Planning from Images with Deep Latent Gaussian Process Dynamics

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Nathanael Bosch *, 1, 2, Jan Achterhold *, 1, Laura Leal-Taixé 2, Jörg Stückler 1

1 Embodied Vision Group, Max Planck Institute for Intelligent Systems, Tübingen, Germany   2 Technical University of Munich, Germany

Overview

Problem setup:

- Learn action-conditioned dynamics of a physical system given images as observations
- Use the learned dynamics to solve a control problem with image feedback (model-based RL)

Our approach:

- Convolutional neural networks for mapping between image space and latent space
- Gaussian process posteriors to model rewards and transitions in the latent space
- MPC with Cross-Entropy Method (CEM) for planning in latent space

Main advantage:

Quick adaptation to changes in environment dynamics without additional training

Contributions

Approach	Dynamics model	Representation model	Reward model	Planning algorithm
PILCO [4]	GP	Identity	Analytic	Policy search
Kalman-VAE [3]	Blended linear	VAE	-	-
PlaNet [2]	RSSM (GRU)	VAE	MLP	MPC/CEM
DLGPD (ours)	GP	VAE	GP	MPC/CEM

- Gaussian process models were shown to be sample efficient for learning control and were sucessfully applied to real-world systems [4]
- Our work joins the two fields of Gaussian processes for sequence modelling and learning control with representation learning techniques to map between image observations and a latent space (Variational Auto-Encoder)
- We show that our model can learn the dynamics of an inverted pendulum from image observations and swing the pendulum up with a model-predictive control algorithm (CEM)
- We demonstrate that the latent Gaussian process dynamics model allows the agent to adapt to environments with modified system dynamics from only a few rollouts and without additional training. Our approach compares favorably to the purely deep-learning based baseline PlaNet [2] in transfer learning experiments

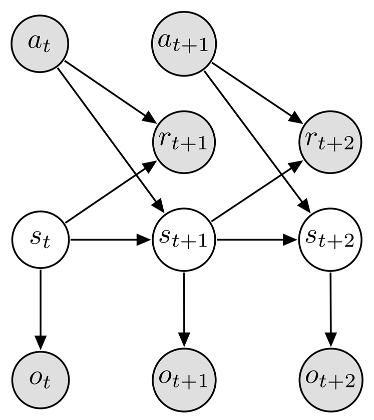
Problem Setup

Dynamical systems:

Stochastic dynamics given by

$$s_{t+1} = f(s_t, a_t) + \epsilon_s,$$
$$r_{t+1} = h(s_t, a_t) + \epsilon_r,$$
$$o_t = g(s_t) + \epsilon_o,$$

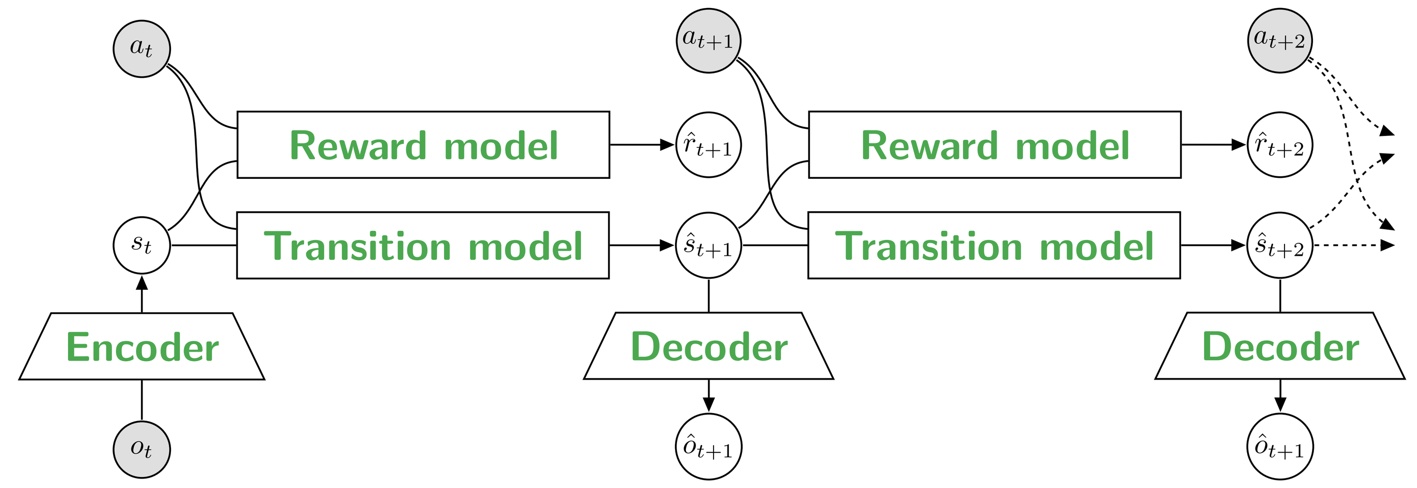
with latent states $s \in \mathbb{R}^D$, actions $a \in \mathbb{R}^K$, rewards $r_t \in \mathbb{R}$, and observations $o \in \mathbb{R}^M$.



Goals:

- Learn low-dimensional, action-conditioned dynamics in a latent space given high-dimensional observations (images)
- Implement a policy $p(a_t | o_{\leq t}, a_{<t})$ that maximizes the expected sum of rewards

Model



- **Transition model:** $f \sim \mathcal{GP}(\mu_f(\cdot), k_f(\cdot,\cdot))$, with mean function $\mu_f : (s_t, a_t) \mapsto s_t$ and radial basis function (RBF) kernel k_f
- **Reward model:** $h \sim \mathcal{GP}(r_{\min}, k_h(\cdot,\cdot))$ where $r_{\min}$ is the minimal reward observed in the collected training data and k_h the RBF kernel
- **Observation model (Decoder):** $p(o_t \mid s_t)$ an approximate Bernoulli with mean $\mathbb{E}_{p(o_t \mid s_t)}[o_t] = g(s_t)$, where $g(\cdot)$ is parametrized by a transposed-convolutional network
- **Encoder:** $q(s_t \mid o_t) \sim \mathcal{N}\left(s_t \mid \mu(o_t), \sigma(o_t)^2 \cdot I\right)$ with vector-valued $\mu(\cdot)$ and $\sigma(\cdot)$ parametrized by a convolutional neural network

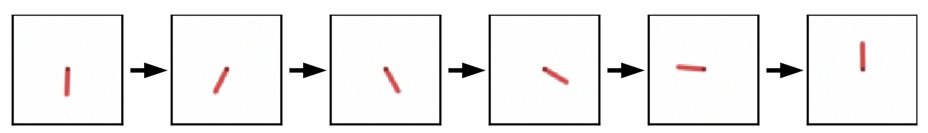
Training Objective

- **Notation:** Given data $\mathcal{D} = \{(o_t, a_t, o_{t+1}, r_{t+1})\}_{t=1}^{T}$, consisting of transitions in observation space, we define $O = \{o_1, \ldots, o_{T-1}\}$, $A = \{a_1, \ldots, a_{T-1}\}$, $O' = \{o_2, \ldots, o_T\}$, and $R' = \{r_2, \ldots, r_T\}$, together with latent states $S = \{s_1, \ldots, s_{T-1}\}$ and $S' = \{s_2, \ldots, s_T\}$.
- **Training objective:** Our joint training objective is to maximize a lower bound on the data log-likelihood:

$$\log p(O', R' \mid O, A) \geq \underbrace{\mathbb{E}_{q(S' \mid O')}\left[\log p(O' \mid S')\right]}_{\text{(I): Reconstruction}} + \underbrace{\mathbb{E}_{q(S' \mid O')}\left[-\log q(S' \mid O')\right]}_{\text{(II): Encoder regularization}}$$
$$+ \underbrace{\mathbb{E}_{q(S' \mid O')q(S \mid O)}\left[\log p(S' \mid S, A)\right]}_{\text{(III): State transitions}} + \underbrace{\mathbb{E}_{q(S \mid O)}\left[\log p(R' \mid S, A)\right]}_{\text{(IV): Reward}}$$

Experiment Setup

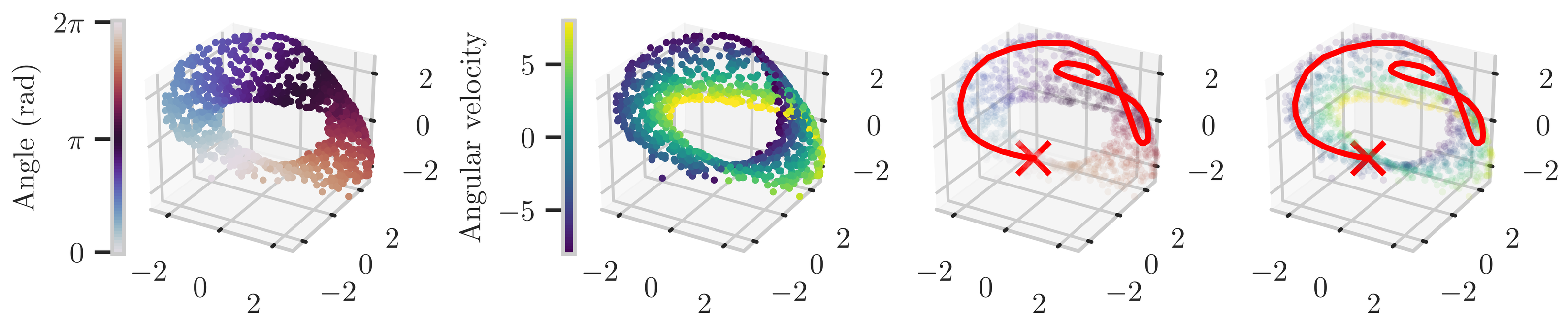
- **Task:** Inverted pendulum swingup (*OpenAI Gym Pendulum-v0*)



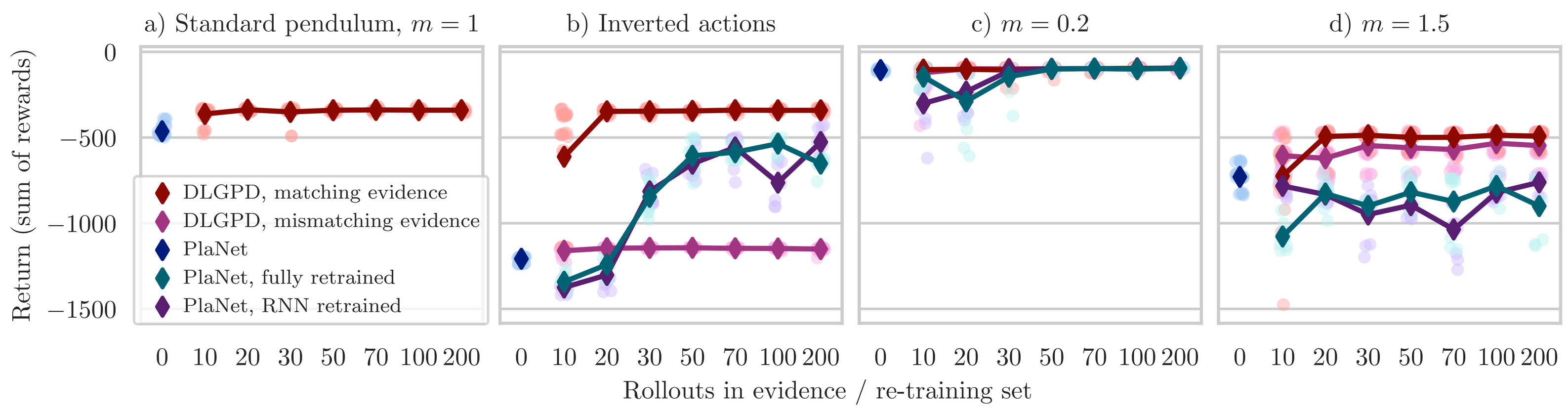
- **Training Data:** Rollouts obtained by applying random actions, 500 for training and 200 for evidence; we choose rollouts from the latter to condition the GPs on.
- **Transfer Learning:** Evaluate model performance on unmodified environment (a) and for the following modifications:
  (b) Inverted actions, (c) Reduced pole mass ($m = 0.2$), (d) Increased pole mass ($m = 1.5$)
  We *condition* the model (more precisely the GPs) on data from these environments
  No additional training is required!
- **Comparison:** PlaNet [2] trained on the training and evidence data (500+200 rollouts; sufficient to achieve good performance on the standard task). For the transfer learning evaluation the model is fully or partially retrained.

Results

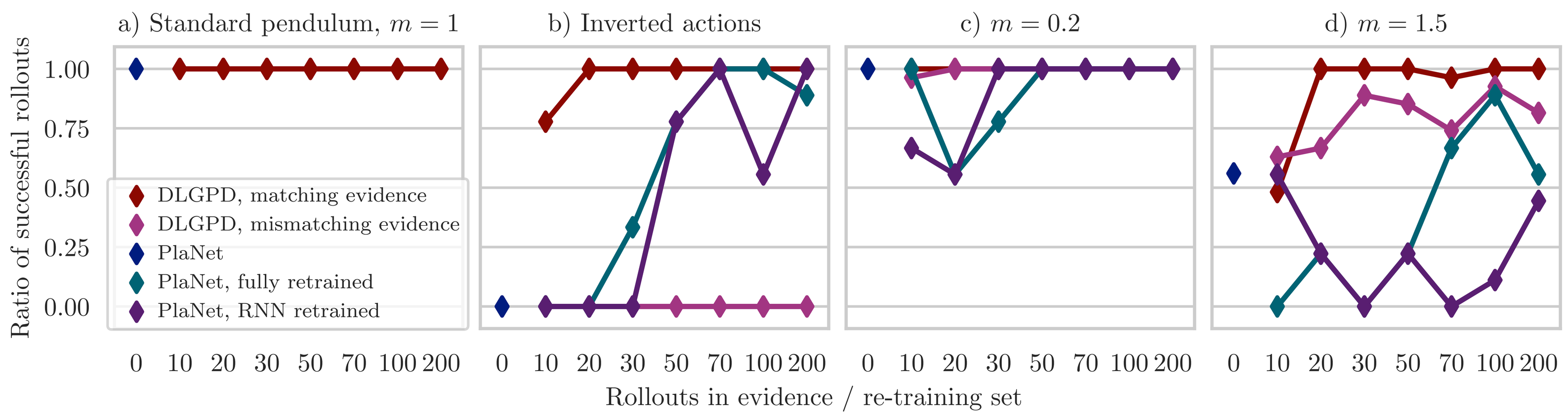
Latent space and planned trajectories



Cumulative rewards



Success rate



Evaluation:

- Structured latent space that allows for good planning
- Good performance on the unmodified environment (a)
- Data-efficient transfer to modified environments (b)-(d):
  - 20 rollouts are enough for (nearly) 100% success rate in all tasks
  - In comparison, PlaNet [2] requires significantly more data to achieve comparable success rates and reaches lower cumulative rewards

Paper and supplementary material available at:
https://dlgpd.is.tue.mpg.de

Acknowledgements:
This work has been supported through the Max Planck Society and Cyber Valley. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Jan Achterhold.

References:

[1] Nathanael Bosch, Jan Achterhold, Laura Leal-Taixe, and Jörg Stückler. Planning from Images with Deep Latent Gaussian Process Dynamics. In *2nd Annual Conference on Learning for Dynamics and Control (L4DC)*, 2020.

[2] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 2555–2565, 2019.

[3] Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. In *Advances in Neural Information Processing Systems, NIPS 2017*, pages 3601–3610, 2017.

[4] Marc Deisenroth and Carl E. Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2011*, pages 465–472, 2011.

nathanael.bosch@uni-tuebingen.com, jan.achterhold@tuebingen.mpg.de

dlgpd.is.tuebingen.mpg.de