Adapting to the Unfamiliar: Leveraging Latent Space in Reinforcement Learning for Cross-Environment Generalization

Lukas von Briel Jona Schulz Pablo Soler

Abstract

We trained a DreamerV2 world model on a multitude of Atari game environments to learn a common latent space dynamics model that extracts general, fundamental features of the seen environments. We then fine-tuned the model on a novel environment while training an actor on top of it and compared the resulting reinforcement learning agent to a model trained from scratch only on the novel environment. Our results show a benefit of using a pre-trained world model for faster learning of the novel environment.

1. Introduction

Leveraging past experience about the dynamics of the world around us is something we humans do all the time. It allows us to quickly adapt to new tasks and environments without having to re-learn everything from scratch. The hippocampus plays a key role in this by learning abstract representations of the state of the environment with the help of grid and place cells (Whittington et al., 2022; O'Keefe & Dostrovsky, 1971; Moser et al., 2017). In the machine learning domain, world models are a promising component to help reinforcement learning (RL) models do exactly that. A world model is trained to encode observations from an environment into a latent space and predict future states from the current state and actions taken (Ha & Schmidhuber, 2018). World-model-based RL agents have repeatedly shown to outperform other RL models in recent years (Hafner et al., 2020; 2022; 2023; Micheli et al., 2023)

Past work has been done on the topic of generalization of RL models (Anand et al., 2021; Cobbe et al., 2019). In our project we aimed at training a world model on a multitude of Atari game environments in order to learn a latent space model that extracts fundamental dynamics common to all of the experienced environments. Our hypothesis is that this allows for a novel environment to be learned in a relatively short fine-tuning procedure ultimately speeding up the training of a new RL agent to act effectively in this new environment.

1.1. Research Questions

The following questions are addressed in our work:

- Does the use of a world model pre-trained on multiple environments allow for a faster training of an RL agent on a novel environment?
- How does a world model encode environment properties that are consistent across different environments?

2. Models and Methods

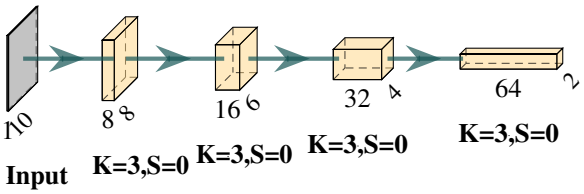
2.1. Environments

Due to the high computational cost of training DreamerV2 models on full-scale Atari game environments, we opted for the smaller MinAtar (Young & Tian, 2019) set of environments for initial experiments. MinAtar is a set of five different Atari game environments reduced to 10-by-10 pixel observations and six possible actions. In a second experiment we then trained on the original Atari game environments provided in the OpenAI Gym framework (Brockman et al., 2016).

2.2. Model

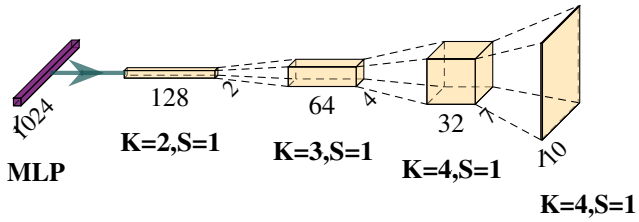
We worked with DreamerV2 (Hafner et al., 2022) which consists of a world model and an actor-critic model. The world model encodes observations into latent state representations. The dynamics of the environment are modeled in this latent space using a recurrent state-space model (RSSM). The model is not only trained to represent a sequence of observations from the environment but is also capable of predicting a sequence of future states from an initial observation in a process called "dreaming". The world model is optimized based on a combination of multiples losses: image, reward and terminal loss measure the quality of image reconstruction, reward prediction and episode termination prediction respectively using a log-probability loss function. An additional KL loss measures the similarity between the model component encoding the current state of the environment and the component predicting the current environment state from the previous one using the two distributions' KL divergence.

2.3. Encoder and Decoder for MinAtar



MinAtar Encoder CNN Structure

In order to process the small observations from MinAtar environments a modification to the observation encoder and decoder parts of DreamerV2 were necessary. The original encoder consists of four convolutional layers with 4x4 kernels, stride 2 and no padding. This does not work with the 10x10 pixel observations from MinAtar. Our alternative architecture consists of four convolutional layers with 3x3 kernels, stride 1 and no padding. This scales the input image in the MinAtar case to a 2x2x64 output which is a lot smaller than the Atari encoding of size 3x3x256.



MinAtar Decoder CNN Structure

Similarly, the observation decoder had to be modified. In the original architecture it consists of four deconvolutional layers, the first two with a 5x5 kernel and the last two with a 6x6 kernel. Again, a stride of 2 and no padding is used. Our alternative architecture consists of four deconvolutional layers with kernel sizes 2x2, 3x3, 4x4 and 4x4, respectively, as well as stride 1 and no padding.

2.4. Training Procedure

The classic DreamerV2 model trains its world model and actor-critic simultaneously and uses the actor in the environment to build its dataset during training. Since our approach requires pre-training of only the world model, there is no need for us to train an actor-critic on each of the environments seen during pre-training. This, however, means that an alternative way of acquiring observations from the environments is needed to build a dataset. We therefore separated the stages of dataset acquisition and world model training resulting, in total, in a three step process:

- **Dataset acquisition**: Observations are gathered from multiple environments using both a random policy and a variety of trained DQN-based agents (see below). Episodes from different environments are mixed such that a single batch may contain sequences from different environments.
- **World model pre-training**: A DreamerV2 world model is then trained on the acquired dataset. The actor-critic part of the DreamerV2 model are not optimized in this stage.
- **Fine-tuning on a novel environment**: A complete DreamerV2 agent is then initialized with the world model learned in the previous step and trained on a novel game environment.

2.5. DQN Agents for Data Acquisition

For data acquisition in Atari environments, DQN agents from a public collection of model checkpoints (Gogianu et al., 2022) are used. Four such model checkpoints, taken from different steps within a single training run, are used for each environment. For MinAtar environments we trained our own DQN agents on all game environments used during world model training. The model architecture in this case is a three-layer MLP. Similarly to the Atari case, several checkpoints are used from each training run. During data acquisition a random switch from one model to another occurs at regular intervals to increase the diversity of the acquired observation sequences.

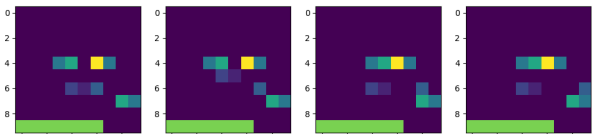
3. Results

All experiments were conducted using the default parameters as proposed by (Hafner et al., 2022), unless stated otherwise.

3.1. MinAtar

For our first set of experiments we trained a DreamerV2 world model for 280k steps on four out of the five available MinAtar game environments (Asterix, Breakout, Freeway and Seaquest).

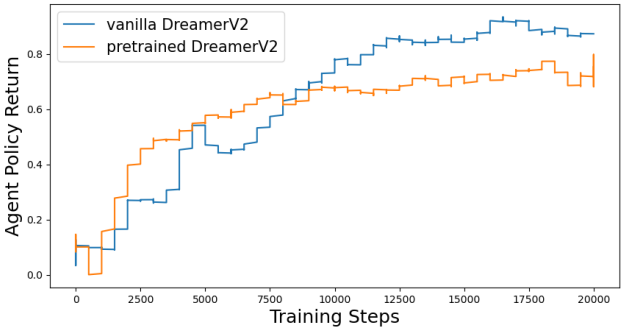
As an example for the quality of the trained world model, the figure below depicts a sequence of dreamed observations of the Seaquest environment. The model correctly predicts the dynamics of moving player, projectiles and other objects.



Frames from a dreamt sequence in Seaquest produced by the multi-environment world model.

To prove the model had been able to learn across the four different games, agents were trained on all of the learned

games while fixing the weights of the world model. The figure below shows a comparison on Seaquest between the rewards per episode of the fixed world model while training the agent and of a Dreamer V2 model trained from scratch. The rewards both start at zero as both agents are initialized randomly. The multi-environment world model agent is able to learn faster and ultimately converges to a slightly worse average return per episode than the DreamerV2 model trained from scratch.

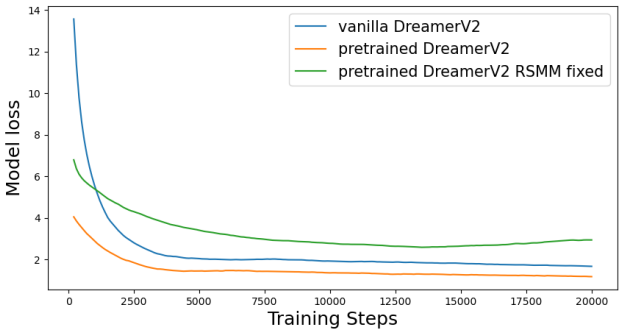


Reward per episode comparison of a DreamerV2 agent trained on top of a pre-trained multi-environment world model (world model weights fixed) and an agent trained from scratch on MinAtar Seaquest. Seaquest is one of the environments the world model was pre-trained on.

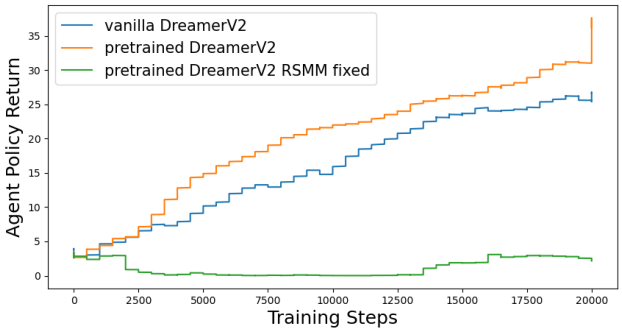
Subsequently, we proceeded to fine-tuned the model on a new environment. Space Invaders was used as the only game the trained model had not seen yet. Two different models were trained. For the first one the weights of the RSSM were fixed thus only the encoder and decoder components of the world model were fine-tuned. The goal was to see if the recurrent neural network performs well on a new environment and therefore has sufficiently learned general concepts and is able to transfer them to new environments. The encoder and decoder were fine-tuned to convert the new game input into the latent space of the fixed RSSM world model. For the second model the entire world model was fine-tuned. This test intends to show if the weights of the trained model can be fine-tuned to a new environment and hence, have general knowledge that can be adapted to new contexts. The actor-critic components of the DreamerV2 agent were trained as well during fine-tuning to produce a complete agent model. Again, for comparison a DreamerV2 model was trained from scratch.

In the figure below one can see that the best and fastest convergence is achieved by the multi-environment model whose weights were all fine-tuned. Also, even though the image reconstruction loss and reward prediction loss are better for the world model with fixed RSSM than for the DreamerV2 model trained from scratch, the overall world model loss is worse. The reason for this is the high KL-loss resulting in poor quality of dreamed episodes.

Furthermore, the second figure below depicts the average return per episode of the different world model agents. Again, the fine-tuned multi-environment model converges faster and to a better return value. Additionally, the agent trained on the world model with the fixed RSSM from the multi-environment model performs poorly compared to the other two world models, likely a result of the poor quality of dreamed episodes that are crucial for actor-critic learning.



World model loss comparison of a DreamerV2 agent with pre-trained world model with and without a fixed RSSM and an agent trained from scratch on MinAtar Space Invaders.

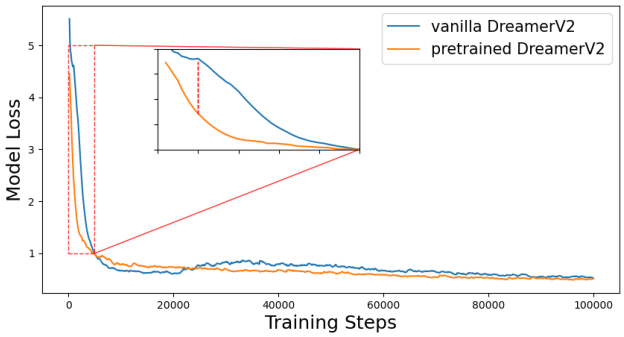


Reward per episode comparison of a DreamerV2 agent with pre-trained world model with and without a fixed RSSM and an agent trained from scratch on MinAtar Space Invaders.

3.2. Atari

We then conducted an experiment on full-scale Atari game environments. We trained a world model on nine different environments (all of them being some variation of a shooter game with one-dimensional player movement) for 165k steps. We then fine-tuned the resulting model on the Breakout game environment for 100k steps. An additional agent was trained from scratch as a comparison. This final test was conducted to verify the consistency of the results from the MinAtar games in more complex environments.

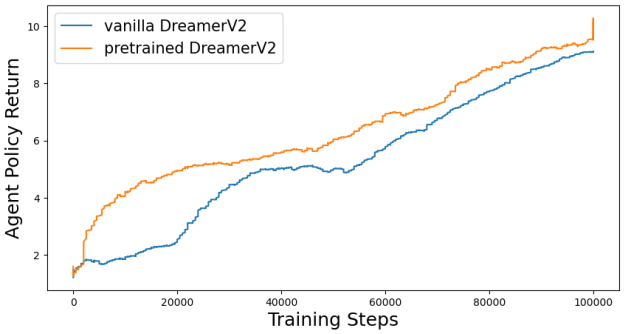
The figure below depicts the world model loss of the fine-tuned multi-environment world model and a DreamerV2 model trained from scratch.



World model loss comparison of a pre-trained and a non-pre-trained DreamerV2 agent during fine-tuning on Breakout.

Looking at the world model training loss in the figure below during the first 5000 training steps, we can see that the pre-trained model loss decreases much faster than the vanilla model loss. The pre-initialised weights allow the Dreamer to infer some information from the seen sequence even in the first 100 training steps. At 1000 steps (marked by the red vertical line) the largest offset between the two losses can be observed. After this point, the vanilla Dreamer catches up and both losses converge to approximately the same value.

The reward per episode of the trained agent acting in the Breakout environment remains higher for the pre-trained model throughout all 100k training steps with the difference in performance being particularly large between 10k and 20k steps.



Reward per episode comparison of a pre-trained and a non-pre-trained DreamerV2 agent during fine-tuning on Breakout.

3.3. Latent Space Analysis

Finally, we inspected the 1280-dimensional latent space of DreamerV2's RSSM to find abstract environment information with different statistical analysis tools. All the tests were conducted in the MinAtar Breakout environment using a set of sequences containing 100,000 frames. We used k-means to find clusters in the feature space and looked for events in the environment that might trigger each cluster but did not find conclusive results. The most relevant test was calculating the canonical cross correlation (CCA). The CCA calculates the linear combinations of the feature space and the player position which have maximum correlation with each other. This resulted in one feature in the feature space having a significantly higher correlation than any other feature having an almost linear dependence to the x coordinate of the player position. This feature also had six times higher correlation than the highest correlation occurring in randomly generated feature space vectors.

4. Discussion

As evident by the results of the fine-tuning experiments, using a pre-trained world model is beneficial for quicker training of DreamerV2 agents on novel environments. The performance difference between a pre-trained and non-pre-trained model is particularly large in the first 10k iterations for MinAtar and the first 20k iterations for the investigated Atari environment. Using a pre-trained world model could therefore be useful in situations where long training in a new environment is not possible or too time-consuming.

Fine-tuning all parts of the world model - RSSM, encoder and decoder - was shown to be critical as freezing the pre-trained RSSM during fine-tuning resulted in little to no performance increase of the agent. We plan to repeat this experiment in the future with a larger set of environments and much longer training duration in order to train a possibly more generalizable latent space dynamics model.

Due to limited computational resources we were unable to train all models to convergence, especially in Atari environments. It is important to verify the conclusions drawn from our experiments by conducting them again on a larger scale.

5. Summary

To conclude, we adapted the DreamerV2 architecture to MinAtar environments and developed a three-stage training framework consisting of a multi-environment dataset generation, a multi-environment world model training and a fine-tuning stage. Our work demonstrates the benefit of a pre-training/fine-tuning procedure for RL environments where fast learning in a new environment is of importance. The generalization capabilities of the world models trained within the scope of this project remain limited, however, possibly due to the limited amount of training steps and number of environments.

References

Anand, A., Walker, J., Li, Y., Vértes, E., Schrittwieser, J., Ozair, S., Weber, T., and Hamrick, J. B. Procedural generalization by planning with self-supervised world models, 2021.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. Quantifying generalization in reinforcement learning, 2019.

Gogianu, F., Berariu, T., Bușoniu, L., and Burceanu, E. Atari agents, 2022. URL https://github.com/floringogianu/atari-agents.

Ha, D. and Schmidhuber, J. World models. 2018. doi: 10.5281/ZENODO.1207631. URL https://zenodo.org/record/1207631.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination, 2020.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models, 2022.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models, 2023.

Micheli, V., Alonso, E., and Fleuret, F. Transformers are sample-efficient world models, 2023.

Moser, E. I., Moser, M.-B., and McNaughton, B. L. Spatial representation in the hippocampal formation: a history, November 2017. URL http://dx.doi.org/10.1038/nn.4653.

O'Keefe, J. and Dostrovsky, J. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, 1971. ISSN 0006-8993. doi: https://doi.org/10.1016/0006-8993(71)90358-1. URL https://www.sciencedirect.com/science/article/pii/0006899371903581.

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., and Behrens, T. E. J. How to build a cognitive map: insights from models of the hippocampal formation, 2022.

Young, K. and Tian, T. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments, 2019.