Learning Latent Dynamics for Planning from Pixels

Danijar Hafner[12], Timothy Lillicrap[3], Ian Fischer[4], Ruben Villegas[15], David Ha[1], Honglak Lee[1], James Davidson[1]

[1] Google Brain [2] University of Toronto [3] DeepMind [4] Google Research [5] University of Michigan

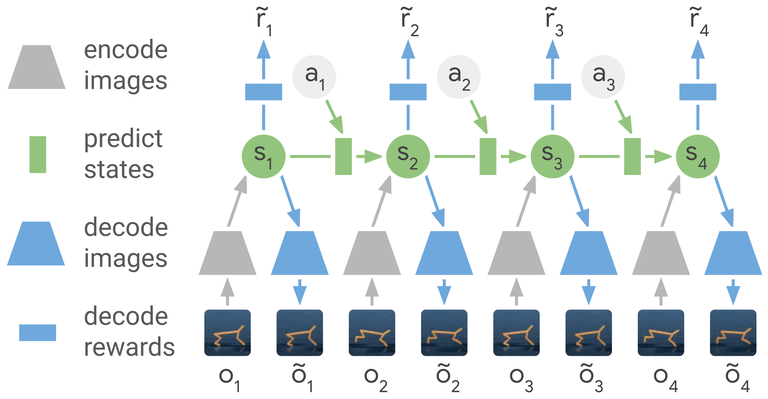
1 We introduce PlaNet

1. Planning is a powerful approach when the rules of the environment are known, for example in AlphaGo.
2. To plan in unknown environments, the agent needs to learn a world model over time from experience.
3. We present the Deep Planning Network (PlaNet), a recipe for scalable model-based RL using latent planning.
4. By planning in latent space, PlaNet reaching top model-free performance on visual control tasks in 200X fewer episodes.
5. We also train one agent to make accurate video predictions and solve all six considered tasks without task labels.

2 Latent Dynamics Model

Our model predicts future images and rewards using a sequence of compact latent states, trained as a sequential VAE.

Reconstructing images provides a rich training signal but is not needed during planning.

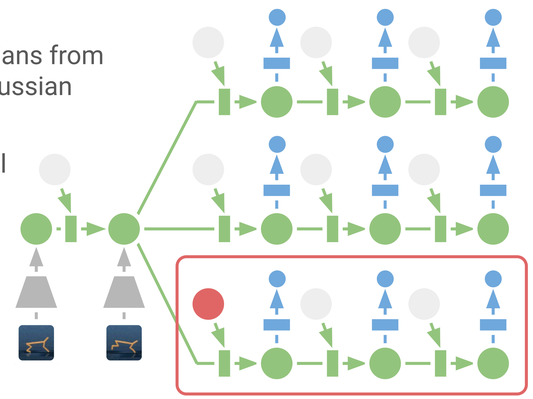


3 Planning in Latent Space

PlaNet selects actions by simple population based search, without needing a policy network. We replan at every step.

We only need future rewards not images, allowing to evaluate thousands of action sequences in parallel in a large batch.

1. Sample 1000 plans from a factorized Gaussian
2. Evaluate them using the model
3. Re-fit Gaussian to the top 100
4. Repeat 1–3 for 10 iterations

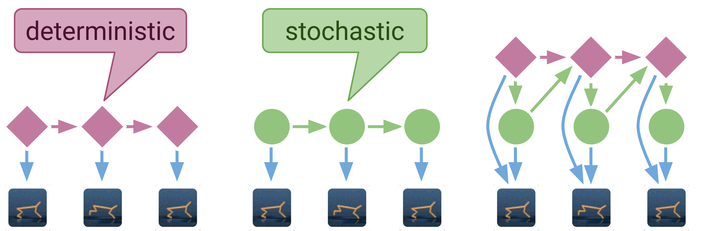


4 Recurrent State Space Model

Deterministic states can pass information far into the future, needed for accurate predictions.

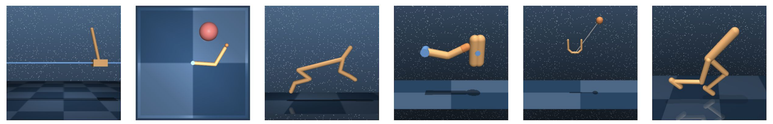
Stochastic states can predict multiple futures, needed for partial observability and model uncertainty.

Our experiments show that having both stochastic and deterministic state elements is crucial for planning performance.



Recurrent Neural Network (RNN) | State Space Model (SSM) | Recurrent State Space Model (RSSM)

5 Comparison to Model-Free Agents



(a) Cartpole (b) Reacher (c) Cheetah (d) Finger (e) Cup (f) Walker

Our visual control tasks include partial observability, contact dynamics, sparse rewards, and many degrees of freedom.

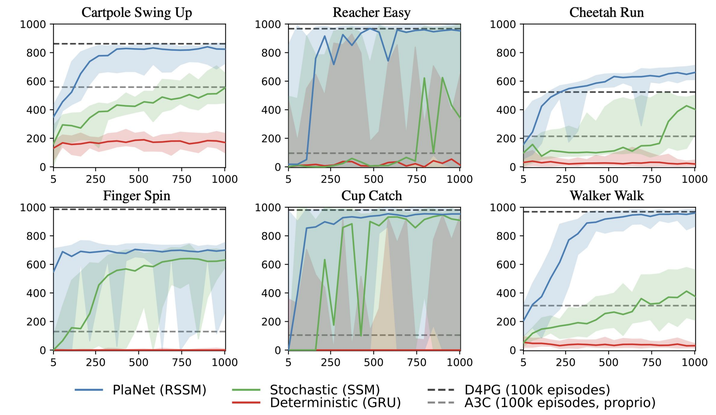
PlaNet reaches the performance of top model-free agents in 200x fewer episodes on average on these tasks.

Method	Modality	Episodes	Cartpole Swing Up	Reacher Easy	Cheetah Run	Finger Spin	Cup Catch	Walker Walk
A3C	proprioceptive	100,000	558	285	214	129	105	311
D4PG	pixels	100,000	862	967	524	985	980	968
PlaNet (ours)	pixels	1,000	821	832	662	700	930	951
Data efficiency gain PlaNet over D4PG (factor)			250	40	500+	300	100	90

Training time: 1 day on a single GPU

6 Comparison of Model Designs

Our paper includes further experiments for the planning method, learned representations, multi-step predictions, and more.



The plots show median performance and percentiles 5 to 95 over 5 seeds and 10 episodes each.

Contact: mail@danijar.com Twitter: @danijarh

Project website with videos: danijar.com/planet