First Paper Summary, EE245 Spring 2025

Kunyi Yu
University of California, Riverside
900 University Ave. Riverside, CA 92521, USA
kyu135@ucr.edu

Abstract

The first paper summary due on Monday, Apr. 28, 2025. The author chose the paper titled "Mastering Diverse Domains through World Models" by Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap.

The following content covers 1) major contributions, 2) prior work, 3) method, 4) major results, 5) strengths, and 6) weaknesses.

1 Summary of Major Contributions

The main contribution of this paper [1] is the introduction of a general reinforcement learning algorithm called Dreamer (specifically, DreamerV3), which contains a world model, a critic neural network (NN), and a actor NN. Dreamer masters a wide range of domains (over 150 tasks) with fixed hyperparameters. Furthermore, the algorithm can learn robustly across different data and compute budgets, making it applicable to a variety of practical applications. It overcomes the challenges of robustly learning through techniques based on normalization, balancing, and transformations.

The paper conducts comprehensive experiments to evaluate the performance of Dreamer across different environments, including Minecraft, Atari, ProcGen, and others. In the task of Minecraft, Dreamer is the first algorithm to collect diamonds from scratch while keeping configurations the same. The algorithm shows the ability to learn farsighted strategies from pixels and sparse rewards in an open world.

Dreamer paves the way for future research directions, including teaching agents world knowledge from internet videos and learning a single world model across domains. This allows artificial agents to build up increasingly general knowledge and competency.

2 Relation to Prior Work

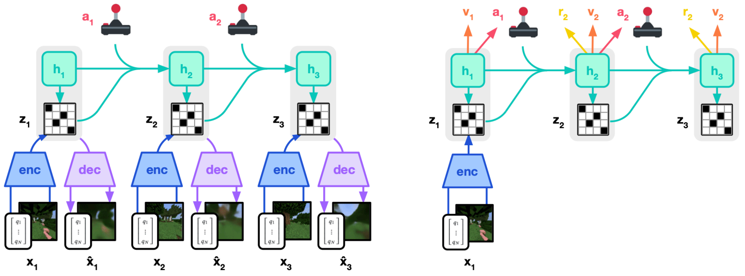
Plenty of prior work has been done in the field of general-purpose reinforcement learning algorithms, such as PPO [2], SAC [3], MuZero [4], and Gato [5]. However, these algorithms face challenges in lower performance, requirements for tuning, no open-source code, or demanding expert data. Dreamer, on the other hand, “masters a diverse range of environments with fixed hyperparameters, does not require expert data, and its implementation is open source” [1].

Besides over 150 tasks conducted in the paper, Dreamer demonstrates its outstanding performance in the Minecraft environment, MALMO (Microsoft version) [7] and MineRL (experiments used version) [6]. Prvious methods like Voyager [8] and VPT [9] indeed show impressive performance in Minecraft, but they require specifically engineered or hundreds of computational resources. In contrast, Dreamer autonomously learns to collect diamonds within very limited resources and without human data.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

3 Summary of the Method

The third generation of Dreamer is a model-based, online, general-purpose reinforcement learning algorithm. The algorithm has three NNs: a world model, a critic, and an actor. All the three NNs are trained concurrently from replayed experience. The figure below shows the architecture of DreamerV3 are copied from the paper:



(a) World Model Learning

(b) Actor Critic Learning

Figure 1: Architecture of DreamerV3 [1]

World model learning is responsible for learning the dynamics of the environment by predicting future representations and rewards for potential actions. The paper implement a Recurrent State-Space Model (RSSM) to learn the world model. Given a batch of inputs, actions, rewards, and continuations flags, the world model are optimized by minimizing the three losses by weights: prediction loss $\beta_{pred}\mathcal{L}_{pred}$, dynamics loss $\beta_{dyn}\mathcal{L}_{dyn}$, and representation loss $\beta_{rep}\mathcal{L}_{rep}$. These losses allows world model with fixed hyperparameters to fit across domains.

Actor Critic learning are conducted entirely within the latent space using abstract trajectories predicted by the world model. The actor and critic operate on model states $s_t \doteq \{h_t, z_t\}$. For each model state, the actor network $a_t \sim \pi_\theta(a_t|s_t)$ aims to maximize the discount return $R_t \doteq \sum_{\tau=1}^{\infty} \gamma^\tau r_{t+\tau}$ with discount factor $\gamma = 0.997$. Meanwhile, the critic network $v_\psi(R_t|s_t)$ estimates the state-value function for each state under the current policy, which can approximate the γ-reward over prediction horizon $T = 16$. The customized A-C framework make DreamerV3 more efficient and robust.

4 Summary of Major Results

The paper conducts comprehensive experiments to evaluate the performance of DreamerV3 across 8 domains – over 150 tasks – using the same hyperparameters. PPO [2] is used as the baseline compared to DreamerV3 and other SOTA algorithms. Noticeably, DreamerV3 also runs on the challenging Minecraft environment. All the experiments are conducted on a single NVIDIA A100 GPU, and the code and results are open-sourced. A summary of the results from the paper is shown in the figure below:

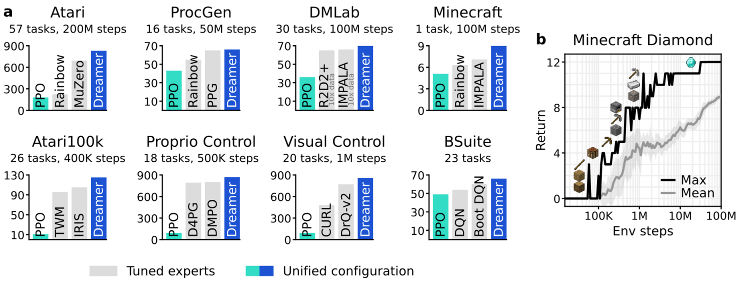


Figure 2: Summary of benchmark results [1]

Shortly, DreamerV3 outperforms all the other SOTA algorithms in most of the tasks. The world model also shows its 1) low computation cost, 2) robustness, 3) generalization, 4) data efficiency, 5) scalability, and 6) capability of handling sparse rewards.

5 Summary of Strengths

* *The current and following sections are the combination of the author's own understanding with internet reviews. Here I omit the references of these reviews.*

Firstly, the paper is very well organized. The introduction section is clear and concise, providing a good overview of the paper's contributions and significance. The arXiv version of the paper contains over 20 pages of appendices and all the code/results are public accessible in their website. The strict research process and outputs provide a good reference for future research.

Furthermore, the DreamerV3 algorithm is a significant improvement over previous versions and other SOTA algorithms. The use of a world model, actor-critic framework, and the ability to learn from pixels and sparse rewards are impressive. The series of Dreamer algorithms proposes an approach to the problem of model inaccuracy that has plagued researchers in model-base RL for a long time.

Last but not least, the performance of DreamerV3 in not only the Minecraft but also other 150 tasks is really impressive. A reduction in computational cost by several orders of magnitude may very likely make (previously computationally expensive) RL more accessible to researchers and practitioners.

6 Summary of Weaknesses

* *I may not be the best person to judge the weaknesses of this paper, but I try my best to summarize some of the weaknesses I found.*

Firstly, the actor-critic framework runs on the latent space, which may cause policy performance limited by the world model accuracy. What's more, the actor-critic framework may still yield out a difficult-to-learn policy in a extremely sparse reward environment. Meanwhile, the algorithm still requires the analysis of explainability, but it is a common challenge in deep RL. The black box nature makes it hard to be well trusted by human users. Lastly, the computation cost of DreamerV3 (specifically in the complex environments like Minecraft) is still need to be improved.

Overall, the series of Dreamer paper are a significant and landmark contribution to the field of reinforcement learning.

References

[1] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. *URL https://arxiv. org/abs/2301.04104*, 2023.

[2] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[4] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[5] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[6] Stephanie Milani, Nicholay Topin, Brandon Houghton, William H Guss, Sharada P Mohanty, Oriol Vinyals, and Noboru Sean Kuno. The minerl competition on sample-efficient reinforcement learning using human priors: A retrospective. *Journal of Machine Learning Research*, 1: 1–10, 2020.

[7] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *Ijcai*, volume 16, pages 4246–4247, 2016.

[8] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[9] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Acknowledgments

This paper summary was independently conducted by the author Kunyi Yu. Meanwhile, the author would like to thank Professor Jiachen Li and TAs for their patient guidance and help.