Model Predictive Curiosity for Self-Supervised Dynamics Models

Michael Lingelbach
Stanford University
mjlbach@stanford.edu

Olivia Y. Lee
Stanford University
oliviayl@stanford.edu

Priscilla Zhao
Stanford University
puzhao@stanford.edu

1 Motivation

Children develop internal models about physical [1] and social [2] dynamics to learn how to interact with the world. A crucial element of child psychological development is play (with objects like blocks, toys, etc.), which is an active, intrinsically-motivated activity through which children learn about the physical dynamics of objects and environments around them through interaction and observation [3]. Therefore, developing an auxiliary reward that encourages exploration in an unconstrained setting is the computational analog of this process, where the agent explores an environment and takes actions that are not conditioned on any particular end-goal. We hope to see how active exploration and self-supervised learning facilitates the learning process of reinforcement learning agents.

Our study is interested in modeling this process of active exploration and self-directed learning, so as to develop artificial intelligence (AI) agents capable of learning flexibly and robustly in a similar way. We present **Model Predictive Curiosity (MPCu)**, a framework inspired by the Model Predictive Control (MPC) paradigm which maps state-action pairs to curiosity values, thereby predicting the error in the forward prediction model. Our model then backpropagates from the predicted curiosity value to select the action which would maximize this value.

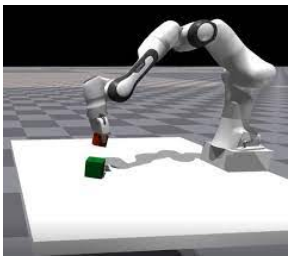
2 Related Work

Prior work has investigated curiosity-driven, closed-loop system that learns forward object dynamics self-supervised and without any human and studying curiosity with regards to physical object manipulation [4] [5]. However, one limitation in this line of previous work is the inefficient sampling-based mechanism for action selection.

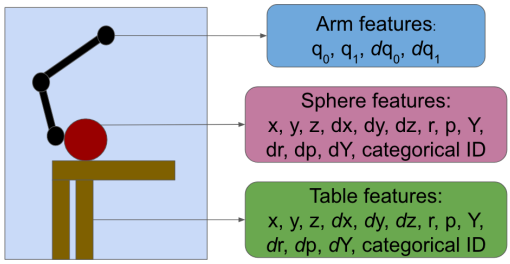
The primary advantage of the MPC approach is that it requires minimal human involvement and can learn in an entirely self-supervised fashion, without a detailed reward function, goal image, or ground truth object pose information. In the canonical instantiation of the MPC algorithm [6], the algorithm enables a robotic agent to plan for actions that move a user-specified object to a user-defined location. It then evaluates the candidate action sequence and chooses the action that maximizes the distribution over the designated pixel's position. As a result, the agent's actions are continuously re-planned as the agent executes the task moves to new states. This work uses a convolutional LSTM to predict future camera images and image-space pixel flow. However, low resolution images may not have the best object representation, which limits the predictive power of the algorithm. Our method draws inspiration from this study and aims to develop a more efficient way of selecting actions to train the forward model.

Our study also draws inspiration from prior research in visual interaction networks [7]. In this study, a visual interaction network is used to generate future trajectories of objects in a physical system from video frames of the system. The network consists of three main components: the visual encoder, which is a convolutional neural network that produces a state code from a sequence of three images, the dynamics predictor, which takes a sequence of state codes and predicts the candidate state code for the next frame, and the state decoder, which converts a state code to a state.

3 Methods



(a) Sample IsaacGym environment with robotic arm, tabletop, and hard block.



(b) Simplified environment: 2-DoF robotic arm, tabletop, and sphere

.

Figure 1: (a) IsaacGym and (b) simplified environment illustrations

Our proposed dynamics model primarily involves object-centric embeddings. We aimed to use NVIDIA's IsaacGym physical simulation engine for an embodied environment comprising of a robotic arm on a tabletop with a variety of potential objects, illustrated in Figure 1a above. Figure 1b illustrates a simplified version of our environment for the purposes of establishing notation, consisting of a robotic arm with two degrees of freedom on a tabletop with a single spherical object. For testing, we use a simplified Box2D environment, a 2-Dimensional physics engine for games. The robot arm features primarily pertain to the robot arm's joint states: each q_i represents the joint state of joint i, and ∂q_i represent the angular velocities of each joint. In Figure 1b, the robotic arm with two degrees of freedom has two joints, with states represented by q_0 and q_1 and angular velocities ∂q_0 and ∂q_1 respectively. For all other objects (e.g. the table, spherical, etc. objects), every object is defined by their kinematic features involving its position in 3-D space (x, y, z), linear velocity (∂x, ∂y, ∂z), rotation (roll r, pitch p, yaw Y), rotation of angular velocity (∂r, ∂p, ∂Y), and categorical ID (which contains object-specific shape information, as our model does not use parameterized shapes).

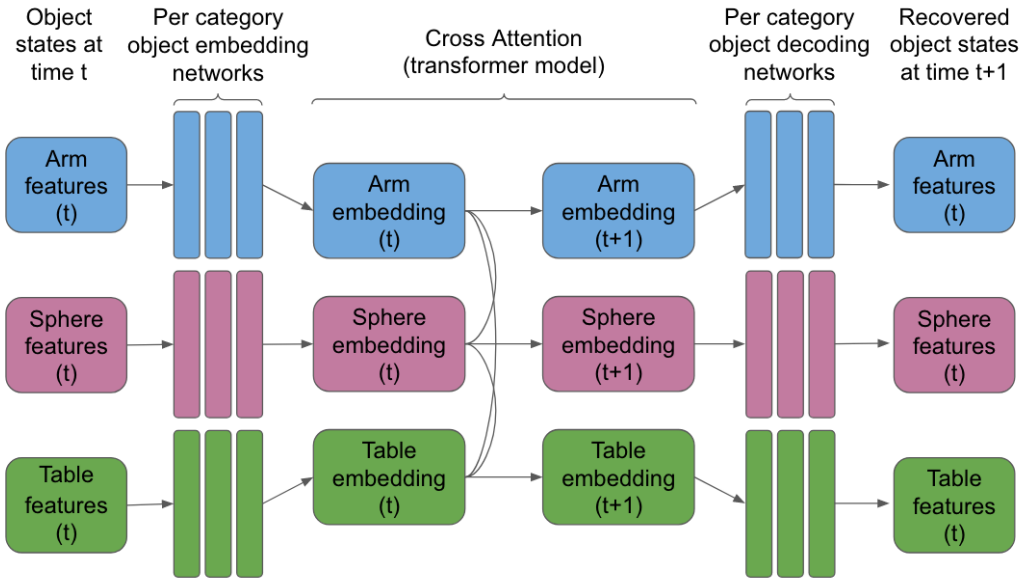
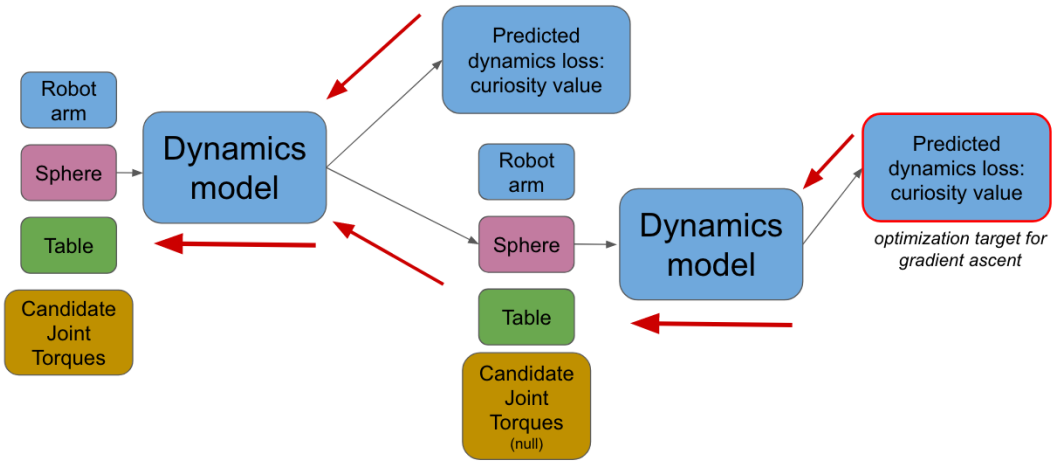
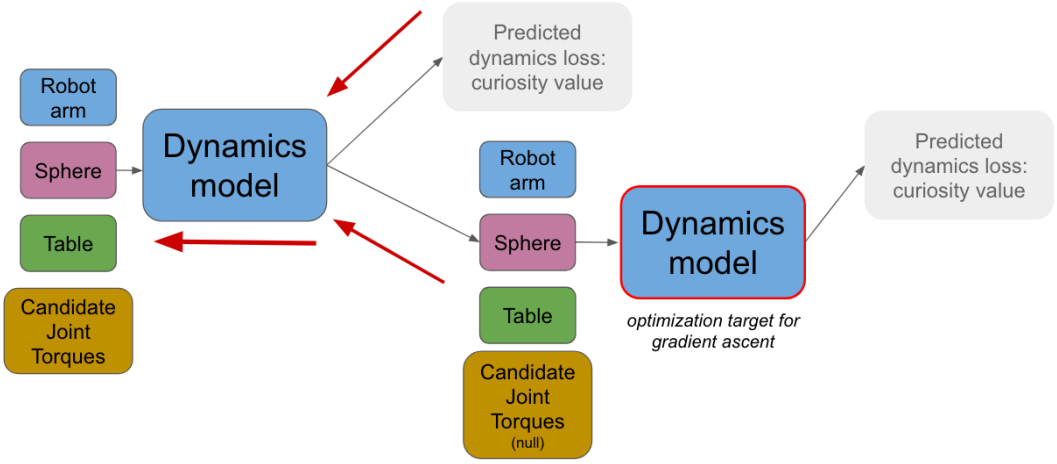


Figure 2: Proposed dynamics model using sequence-to-sequence based forward prediction on object-centric embeddings

Using this object-centric embedding framework, our dynamics model engages in sequence-to-sequence based forward prediction, inspired by sequence-to-sequence language models (Figure 2). The object states at time t (as defined in Figure 1b) are fed into embedding networks specific to each category of object, generating the object embeddings corresponding to the provided states. These object embeddings are then fed into a cross-attention transformer model, which outputs the predicted object embeddings at the next timestep $t + 1$. The predicted embeddings are then fed into category object-specific decoding networks to recover the object states at time $t + 1$.



(a) MPCu: sequence-to-sequence-based curiosity estimation



(b) Sequence-to-sequence based model predictive control

.

Figure 3: Proposed methods for (a) MPCu and (b) traditional MPC

Our proposed method is Model Predictive Curiosity (MPCu) using sequence-to-sequence-based curiosity estimation (Figure 3a). By sampling several rollouts from the dynamics model, we can compute expected loss over multiple timesteps. We can them maximize the predicted loss in the dynamics model using the curiosity value as the optimization target for gradient ascent. Like model predictive control, we predict both the future world state at the next timestep $t + 1$, along with the predicted loss as measured by the L2 pose estimation error. From the prediction target, we perform gradient ascent via backpropagation through the dynamics model into the candidate joint torques, in order to choose a set of joint torques which maximally antagonize the dynamics model.

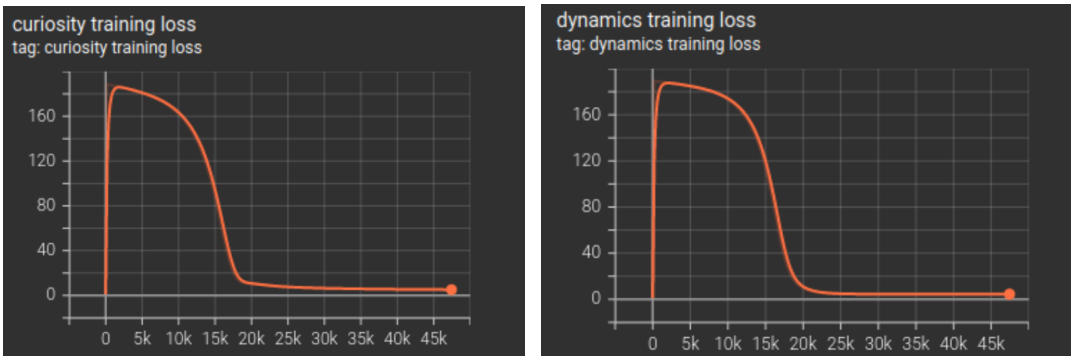
We can also directly leverage our dynamics model for control by performing traditional model predictive control, as shown in Figure 3b. In this mode, we ignore the forward pass through the curiosity model and use the desired world state as the optimization target for gradient ascent, specifically to minimize the delta between desired world state and predicted world state.

In sum, we train our dynamics model using the curiosity value function as our initial optimization target for gradient ascent. We then exploit the trained model at test time to adopt the standard model predictive control paradigm, differentiating back into candidate joint torques to minimize the divergence between the desired and predicted world states.

4 Results & Discussion

We use the Box2D environment to test a simplified version of our proposed model, rendered using PyGame. In a Box2D environment, we generated 50,000 training scenarios of a force being applied to a circle adjacent to a tower of three squares. We then trained a dynamics model to predict the forward motion of circle based on a single timestep, and a curiosity-based model to predict the loss in the aforementioned dynamics model.

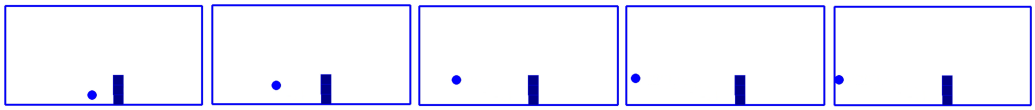
At test time, we performed a backward pass through the curiosity-based model into the action space, and performed gradient ascent on the force vector that is applied to the circle to select the force (i.e. action on the circle) that maximizes this curiosity value. Note that this force vector is distinct from the origin of the force, which is always fixed to the center of the circle.



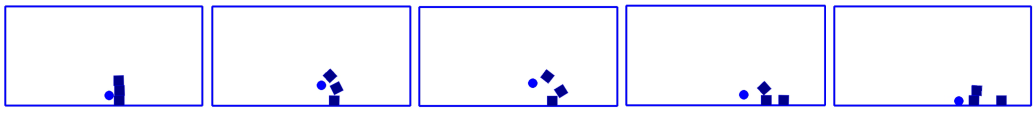
(a) Curiosity Model Loss

(b) Dynamics Model Loss

Figure 4: Loss curves for (a) curiosity model and (b) dynamics model



(a) Pre-MPCu Optimization Trajectory



(b) Post-MPCu Optimization Trajectory

Figure 5: Trajectories based on rollouts (a) pre-MPCu and (b) post-MPCu

We observe that the training losses for both the curiosity-based model and the dynamics model converge at $\sim$ 20,000 steps, as seen in Figure 4a and 4b respectively. This demonstrates the success of the MPCu approach in learning successive prediction models for the forward motion of the circle and loss of the dynamics model.

We can also engage in a more qualitative assessment of the model's performance by comparing the forward motion of the circle before and after optimization of the circle's trajectory with MPCu. We see that unlike the motion of the circle away from the tower of squares without interacting with them (Figure 5a), the circle proceeds to "knock over" the tower of three squares after training with MPCu in a manner consistent with our traditional understanding of physical dynamics (Figure 5b). Therefore, we see that the forward motion of the circle is able to optimize for high curiosity action values, enhancing exploration through interactions between objects.

Hence, we see that in this simplified Box2D environment, MPCu enriches vector forces that cause multi-object interactions. Furthermore, the performance of MPCu depends on three main factors: the predictive power of the forward dynamics model, the accuracy of the curiosity-based prediction network, and the inductive bias for action selection which initializes actions on the objects.

5 Conclusion & Future Work

We demonstrate that our proposed MPCu approach succeeds in facilitating exploratory multi-object interactions in a 2D environment. We implement this approach by first training a forward dynamics model and curiosity model to predict the motion of a primary subject and the loss in the dynamics model respectively, then backpropagating from the predicted curiosity value to select actions that maximize this value.

Given these promising observations in the simplified Box2D environments, a natural next step is to extend these experiments to an embodied 3D environment using NVIDIA's IsaacGym physics simulation engine. This will allow us to add complexity to our model as the embodied 3D environment has a higher dimensional action space (as opposed to a 2D action space of force vectors) for a robot arm with multiple degrees of freedom, as well as a larger variety of objects with different physical properties. This makes both forward prediction and backpropagation into the action space more complex, and will be an important step to seeing if MPCu can be applied to areas such as robotics that engage with high dimensional action spaces.

In addition, the performance of the system is hindered by the choice of dynamics model, which remains fairly simple in the Box2D environment. Future work in developing a robust dynamics model for MPCu, for instance by incorporating sequence-to-sequence based forward prediction as outlined in Figure 2, can further enhance the performance of the MPCu approach.

Finally, another extension to this project can involve incorporating longer horizon rollouts into the action selection phase. The current approach uses a single timestep, which is appropriate given the frame rate of the simplified environment. However, especially for more complicated environments such as embodied 3D environments, action selection over trajectories and not just based on one timestep will be important for optimizing action selection.

Overall, our proposed curiosity-based MPCu approach shows promising effects in facilitating exploration of physical object dynamics in a 2D environment, and results in interesting qualitative behavior in multi-object interactions in a 2D environment. Future research should focus on scaling this approach to incorporate more complex predictive and action-selection models, which in turn facilitates its applicability to more complex environments with higher dimensional action spaces.

References

[1] Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

[2] Jessica Sommerville, Amanda Woodward, and Amy Needham. Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, 96:B1–11, 06 2005.

[3] Alison Gopnik, Andrew Meltzoff, and Patricia Kuhl. *The Scientist in the Crib: Minds, Brains and How Children Learn*, volume 189. 03 2001.

[4] Michael John Lingelbach, Damian Mrowca, Nick Haber, Li Fei-Fei, and Daniel L. K. Yamins. Towards curiosity-driven learning of physical dynamics. In *Bridging AI and Cognitive Science*. International Conference on Learning Representations, 2020.

[5] Geroge Kachergis, Samaher Radwan, Bria Long, Judith Fan, Michael Lingelbach, Daniel M. Bear, Daniel L. K. Yamins, and Michael C. Frank. Predicting children's and adults' preferences in physical interactions via physics simulation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43. UC Merced.

[6] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. *CoRR*, abs/1610.00696, 2016.

[7] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.