UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Emergence of Structured Behaviors from Curiosity-Based Intrinsic Motivation

Permalink
https://escholarship.org/uc/item/0hk971rx

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors
Haber, Nick
Mrowca, Damian
Fei-Fei, Li
et al.

Publication Date
2018

eScholarship.org Powered by the California Digital Library
University of California

Emergence of Structured Behaviors from Curiosity-Based Intrinsic Motivation

Nick Haber, Damian Mrowca, Li Fei-Fei, Daniel L. K. Yamins
([nhaber, mrowca, feifeili, yamins]@stanford.edu)
Department of Psychology and Computer Science, Stanford University, Stanford, CA 94305, USA

Abstract

Infants are experts at playing, with an amazing ability to generate novel structured behaviors in unstructured environments that lack clear extrinsic reward signals. We seek to replicate some of these abilities with a neural network that implements curiosity-driven intrinsic motivation. Using a simple but ecologically naturalistic simulated environment in which the agent can move and interact with objects it sees, the agent learns a world model predicting the dynamic consequences of its actions. Simultaneously, the agent learns to take actions that adversarially challenge the developing world model, pushing the agent to explore novel and informative interactions with its environment. We demonstrate that this policy leads to the self-supervised emergence of a spectrum of complex behaviors, including ego motion prediction, object attention, and object gathering. Moreover, the world model that the agent learns supports improved performance on object dynamics prediction and localization tasks. Our results are a proof-of-principle that computational models of intrinsic motivation might account for key features of developmental visuomotor learning in infants. **Keywords:** Development learning, Curiosity, Neural Network Models

Introduction

Within the first year of life, humans exhibit a wide range of interesting, apparently spontaneous, visuomotor behaviors — including navigating their environment, seeking out and attending to objects, and engaging physically with these objects in novel and surprising ways (Fantz, 1964; Twomey & Westermann, 2017; Hurley et al., 2010; Hurley & Oakes, 2015; Goupil et al., 2016; Begus et al., 2014; Gopnik et al., 2009). In short, young children are excellent at playing, and their ability to make sense of and (re)structure their environments sets them apart from even the most advanced autonomous robots. Play capacity in this period likely interacts with infants' powerful abilities to understand and model their environment. By six months of age or younger, infants can orient themselves in a complex environment, account for the presence, number and visual properties of objects they interact with, and have a sense of how these objects behave dynamically Spelke (1985); Stahl & Feigenson (2015); Baillargeon (2007).

But how exactly do such young children know how to play? And how do such behaviors relate to their world model building abilities? One natural idea is that infants' world modeling capacities are the result of built-in core systems, including those for e.g. object attention and permanence, self-localization, number sense, and intuitive physics (Spelke & Kinzler, 2007). Once operational, such systems would naturally give the infant a basis on which to make judgments about which sequences of actions would be interesting to perform.

A related but alternative idea is that the intrinsic motivation of curiosity can itself drive the development of world model making (Schmidhuber, 2010). This idea relies on a virtuous cycle in which, by seeking out novel but replicable interactions, the child pushes the boundaries of what its world model-prediction systems can achieve, giving itself useful data on which to improve and develop these systems. As world modeling capacity improves, what used to be novel becomes old hat, and the cycle starts again. Related to the conception of the "scientist in the crib" (Gopnik et al., 2009), in this account, aside from being fun, play behaviors may be extremely useful and highly structured, driving the self-supervised learning of a variety of representations underlying sensory judgments and motor planning capacities (Mitash et al., 2017).

Figure 1: Problem setting. How do agents learn from object interactions while moving around in the physical world?

Building on recent work in artificial intelligence (Schmidhuber, 2010; Pathak et al., 2017; Kulkarni et al., 2016; Jaderberg et al., 2016; Oudeyer et al., 2007) we make a computational model of an agent driven by curiosity-based intrinsic motivation — an agent acting to maximize an intrinsic reward derived from a measure of what its internal model of the world finds, by some operationalization, *interesting*, rather than an extrinsic reward. Related are approaches that attempt to learn "options" useful for exploration (Machado et al., 2017) and those that learn a predictive coding amidst random goal setting and imitation (Najnin & Banerjee, 2017). We present a simple simulated interactive environment in which an agent can move around and physically act on objects it sees (Fig. 1). In this world, interesting interactions are sparse unless actively sought after. We then describe a neural network architecture through which the agent learns a world model that seeks to predict the consequences of its own actions. In addition, as the agent optimizes the accuracy of its world model, a separate neural network simultaneously learns an agent action policy that seeks to take actions that adversarially challenge the current state of its world model. We demonstrate that this architecture stably engages in the virtuous reinforcement learning cycle described above, spontaneously learning to understand self-generated ego motion and to selectively pay attention to, localize, and interact with objects, without having to have any of

these concepts built in. This learning occurs through emergent self-curricularization in which new capacities arise at distinct "developmental milestones". Unlike previous work, we show the learned representation transfers to improved performance on analogs of real-world visual tasks such as object localization and detection. To our knowledge, a self-supervised setup in which an explicitly self-modeling agent uses intrinsic motivation to learn about and restructure its environment has not been explored prior to this work.

Agent Architecture and Environment

We place an agent in a physically realistic simulated *environment* built in the Unity 3D simulation framework. The agent consists of a world model and a loss model. The *world model* is tasked to learn dynamics from visuomotor inputs. The *loss model* tries to estimate the world model's losses several time steps into the future to choose actions that antagonize the world model's learning. Our self-supervised curiosity system is depicted in Figure 2. We emphasize that we do not initialize our model with pretrained weights so as to explore what world representation and behaviors emerge from this simple antagonistic setup in a physically embodied environment.

Interaction environment

Our environment consists of a simple square room in which an agent and several objects are initially placed randomly. The agent is modeled as an invisible sphere that can move around and, in discrete time steps, receives an RGB image from its forward-facing camera. In order to model interaction with objects requiring some attention and proximity, the agent can apply forces and torques in all three dimensions to objects that are both in view and within a fixed distance δ. We refer to a state in which the agent can act on the object as a *play state*, and the object as *in play*. Although the remaining environment is static, the agent and objects can collide with every part.

We define a state in the state space $s_t \in S$ to consist of the images captured at times t and $t-1$ by the agent. In state s_t, the agent specifies an action $a_t \in A$, which leads to the next state s_{t+1}. The action space $A \in \mathbb{R}^{2+6N}$ is continuous. The first 2 dimensions specify ego motion, restricting agent movement to forward/backward motion v_{fwd} and horizontal planar rotation v_θ. The remaining $6N$ dimensions specify the forces f_x, f_y, f_z and torques $\tau_x, \tau_y \tau_z$ applied to N objects sorted from the lower-leftmost to the upper-rightmost object relative to the agent's field of view. This representation is unambiguous as objects can only be acted on when in view, and if $k < N$ objects are in play, the environment accepts all $2+6N$-tuples but only applies the leftmost k force-torque pairs. All coordinates are bounded by constants and normalized to 1.

World model

Given a slice of history $h_t = (s_{t-k_b}, a_{t-k_b} \dots s_{t+k_f}, a_{t+k_f}) \in H$, we can describe a *generalized dynamical problem* by an input map $\xi : H \to X$ and a true-value map $\eta : H \to Y$ and require the world model (blue in Figure 2) map $\xi(h)$ to $\eta(h)$ — regardless of whether this is well-defined. Let ω denote this world model, so that $\omega(\xi(h)) \in Y$. For each prediction, a loss $L_{wm}(\omega(\xi(h)), \eta(h))$ is incurred. Note that both the input and true-value maps are determined exclusively by the agent's experience and require no outside processing or labels. In practice, we find *inverse dynamical prediction* useful — filling in a missing action, instead of predicting the future — and concretely in our case, $H = \{(a_{t-2}, s_{t-1}, a_{t-1}, s_t, a_t, s_{t+1}, a_{t+1}, s_{t+2})\}$ with $\xi(h)$ excluding a_t and $\eta(h) = \zeta(a_t)$. Here ζ zeros out the force and torque components not corresponding to objects in play, as they have no observable effect, and bins the action into classes, with L_{wm} the softmax cross-entropy loss. We bin each dimension by $x < -0.1, -0.1 \leq x \leq 0.1$, and $x > 0.1$.

We train a convolutional neural network ω_θ for this task from scratch with stochastic gradient descent, with randomly initialized parameters θ. We use twelve convolutional layers, with two-stride max pools every other layer and one hidden layer to encode states into a latent space with shared weights. The latent states $\{\lambda_{t+i}\}$ concatenated with the given actions $\{a_{t+j}\}$ are then input to a two-layer MLP to predict a_t.

Loss model

The agent's goal is to antagonize the world model, so if it could predict the loss incurred at future time steps as a function of its options, a policy could be made. In practice, we do this explicitly, except predicting only a discretization of the loss for ease of training. Given s_t and a proposed next action a, the loss model predicts, for $u = 1 \dots T$ timesteps into the future, probability distributions $p_u(c|s_t, a)$ over discrete (via thresholding) classes of loss $c \in C_l$. It is penalized with softmax cross-entropy loss, averaged over timesteps. We use a separate convolutional neural network Λ_ψ with parameters ψ, with twelve convolutional layers with two-stride max pools every other layer and one hidden layer to encode the state which is then concatenated with the proposed a_t. This representation is then used as input to a two-hidden-layer MLP to infer the prediction. All future losses aside from the first one depend not only on the state of the world model but also on future actions taken. The loss model hence needs to predict in expectation over future policy. The loss predictions are usefully interpreted as *loss prediction maps* $\Lambda_{s_t}(a)$ on action space given a current state s_t as depicted in Figure 4.

Action policy

Given the loss prediction model, the agent can use a simple mechanism to choose its actions. Given a real-valued function σ of these T probability distributions, we can define our policy $\pi(a|s_t)$ as a distribution

$$\pi(a|s_t) \sim \exp(\beta\sigma(p_1, \dots p_T)(a, s_t)), \tag{1}$$

with hyperparameter β. We take σ to be a sum over expectation values in loss class $\sigma(a)[s_t] = \sum_i \sum_{c \in C} c p_i(c|a, s_t)$, though other methods (Feigelis et al., 2018) have proven useful in some contexts. We evaluate σ for $K = 1000$ uniform random samples in A. We then sample from a K-way discrete distribution with probabilities proportional to equation (1).

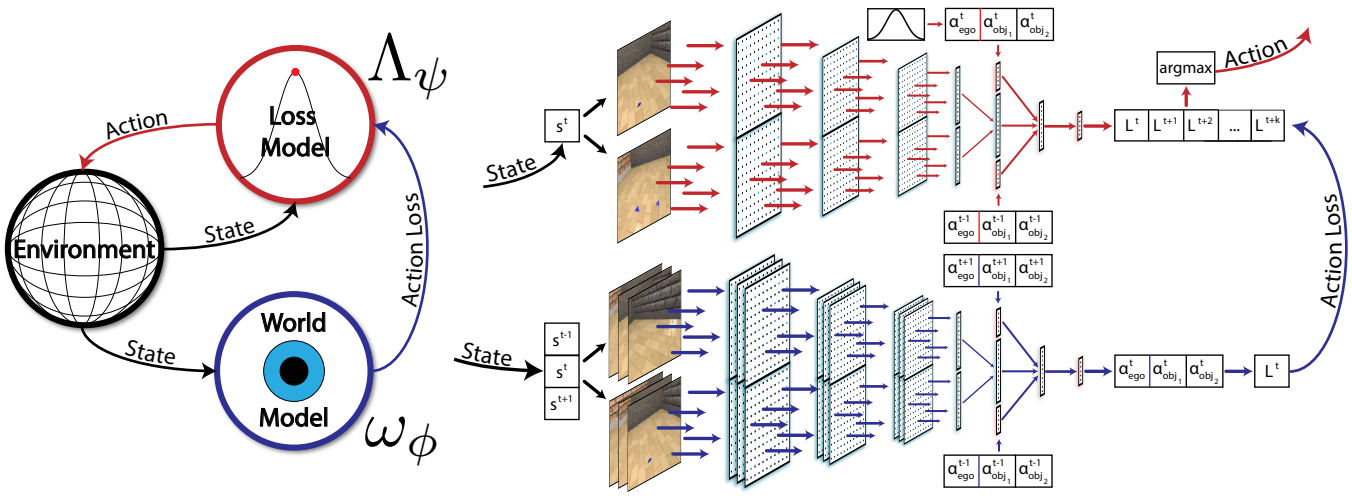


Figure 2: Self-supervised curiosity model. We train a dynamical world model (blue), while simultaneously learning a loss model (red) that predicts the world model's loss to choose actions that lead to novel and surprising events in the environment (black). Models both consist of convolutional neural networks and are fed one or more states consisting of two successive images from the agent's camera.

Experiments

We randomly situate the agent and up to two objects in a square 10x10 unit room with play distance $\delta = 2$. The agent trains on 16 blue objects with different shapes, i.e. cones, cylinders, cuboids, pyramids, and spheres of varied aspect ratios. We reinitialize the play scene every $8,000$ to $30,000$ steps.

We first place the agent with one object in the room and show that it learns to predict ego motion, and attend, localize and navigate towards objects by evaluating the world model's training loss, the agent's play state frequency, and the inverse dynamics prediction performance on a fixed validation set. In a second experiment we increase the number of objects to two, and demonstrate that the agent learns to gather both objects and further prefers to play with two objects over one object by looking at the frequency of 1 and 2 object play states and object-agent distances. These behaviors emerge in a specific order akin to developmental milestones.

We compare our learned world model with curious policy with $T = 40$ (LW/CP-40) against a baseline with a world model with fixed random weights following a random policy (RW/RP) and a baseline where the world model weights are learned with a random policy (LW/RP).

Ego motion learning

Figure 3 (a) shows the training loss curves of LW/CP-40 and the baselines. RW/RP does not learn well since most of its weights are fixed to be random. LW/RP quickly converges to a low value as it learns from a constant random distribution without an antagonistic policy. The LW/CP-40 loss dips before increasing as the loss model first needs to learn which actions lead to higher loss before being able to antagonize the world model effectively. This first dip in loss corresponds to the world model learning ego motion. The ego motion error reported in Table 1 is close to the error reached at this point.

Emergence of object attention

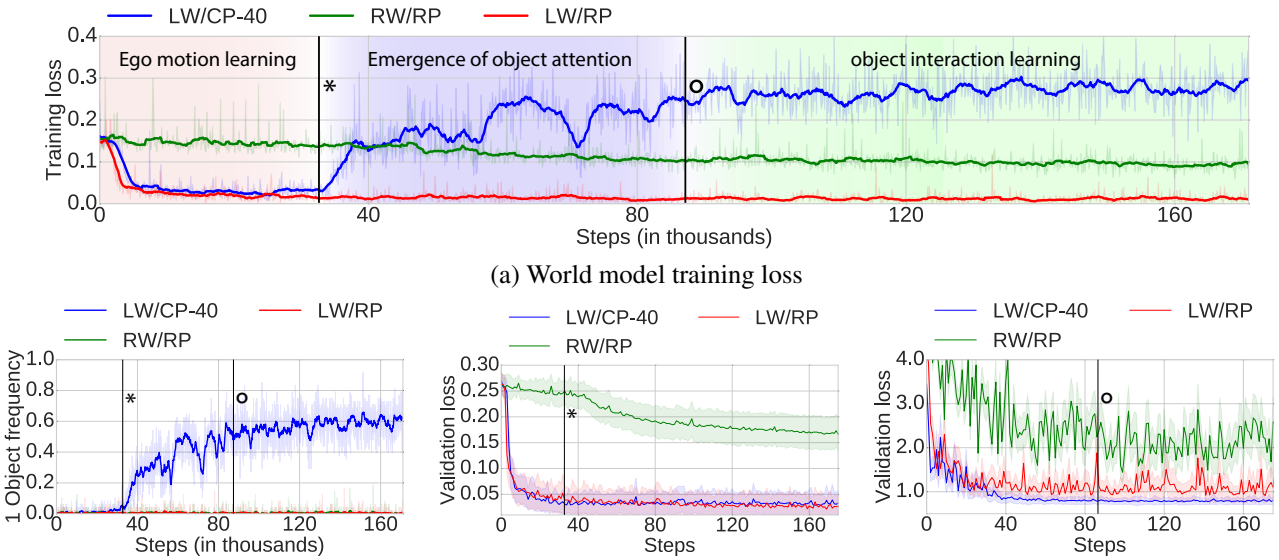
As the LW/CP-40 loss increases after the initial ego motion dip the agent starts to attend to objects which is reflected in an increase of object interactions as shown in Figure 3 (b). At the final stage, the agent interacts with the object about 60% of the time which indicates that it learns to localize and attend to the object (Table 1). At the same time, the increasing world model loss shows that these object interactions are much harder to predict than ego motion. The baselines almost never interact with the object and thus experience lower ego motion losses.

Improved inverse dynamics prediction

We evaluate the inverse dynamics prediction performance on a held out validation set gathered from the environment while following a random action policy. To measure the models' performance on predicting ego motion and object actions separately, we divide the validation set into two sets. The first set contains all examples in which the agent is in a play state. The second set consists of all remaining examples. As we can see in Figure 3 (c) and in Table 1, LW/CP-40 and LW/RP perform well on predicting ego motion as no antagonistic policy is necessary to encounter ego motion. However, our policy outperforms the baselines on predicting object interactions by a significant margin showing that focusing on object interactions does indeed improve inverse dynamics prediction performance as seen in Figure 3 (d) and Table 1.

Improved object detection and localization

To quantify the world model's object presence and localization performance, we train a linear regression/logistic regression with elastic net regularization on features from various world model layers on a dataset generated by gathering data online while following a random action policy. Half of object presence training data contains an object. For localization, the second image is guaranteed to contain the object. Both



(a) World model training loss

(b) 1 object play frequencies

(c) Ego motion validation loss

(d) 1 object interaction validation loss

Figure 3: 1 object experiments. Learned world model with curious policy (LW/CP-40) is compared against learning a world model following a random policy (LW/RP) and a world model with random weights with a random policy (RW/RP). (a) World model cross-entropy loss during training. (b) Object play state frequency in %. (c) Ego motion prediction cross-entropy loss on held out validation set. (d) 1 object interaction prediction cross-entropy loss on held out validation set.

Table 1: Performance comparison. Learned world model with a curious policy (LW/CP-40) is compared against learned world model with random policy (LW/RP) and random world model with random policy (RW/RP). Ego motion (v_{fwd}, v_θ) and interaction (f, τ) accuracy in % is compared for play and non-play states. Object frequency and presence are measured in % and localization in mean pixel error.

Task	RW/RP	LW/RP	LW/CP-40
Play v_{fwd} accuracy	61.8	84.5	**93.6**
Play v_θ accuracy	78.8	95.2	**97.9**
Play f, τ accuracy	20.8	39.8	**44.4**
Non-play v_{fwd} accuracy	59.2	**94.2**	92.3
Non-play v_θ accuracy	75.0	**97.9**	**97.9**
Object frequency	0.2	0.4	**59.0**
Presence error	3.9	1.2	**0.6**
Localization error [px]	15.14	5.99	**4.80**

training datasets consists of 16,000 image pairs labeled with the object's presence or pixel-wise 2d position of its centroid respectively. The respective validation and test sets comprise 8,000 image pairs each. As can be seen in Table 1 our model outperforms the baselines on the object presence and localization task, indicating that it learns better visual features.

Navigation and planning

In addition to object localization, the agent also exhibits navigation and planning abilities. In Figure 4 we give visualizations of loss maps projected onto the agent's position at the respective time. The *loss prediction maps* are generated by uniformly sampling 1000 actions a from the action space A, evaluating $\Lambda_{s_t}(a)$ and applying a postprocessing smoothing algorithm. We truncate the figure at five out of the 40 time steps our loss model predicts. The loss maps show the agent predicting a higher loss (red) for actions moving it towards the object to reach a play state. Consequently, our curious policy will take actions that navigate the agent closer to the object.

Emergence of multi-object interactions

At the beginning of the training of the 2 object experiment we observe similar stages as for the 1 object experiment (Figure 5 (a)). The loss dips as the agent learns to predict its ego motion and rises when its attention shifts towards objects which it then interacts with. This is followed by a another loss increase which corresponds to the agent gathering and playing with both objects simultaneously. This is reflected in an increase in 2 object play time (Figure 5 (c)) over 1 object play time (Figure 5 (b)). Consequently, the average distance between the agent and the objects decreases over time as seen in Figure 5 (c). It drops to about 2 units, the maximum interaction distance. The LW/RP baseline quickly drops and flattens out. LW/CP-20 with $T = 20$ instead of $T = 40$ learns to interact with one object but not with two objects simultaneously.

Discussion and Future Work

We observe that a simple and general intrinsic motivation mechanism based on adversarially antagonizing the loss of a dynamically constructed model of the world allows an agent to stably generate a spectrum of emergent naturalistic behaviors. Through self-curricularization in an active learning Settles (2011) process the agent achieves several "developmental milestones" of suitably increasing complexity as it learns to "play". Starting with random actions, it quickly learns the dynamics of its own ego motion. Then, without being given an explicit

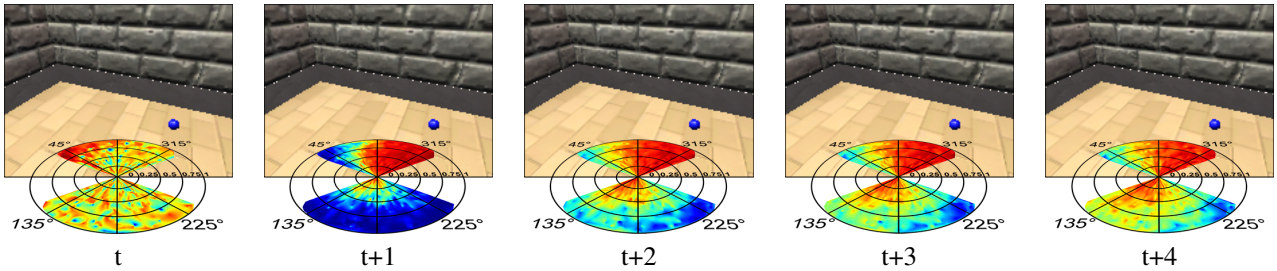
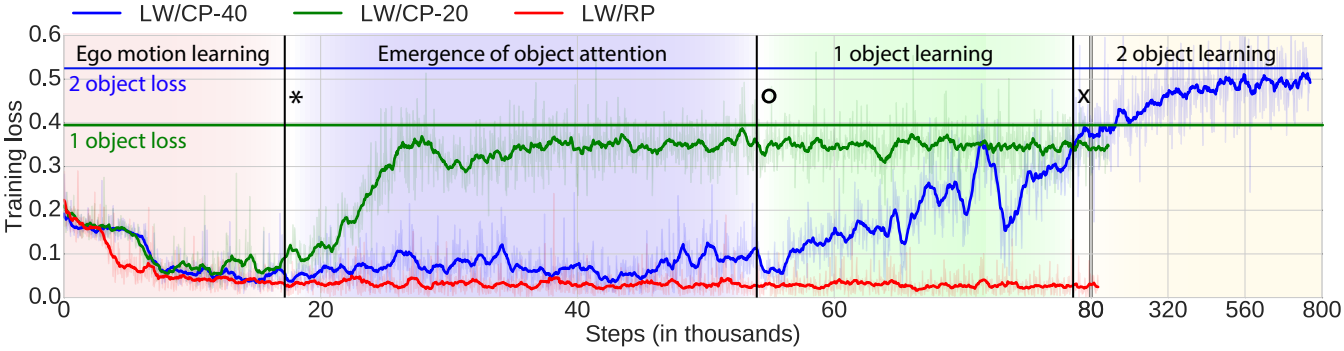
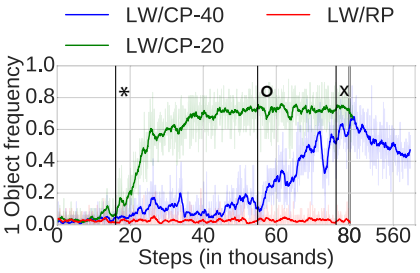


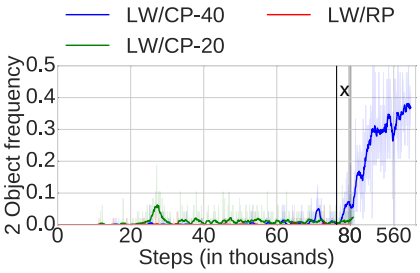
Figure 4: Navigation and planning behavior. The loss model predicts higher loss if the agent turns towards the object. Red colors correspond to high and blue colors to low loss predictions. The center of the heat map corresponds to the agents position.



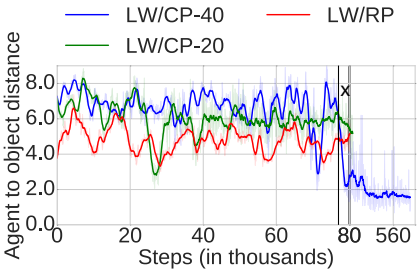
(a) World model training loss



(b) 1 object play frequencies



(c) 2 object play frequencies



(d) Object-agent distances

Figure 5: 2 object experiments. Learned world model with curious policy (LW/CP-40) is compared against the same setup but with $T = 20$ (LW/CP-20) and a learned world model following a random policy (LW/RP). (a) World model cross-entropy loss during training. (b) Object play state frequency for 1 object in %. (c) Object play state frequency for 2 objects in %. (d) Average distance between agent and objects in Unity units.

supervision signal as to the presence or location of an object, it discards ego motion prediction as boring and begins to focus its attention on objects, which are more interesting. Lastly, when multiple objects are available, it gathers the objects so as to interact with both. This policy leads to performance gains in object dynamics prediction as well as other tasks which the system was not explicitly learning.

This occurs without any pretrained visual backbone — the visual system world model was intentionally not initialized with filter weights pretrained on (e.g.) ImageNet classification. This constitutes partial progress in replacing the training of a visual backbone through a task such as large-scale image classification with an interactive self-supervised task and is a proof-of-concept that more complex milestones can be potentially reached while developing an understanding of object categories and physical relations.

This combination of spontaneous behavior leading to an improved world model is well suited to designing agents that must act effectively in the many real-world reinforcement learning scenarios in which rewards are sparse or potentially unknown. Here, we ultimately seek to develop algorithms that will control autonomous robots that learn to operate in complex unpredictable environments. From a cognitive science perspective, these results suggest a route toward using intrinsically motivated learning systems to model emergence of spontaneous behavior in young children. In this domain, we seek to make computational models that describe key aspects of real infant learning.

In order to have "developmental milestones" that may be mapped onto real child development, a variety of limitations will need to be overcome. First, to make the connection to cognitive science more realistic, our environment and agent need to be more realistic. Better graphics and physics, with more interesting visual objects, will be important to allow

better transfer the learned behavior to real-world visuomotor interactions. It will also be important to create a properly embodied agent with visible arms and tactile feedback, allowing for more realistic interactions. In our current work, we fail to address the fact that infants have severely limited mobility and motor control. A realistic model of this is critical to make realistic predictions for actual infant development. Including other animate agents would lead to more complex interactions and the possibility of learning through imitation.

Second, our current approach suffers from *degeneracy*: that a well-behaved enough map from inputs X to outputs Y may not exist. The inverse dynamical prediction problem exhibits this: when given an image sequence in which an object rests on the ground, action is under-determined: the agent could have been pushing down on the object, or not. Extending beyond our environment and to other dynamical problems, the existence of unpredictable stimuli, referred to as the *white noise problem* Schmidhuber (2010), further contributes to degeneracy, and our method can easily break down in this setting as the agent seeks out experience it cannot learn from. Several approaches to address this have been proposed Schmidhuber (2010); Pathak et al. (2017), and TODO cite Oudeyer Pathak et al. (2017) have demonstrated some success, it remains to be seen whether such methods lead to gains in environments that approach ecological realism. Taking these next steps will not only help us to understand how infants learn, but also to develop AI systems that learn without human supervision.

Acknowledgements

This work was supported by an Understanding Human Cognition award from the James S. McDonnell foundation (DLKY), a Simons Collaboration on the Global Brain grant (DLKY), a Berry Foundation postdoctoral fellowship (NH), and by hardware donation from the NVIDIA Corporation.

References

Baillargeon, R. (2007). The acquisition of physical knowledge in infancy: A summary in eight lessons. In *Blackwell handbook of childhood cognitive development* (pp. 47–83). Blackwell Publishers Ltd.

Begus, K., Gliga, T., & Southgate, V. (2014, 10). Infants learn what they want to learn: Responding to infant pointing leads to superior learning. *PLOS ONE*, *9*(10), 1-4.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668-670.

Feigelis, K. T., Sheffer, B., & Yamins, D. L. K. (2018). Modular continual learning in a unified visual environment. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=rkPLzgZAZ`

Gopnik, A., Meltzoff, A., & Kuhl, P. (2009). *The scientist in the crib: Minds, brains, and how children learn*. Harper-Collins.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they dont know. *Proceedings of the National Academy of Sciences*, *113*(13).

Hurley, K. B., Kovack-Lesh, K. A., & Oakes, L. M. (2010, Dec). The influence of pets on infants' processing of cat and dog images. *Infant Behav Dev*, *33*(4), 619–628.

Hurley, K. B., & Oakes, L. M. (2015, Jan). Experience and distribution of attention: Pet exposure and infants' scanning of animal images. *J Cogn Dev*, *16*(1), 11–30.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., & Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., & Tenenbaum, J. B. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *CoRR*, *abs/1604.06057*.

Machado, M. C., Bellemare, M. G., & Bowling, M. (2017). A laplacian framework for option discovery in reinforcement learning. *arXiv preprint arXiv:1703.00956*.

Mitash, C., E. Bekris, K., & Boularias, A. (2017, 03). A self-supervised learning system for object detection using physics simulation and multi-view pose estimation.

Najnin, S., & Banerjee, B. (2017). A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Communication*, *92*, 24 - 41. Retrieved from `http://www.sciencedirect.com/science/article/pii/S016763931630139X` doi: https://doi.org/10.1016/j.specom.2017.05.002

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, *11*(2), 265–286.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363*.

Schmidhuber, J. (2010, Sept). Formal theory of creativity, fun, and intrinsic motivation (1990 – 2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230-247.

Settles, B. (2011). *Active learning* (Vol. 18). Morgan & Claypool Publishers.

Spelke, E. S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Stahl, A. E., & Feigenson, L. (2015, Apr). Cognitive development. Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Twomey, K. E., & Westermann, G. (2017, Oct). Curiosity-based learning in infants: a neurocomputational approach. *Dev Sci*.