Learning Unsupervised Latent Dynamics Models for Multi-task Continuous Control from Pixels

Danijar Hafner, Ian Fischer, Timothy Lillicrap, David Ha, James Davidson, Honglak Lee

Google AI DeepMind

1. Overview

Learning a dynamics model in principle provides a natural approach to multi-task RL.

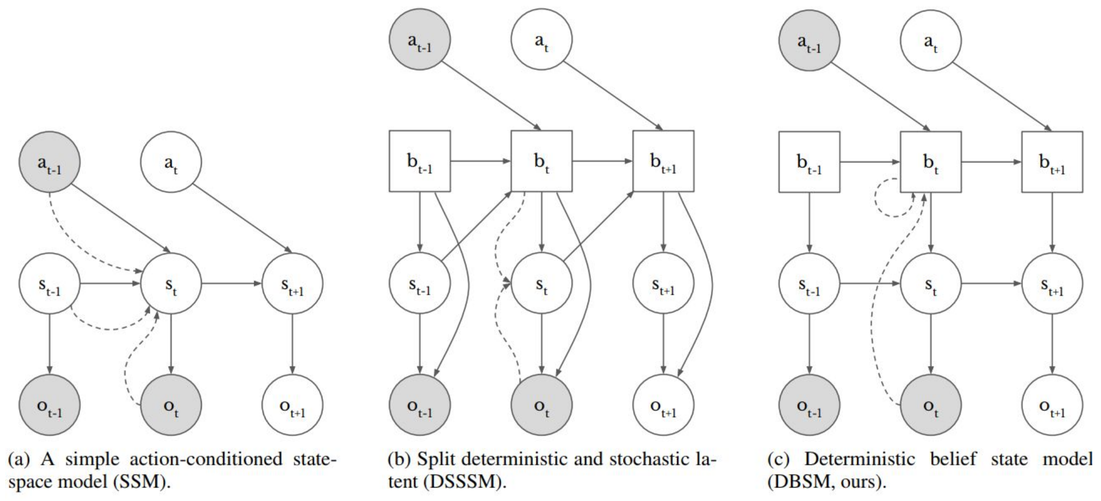
Training accurate models is an open problem.

Latent dynamics models enable fast planning and reduce accumulating errors in pixel domains.

Contributions:

- DBSM: Variational sequence model that avoids unnecessary sampling steps.
- Variational overshooting: Generalization of ELBO to train models on multi-step predictions.
- Multi task: Learn multiple reward predictors.
- Online learning: Gaits in a handful of episodes.

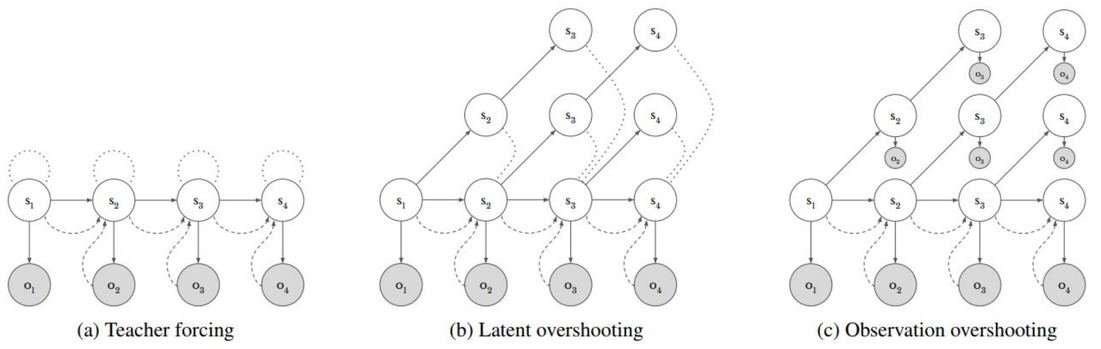
2. Deterministic Belief State Model



We maintain a deterministic belief that has access to the full history and can be predicted forward in time without sampling.

The beliefs guide sampled state sequences.

3. Variational Overshooting

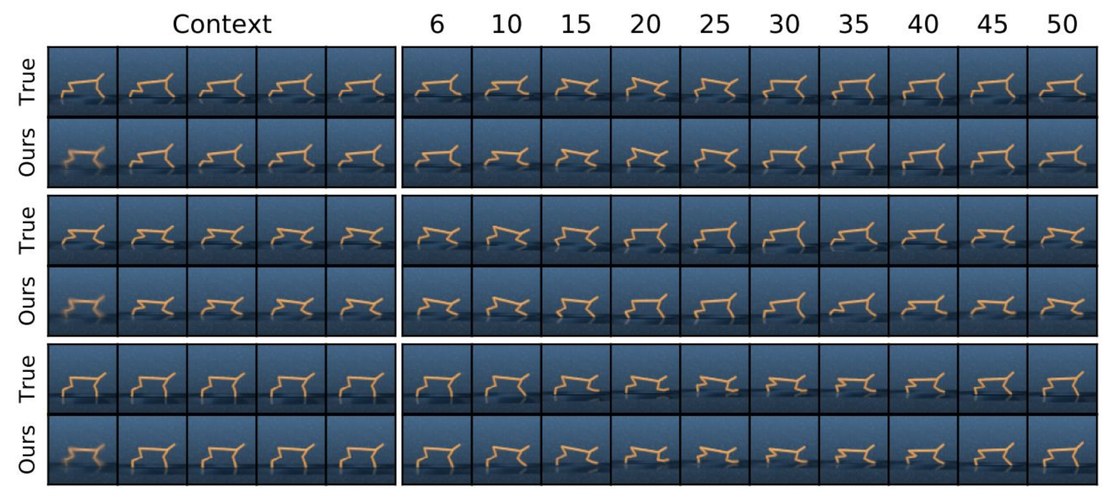


Normal ELBO for sequence models only trains the model to do 1-step predictions.

Our generalization trains the model on 1...H-step predictions, inspired by Amos et al. 2018.

This adds extra KL losses (latent overshooting) and likelihood losses (observation overshooting).

4. Accurate Video Prediction



DBSM infers velocities better due to deterministic access to the entire history.

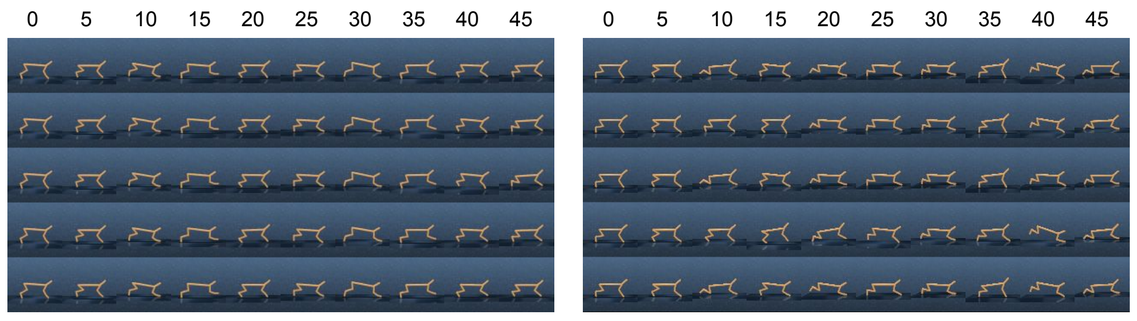
Latent overshooting allows accurate long-term predictions and is fast to compute.

5. Multi-task Learning

A large data set of experience without rewards is available to train the dynamics.

After convergence, train a reward predictor from the latent space for each task.

Use the cross-entropy method (CEM) for planning in latent space to solve the task.



Cheetah forward

Cheetah backward

6. Online Data Acquisition

Actively collect data using CEM with the partially trained dynamics and reward models.

Corrects model biases quickly because collected episodes serve as new training data.

Peak performance so far is comparable to A3C trained on 100k episodes from states. Our model was trained on 120 episodes from pixels.

Method	Modality	Control frequency	Episodes	Mean final score
DBSM + CEM (ours)	pixels	25 Hz	120	213±2.1
PPO	proprioceptive	25 Hz	10,000	248
PPO	proprioceptive	100 Hz	10,000	190
A3C	proprioceptive	100 Hz	100,000	214±1.6
D4PG	proprioceptive	100 Hz	100,000	737±4.4
D4PG	pixels	100 Hz	100,000	524±6.8