions (only in MacroGrid), use different object sprites (only in MacroGrid).

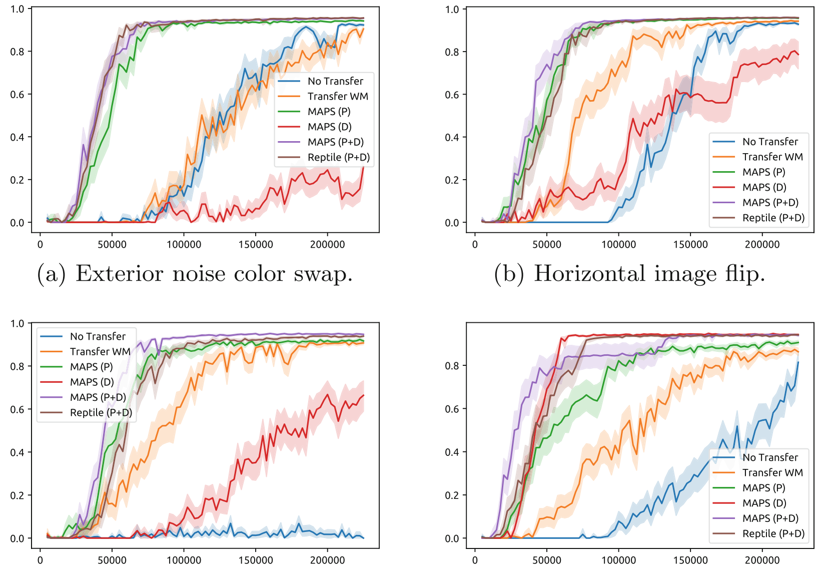
We employ a subset of 5 different augmentations as modified tasks: two perceptual augmentations (exterior noisy color swap and horizontal image flip), one dynamic augmentation (random NoOp action) and two semantic tasks (random black flicker, MacroGrid with semantic data).

Furthermore, we also test the MAPS framework in the Atari game environment [2], as shown in Fig. 3c, d , with image-based augmentations as the previous case. We evaluate the sample-efficiency of MAPS by following the metrics presented in [26]: an algorithm is more sample efficient than another if it reaches a higher performance in the same training window. If the algorithms present similar asymptotic performances, then we compare the jump-start performance and area under the curve.

4.2 Pre-training of Model-Based RL Agents

We introduce MAPS in the pre-training phase for a successful transfer of model-based agents. We initially consider two transfer scenarios: we pre-train the agent in the MicroGrid or MacroGrid scenario for 225k or 250k time steps, respectively. Then we transfer the learned world model to learn the downstream task, that consists of the original task with an augmentation previously not seen during the pre-training, following the augmentations in Sect. 4.1. We compare the fine-tuning performance of the agents that were pre-trained with MAPS against agents pre-trained without MAPS and agents without pre-training (learning from scratch). Furthermore, we also compare the MAPS approach to a meta learning approach that uses the same data augmentations, as meta learning as been successfully employed for transfer learning. As such we employ Reptile [19], a state-of-the-art first-order algorithm, as a baseline.

For these environments, we attempt multiple augmentation settings to better ascertain how the increase of pre-training variability can help transfer: single augmentations in perceptual (P) and dynamical (D) categories, combinations of augmentations like perceptual-dynamical (P+D) and complete perceptual-dynamical-semantic augmentations (P+D+S). We present the evaluation results for the MicroGrid scenario in Fig. 4. Overall, the results show that MAPS has a significant contribution to a positive transfer to the downstream task. This improvement is clearly seen for the *Exterior noisy color swap* downstream tasks,



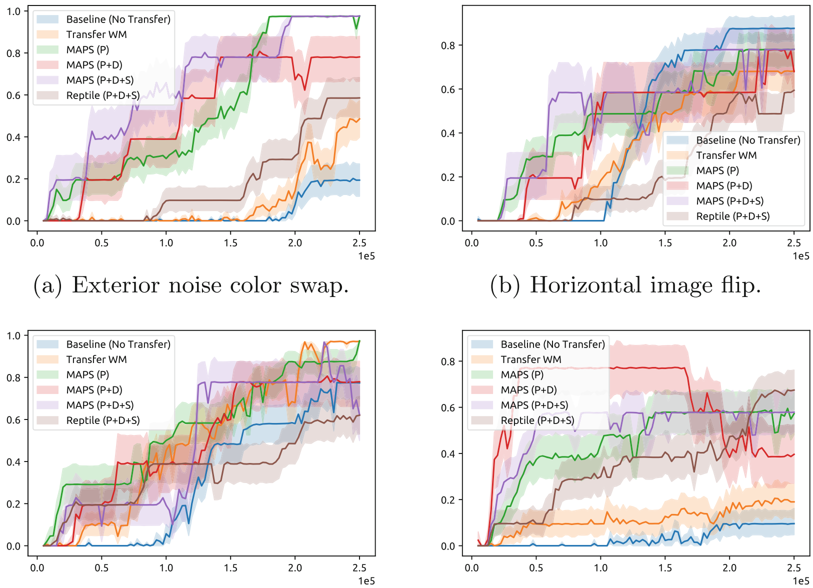
(a) Exterior noise color swap.

(b) Horizontal image flip.

(c) Random black flicker.

(d) Random NoOp action.

Fig. 4. Transfer performance of pretrained agents in MicroGrid to an augmentation task ($T_p \neq T_d$). Results averaged over 10 randomly-seeded runs.



(a) Exterior noise color swap.

(b) Horizontal image flip.

(c) Random black flicker.

(d) Random NoOp action.

Fig. 5. Transfer performance of pretrained agents in MacroGrid to an augmentation task ($T_p \neq T_d$). Results averaged over 10 randomly-seeded runs.

where introducing perceptual augmentations with MAPS during pre-training allows the agent to efficiently adapt to the downstream task. The improvement can be extended to all other augmentations, where transferring the world model pre-trained with MAPS results either in a higher asymptotic performance or/and a better jump-start learning performance. The same results are valid when comparing MAPS with Reptile, where MAPS always has a higher asymptotic performance or/and a better jump-start learning performance.

We verify a similar trend in the results for the MacroGrid scenario, presented in Fig. 5. The results show, once again, that transferring the world model pre-trained with MAPS results overall in a higher asymptotic fine-tuning performance or/and a better jump-start fine-tuning performance. Moreover, in the MacroGrid scenario we can also evaluate the transfer to tasks with distinct semantic features (such as when changing the sprites of the objects in the environment). We present such results in Fig. 6: only the agent that pre-trains with MAPS, considering perceptual and semantic augmentations, is able to positively transfer to this challenging task, with a significant jump-start fine-tuning performance over the baselines. The results also show that MAPS outperforms or has similar performance to the meta-learning approach of Reptile, thus showing that joint-training of tasks can still be a strong alternative over meta-learning methods. Overall, the results attest to the importance of introducing variability regarding perceptual, dynamical and semantic features using MAPS during the pre-training phase to a positive transfer of model-based RL agents.

4.3 Atari Games

Finally, to understand how the performance of MAPS scales to more complex scenarios, we evaluate our approach in Atari Games [2]. In this complex scenarios we both pre-train and fine-tune all agents up to 5M steps.

For this environment we considered general perceptual and dynamical augmentations, which can be employed in any image-based scenario, regardless of its complexity. Furthermore, due to the intrinsic difficulty of creating semantic augmentations in complex scenarios, we explore the use of similar tasks in order to exploit semantic features during the pre-training of MAPS. Therefore, we created and use a set of general perceptual, dynamical and semantical augmentations for any image based environments that we refer only as *MAPS*.

We consider two different training scenarios: initially, we select Pepsi Invaders as the pre-training task and Space Invaders as our downstream task, to evaluate the role of perceptual and dynamical augmentations in the performance of MAPS, a scenario we denote by *Single Task Transfer*. Secondly, we select a group of three similar games (Pepsi Invaders, Galaxian, Pigs in Space) as pre-training

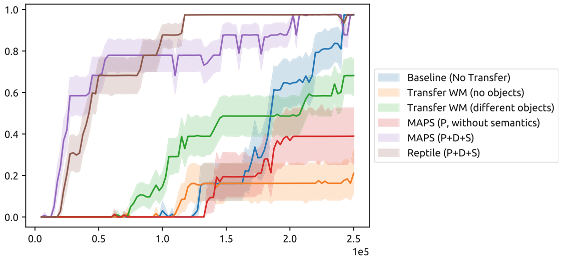
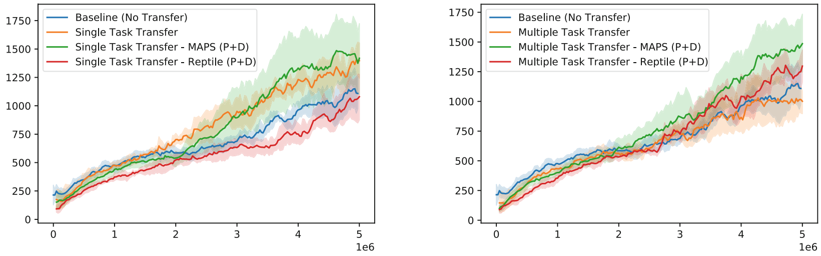


Fig. 6. Transfer performance of pretrained agents in MacroGrid to a semantically-augmented task ($T_p \neq T_d$). Results averaged over 10 randomly-seeded runs.



(a) Single Task Transfer

(b) Multiple Task Transfer

Fig. 7. Transfer performance of pretrained agents in **a** Pepsi and **b** 3 similar task to Space Invaders. Results averaged over 5 randomly-seeded runs

Table 1. Comparison of the final score at step 5M in Space Invaders.

Method	5M Avg. Score
DQN [18]	808 ± 38
C51 [1]	1035 ± 30
Rainbow [12]	1086 ± 81
IQN [6]	**1602 ± 153**
DreamerV2 [9]	1141 ± 295
Single Task Transfer	1393 ± 466
Single Task Transfer—MAPS (P+D)	**2032 ± 774**
Multiple Task Transfer	863 ± 335
Multiple Task Transfer—MAPS (P+D)	**1813 ± 576**
Single Task Transfer—Reptile (P+D)	1226 ± 458
Multiple Task Transfer—Reptile (P+D)	1719 ± 483

tasks, to exploit semantic variability, and transfer the agent to Space Invaders, a scenario we refer as *Multiple Task Transfer*. The selected pre-training tasks share the same grid-like structure of the enemies as the fine-tuning task.

We present our results in Fig. 7. In the Single Task Transfer we verify that pre-training on the similar Pepsi task yields a positive improvement, while pre-training jointly on Multiple Task Transfer has a negligible to negative performance over training from scratch. Using MAPS brings a significant improvement over the Single Task pre-training, making it the best performing method for training the Space Invaders tasks with 5M time-steps when compared with other publicly available results in Table 1. In the Multiple Task Transfer scenario (Fig. 7), the results show once again that employing MAPS allows for a significant positive transfer over the baseline and over the naive pre-training approach. Is worth mentioning that while both pre-training methods with MAPS have a high mean, both also have a big variance in the results, as seen in Table 1. This higher variance in the performance results also seems to be a characteristic of the DreamerV2 method. Thus, we can ascertain that our methods using MAPS are better with a 95% CI than DQN [18], C51 [1] and Rainbow [12], while being competitive with IQN [6] and baseline DreamerV2 [9]. On another hand, we can also conclude that learning using Multiple Task Transfer with MAPS is significantly better than without.

5 Conclusions

In this work, we investigated the introduction of perceptual, dynamical and semantic variability during the pre-training of model-based RL agents for an efficient transfer of the agents to novel tasks. We contributed with MAPS, a novel pre-training scheme that introduces augmentations over the observations of the agent to take advantage of such variability. Our results show that MAPS improves the fine-tuning efficiency of pre-trained agents to novel downstream tasks. In future work, we will explore how MAPS can be used to improve transfer in model-free RL, as well as accessing MAPS in other model-based agents, which might employ different auxiliary losses like contrastive learning methods.

References

1. Bellemare, M., Dabney, W., Munos, R.: A distributional perspective on reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning, pp. 449–458 (2017)
2. Bellemare, M., Naddaf, Y., Veness, J., Bowling, M.: The arcade learning environment: an evaluation platform for general agents. J. Artif. Intell. Res. **47**, 253–279 (2013)
3. Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. CoRR abs/2108.07258 (2021)
4. Brown, T., et al.: Language models are few-shot learners. CoRR abs/2005.14165 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, pp. 1597–1607 (2020)
6. Dabney, W., Ostrovski, G., Silver, D., Munos, R.: Implicit quantile networks for distributional reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning, pp. 1096–1105 (2018)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018)
8. Grill, J., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 21271–21284 (2020)
9. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering Atari with discrete world models. CoRR abs/2010.02193 (2020)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
11. Henderson, P., et al.: Deep reinforcement learning that matters. In: Proceedings of the 32nd AAAI Conf. Artificial Intelligence, pp. 3207–3214 (2018)
12. Hessel, M., et al.: Rainbow: combining improvements in deep reinforcement learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 3215–3222 (2018)
13. Lake, B., Ullman, T., Tenenbaum, J., Gershman, S.: Building machines that learn and think like people. Behav. Brain Sci. **40** (2017)
14. Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., Srinivas, A.: Reinforcement learning with augmented data. In: Advances in Neural Information Processing Systems, vol. 33, pp. 19884–19895 (2020)
15. Laskin, M., Srinivas, A., Abbeel, P.: CURL: contrastive unsupervised representations for reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning, pp. 5639–5650 (2020)
16. Levine, S., Finn, C., Darrell, T., Abbeel, P.: End-to-end training of deep visuomotor policies. CoRR abs/1504.00702 (2016)

17. Lillicrap, T., et al.: Continuous control with deep reinforcement learning. CoRR abs/1509.02971 (2015)
18. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
19. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. CoRR abs/1803.02999 (2018)
20. Obando-Ceron, J., Castro, P.: Revisiting Rainbow: Promoting more insightful and inclusive deep reinforcement learning research. In: Proceedings of the 38th International Conference on Machine Learning, pp. 1373–1383 (2021)
21. Schwarzer, M., et al.: Pretraining representations for data-efficient reinforcement learning. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
22. Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., Pathak, D.: Planning to explore via self-supervised world models. CoRR abs/2005.05960 (2020)
23. Silver, D., et al.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
24. Stooke, A., Lee, K., Abbeel, P., Laskin, M.: Decoupling representation learning from reinforcement learning. In: Proceedings of the 38th International Conference on Machine Learning, pp. 9870–9879 (2021)
25. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Proceedings of the International Conference on Artificial Neural Networks, pp. 270–279 (2018)
26. Taylor, M., Stone, P.: Transfer learning for reinforcement learning domains: a survey. J. Mach. Learn.Res. **10**(7) (2009)