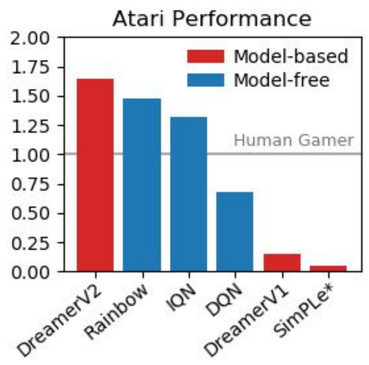
Mastering Atari with Discrete World Models

Danijar Hafner[12], Timothy Lillicrap[3], Mohammad Norouzi[1], Jimmy Ba[2]

[1] Google Brain [2] University of Toronto [3] DeepMind

1 Introducing DreamerV2

1. RL agent that learns purely from latent space predictions of a learned **world model**
2. First world model to achieve **human-level** performance on the Atari 200M benchmark
3. **Outperforms** top model-free agents with the same amount of data and computation time
4. Vector of **categorical latents** enforces learning sparse state representations
5. **KL balancing** scales prior vs posterior loss to encourage accurate prior dynamics



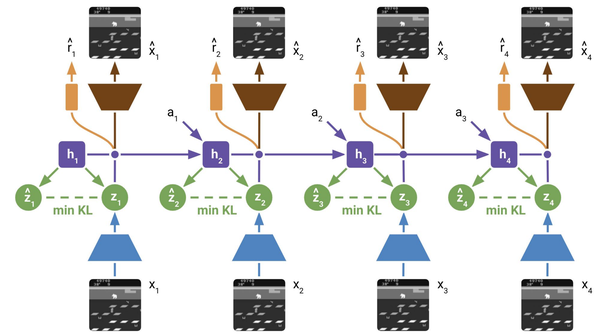
Gamer normalized task median over 55 games at 200M frames with sticky actions.

According to its authors, SimPLe trained for fewer steps but does not improve further after that.

2 World Model Learning

Based on the **latent dynamics model** of PlaNet that predicts ahead in a compact latent space and is trained as a sequential VAE

Each latent state is a vector of 32 **categoricals** with 32 classes each trained via **straight-through** gradients (1 line of code)



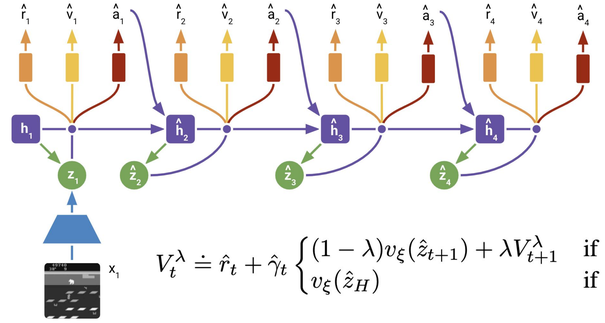
KL balancing scales the two terms in the KL, the posterior entropy and the prior cross entropy, to encourage accurate prior dynamics

$$\mathrm{KL}_{\eta_q,\eta_p}\big[q_t \,\|\, p_t\big] \doteq \eta_p \mathrm{E}_{q_t}\big[-\ln p_t\big] - \eta_q \mathrm{H}\big[q_t\big] \qquad \begin{aligned} p_t &\doteq p_\phi(\hat{z}_t \mid h_t) \\ q_t &\doteq q_\phi(z_t \mid h_t, x_t) \end{aligned}$$

3 Actor Critic Learning

Learn **neural network actor** and **state-value critic** from imagined rollouts in the compact state space of the world model

Compact states enable highly efficient predictions of several thousands of rollouts with a **large batch size** on a single GPU

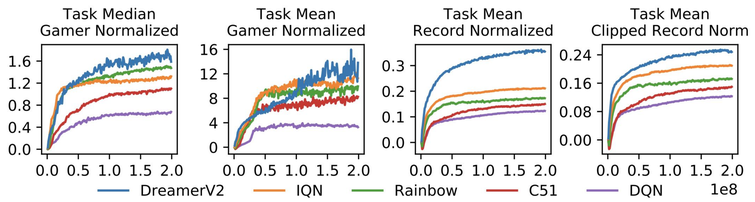


$$V_t^\lambda \doteq \hat{r}_t + \hat{\gamma}_t \begin{cases} (1-\lambda) v_\xi(\hat{z}_{t+1}) + \lambda V_{t+1}^\lambda & \text{if} \quad t < H, \\ v_\xi(\hat{z}_H) & \text{if} \quad t = H. \end{cases}$$

For each imagined trajectory, we compute the **lambda return** that is an average of n-step returns of all horizons n

- The critic is trained to predict the lambda return via squared loss
- The actor is trained to maximize the lambda return, using both **reinforce** gradients and **straight-through** gradients

4 Atari Benchmark

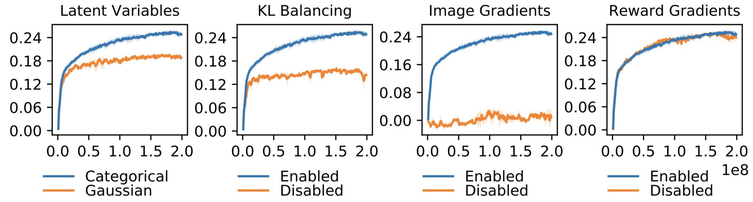


Comparison on the established Atari 200M benchmark of 55 games with sticky actions. We consider the following score aggregations:

- **Task Median, Gamer Normalized** Most common in the literature but median score ignores almost half of the games
- **Task Mean, Gamer Normalized** Dominated by few games with low gamer baseline (Crazy Climber, Video Pinball, James Bond)
- **Task Mean, Record Normalized** Suggested by Toromanoff et al. (2019) but few games still contribute the most
- **Task Mean, Clipped Record Norm** We recommend clipping scores to not exceed world record, which results in a robust metric

DreamerV2 outperforms the top single-GPU algorithms **Rainbow** and **IQN** in all four metrics

5 Ablation Study



To understand the contributions to performance, we disable key components of the world model:

- **Categorical latents** contribute substantially to performance
- **KL balancing** contributes substantially to performance
- **Image gradients** are essential for representations in DreamerV2
- **Reward gradients** are not necessary for learning representations

Agent	Task Median Gamer Normalized	Task Mean Gamer Normalized	Task Mean Record Normalized	Task Mean Clipped Record Norm
DreamerV2	1.59	13.81	0.35	0.25
No Reward Gradients	1.55	13.85	0.31	0.24
No KL Balancing	0.85	4.08	0.15	0.14
No Discrete Latents	0.85	3.96	0.24	0.19
No Image Gradients	0.06	0.37	0.01	0.01

6 Related & Future Work

Algorithm	Reward Modeling	Image Modeling	Latent Transitions	Single GPU	Trainable Parameters	Atari Frames	Accelerator Days
DreamerV2	✓	✓	✓	✓	22M	200M	10
SimPLe	✓	✓	✗	✓	74M	4M	40
MuZero	✓	✗	✓	✗	40M	20B	80
MuZero Reanalyze	✓	✗	✓	✗	40M	200M	80

SimPLe Learns a large world model that predicts forward in image space, making it slow to learn the policy. The highest performance of SimPLe, which is achieved after 2M frames, is far behind that of a human gamer or model-free agents.

MuZero Learns a task-specific sequential value function similar to VPN without leveraging image information. Uses sophisticated MCTS tree search that would be compatible with DreamerV2. Requires an expensive distributed training setup and is closed source.

What did not work Vector of binary latents, single categorical latents, long-term entropy bonus as part of the value function, free bits, KL annealing, using only straight-through policy gradients

Future work Multi-task transfer, long-term memory, global exploration using model uncertainty, hierarchical latents and temporal abstraction, environments with real-world visual complexity, better policy optimization using techniques from the model-free literature

@danijarh

Website with videos and code: danijar.com/dreamerv2