A Compositional Object-Based Approach to Learning Physical Dynamics

Michael Chang[*], **Tomer Ullman**[**], **Antonio Torralba**[*], and **Joshua B. Tenenbaum**[**]
[*]Department of Electrical Engineering and Computer Science, MIT
[**]Department of Brain and Cognitive Sciences, MIT
{mbchang,tomeru,torralba,jbt}@mit.edu

Abstract

This paper presents the Neural Physics Engine (NPE), an object-based neural network architecture for learning predictive models of intuitive physics. The NPE draws on the strengths of both symbolic and neural approaches: like a symbolic physics engine, it is endowed with generic notions of objects and their interactions, but as a neural network it can also be trained via stochastic gradient descent to adapt to specific object properties and dynamics of different worlds. We evaluate the efficacy of our approach on simple rigid body dynamics in two-dimensional worlds of bouncing balls. By comparing to less structured architectures, we show that the NPE's compositional representation of the causal structure in physical interactions improves its ability to predict movement, generalize to different numbers of objects, and infer latent properties of objects such as mass.

1 Introduction

A sense of intuitive physics can be seen as a program [7] that takes in input provided by a physical scene and the past states of objects and then outputs the future states and physical properties of relevant objects for a given task. At least two general approaches have emerged in the search for such a program that captures common-sense physical reasoning. The top-down approach [3, 16, 17] formulates the problem as inference over the parameters of a symbolic physics engine, while the bottom-up approach [1, 5, 8–11, 13] learns to directly map physical observations to motion prediction or physical judgments. A program under the top-down approach can express and generalize across any scenario supported by the entities and operators in its description language. However, it may be brittle under scenarios not supported by its description language, and adapting to these new scenarios requires modifying the code or generating new code for the physics engine itself. In contrast, the same model architecture and learning algorithm under gradient-based bottom-up approaches can be applied to new scenarios without requiring the physical dynamics of the scenario to be pre-specified. However, such models require extensive amounts of data, and oftentimes transferring knowledge to new scenes requires retraining, even in cases that seem trivial to human reasoning.

This paper takes a step toward bridging this gap between expressivity and adaptability by proposing a model that combines rough symbolic structure with gradient-based learning. We present the Neural Physics Engine (NPE), a predictive model of physical dynamics. It exhibits several strong inductive biases that are explicitly present in symbolic physics engines, such as a notion of objects and object interactions. It is also end-to-end differentiable and thus is also flexible to tailor itself to the specific object properties and dynamics of a given world through training. This approach – starting with a general sketch of a program and filling in the specifics – is similar to ideas presented by [12, 14]. The NPE's general sketch is the structure of its architecture, and it extends and enriches this sketch to model the specifics of a particular scene by training on observed trajectories from that scene.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Neural Physics Engine

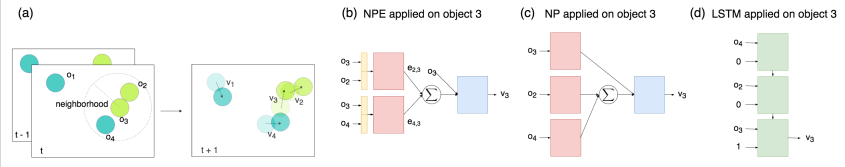

Figure 1: **Scenario and Models:** This figure compares the NPE, the NP and the LSTM architectures in predicting the velocity of object 3 for an example scenario [**a**] of two heavy balls (cyan) and two light balls (yellow-green). Objects 2 and 4 are in object 3's neighborhood, so object 1 is ignored. [**b**]: The NPE encoder comprises a pairwise layer (yellow) and a feedforward network (red) and its decoder (blue) is also a feedforward network. [**c**]: The NP encoder is the same as the NPE encoder, but without the pairwise layer. The NP decoder is the same as the NPE decoder. [**d**]: We shuffle the context objects inputted into the LSTM and use a binary flag to indicate whether an object is a context or focus object.

This paper's contribution links two levels of factorization and composition in learning physical dynamics. On the level of the physical scene, we factorize the scene into object-based representations (Fig 1a), and compose smaller building blocks to form larger objects (Fig 2c). This framework of representation adapts to complex scenes and configurations with variable object count. On the level of the physics program, the NPE architecture explicitly reflects a causal structure in object interactions by factorizing object dynamics into pairwise interactions. As a predictive model of physical dynamics, the NPE models the future velocity $v_f^{[t+1]}$ of a single *focus* object f as a function composition of the pairwise interactions between itself and other neighboring *context* objects c_k in the scene. This structure guides learning towards object-based reasoning, and by design allows physical knowledge to transfer across variable number of objects and for object properties to be explicitly inferred.

The input is represented as pairs of object state vectors $\{(o_f, o_{c_1})^{[t-1,t]}, (o_f, o_{c_2})^{[t-1,t]}, ...\}$. A state vector comprises extrinsic properties (position, velocity, orientation, angular velocity), intrinsic properties (mass, object type, object size), and global properties (gravitational, frictional, and pairwise forces). The NPE also predicts angular velocity along with velocity, but for our experiments we always set angular velocity, as well as gravity, friction, and pairwise forces, to zero. As shown in Fig. 1b, the NPE is a composition of an encoder function f_{enc} that summarizes the interaction of a single object pair and a decoder function that takes the sum of encodings of all pairs to predict the velocity.

How f_{enc} and f_{dec} are composed emulates the high-level formulation of many symbolic physics engines. We provide a loose interpretation of the encoder output $e_{f,c}$ as the *effect* of object c on object f, and require that these effects are additive as forces are, allowing the NPE to scale naturally to different numbers of neighboring context objects. These inductive biases have the effect of strongly constraining the space of possible programs of predictive models that the NPE can learn, focusing on compositional programs that reflect pairwise causal structure in object interactions.

We compared the NPE to two baselines (Fig. 1c,d). The No-Pairwise (NP) baseline is a Markovian variant of the Social LSTM presented by [2]; it sums the encodings of context objects after encoding each object independently. It most directly highlights the value of the NPE's pairwise factorization. Because it moves through the object space sequentially, the LSTM baseline's lack of factorized compositional structure highlights the value of the NPE's function composition of the independent interactions between an object and its neighbors.

3 Evaluation

Using the matter-js physics engine [4], we evaluate the NPE on worlds of bouncing balls. These worlds exhibit self-evident dynamics and support a wide set of scenarios that reflect everyday physics. Bouncing balls have been of interest in cognitive science to study causality and counterfactual reasoning, as in [6]. We trained on 3-timestep windows in trajectories of 60 timesteps (10 timesteps $\approx$ 1 second) using rmsprop [15] with a Euclidean loss. Experimental results on held-out test data are summarized in Fig. 2. Randomly selected simulation videos are at `https://drive.google.com/drive/folders/0BxCJLi4FnT_6QW4tcF94d1doLWs?usp=sharing`.

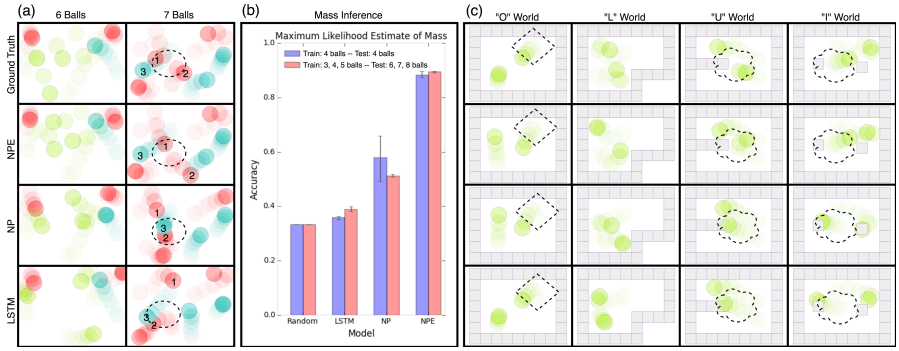

Figure 2: **Results:** The NPE scales to complex dynamics and world configurations while the NP and LSTM cannot. The masses are visualized as: cyan = 25, red = 5, yellow-green = 1. [**a**] Consider the collision in the 7-balls world (circled). In the ground truth, the collision happens between balls 1 and 2, and the NPE correctly predicts this. The NP predicts a slower movement for ball 1, so ball 2 overlaps with ball 3. The LSTM predicts a slower movement and incorrect angle off the world boundary, so ball 2 overlaps with ball 3. [**b**]: In mass inference, the NPE notably performs similarly well whether in a world it has seen before or in a world with a number of objects it hasn't trained on, further showcasing its strong generalization capabilities. [**c**] At first glance, all models seem to handle collisions well in the "O" world (diamond), but when there are internal obstacles (cloud), only the NPE can successfully resolve collisions. This suggests that the NPE pairwise factorization handles object interactions well, letting it generalize to different world configurations, whereas the NP and LSTM have only memorized the geometry of the "O" world.

Generalization and knowledge transfer We test whether learned knowledge of these simple physics concepts can be transferred to worlds with a number of objects previously unseen. We train on worlds with 3, 4, and 5 balls and test on more complex worlds with 6, 7, and 8 balls, all with equal mass. As shown in Fig. 2a, the NPE exhibits much cleaner extrapolation to worlds with more objects. The NPE's performance of this generalization task suggests that its architectural inductive biases are useful for generalizing knowledge learned in Markovian domains with causal structure in object interactions.

Mass inference We show that the NPE infers latent properties such as mass (Fig. 2b). We uniformly sampled the mass for each ball from the log-spaced set $\{1, 5, 25\}$. For evaluation, we select scenarios exhibiting collisions with the focus object, fix the masses of all objects except that of the focus object, and score the NPE's prediction under all possible mass hypotheses for the focus object. The prediction is scored against the ground-truth under the same Euclidean loss used in training. The mass hypothesis whose prediction yielded the lowest error is the NPE's maximum likelihood estimate of the mass for the focus object. The NPE achieves about 90% accuracy, while a random model would guess the correct mass with 33% accuracy.

Different scene configurations We demonstrate representing large structures as a composition of smaller objects as building blocks in a world of balls and obstacles. These worlds contain 2 balls bouncing around in variations of 4 different wall geometries. "O" and "L" geometries have no internal obstacles and are in the shape of a rectangle and "L" respectively. "U" and "I" have internal obstacles. Obstacles in "U" are linearly attached to the wall like a protrusion, while obstacles in "I" have no constraint in position. We randomly vary the position and orientation of the "L" concavity and the "U" protrusion. We randomly sample the positions of the "I" internal obstacles.

We train on the conceptually simpler "O" and "L" worlds and test on the more complex "U" and "I" worlds (Fig. 2c). Variations in wall geometries adds to the difficulty of this extrapolation task. However, our state space representation was designed to be flexible to this variation, by representing walls as composed of uniformly-sized obstacles, similarly to how many real-world objects are composed of smaller components. At most 12 context objects are present in the focus object's neighborhood at a time. The "U" geometries have 33 objects in the scene, the most out of all the wall geometries. Using such a compositional representation of the scene allows the NPE to scale to different configurations, which would not be straightforward to do without such a representation.

4 Discussion

We have demonstrated a compositional object-based approach to learning physical dynamics in worlds of bouncing balls in several tasks ranging in complexity. Further work includes generalization to unseen object types and physical laws such as in worlds with immovable obstacles and stacked block towers. Because the NPE is differentiable, we expect that by backpropagating prediction error to its input, it may be able to infer the positions of "invisible" objects, whose effects are felt but whose position is unknown. Our results invite questions on how much prior information and structure should and could be given to bottom-up neural networks, and what can be learned without inducing such structure. It would be interesting to explore how similar models to the NPE can be used as subprograms that can be called by parent programs to evolve entity states through time for applications in areas such as model-based planning and model-based reinforcement learning.

References

[1] P. Agrawal, A. Nair, P. Abbeel, J. Malik, and S. Levine. Learning to poke by poking: Experiential learning of intuitive physics. *arXiv preprint arXiv:1606.07419*, 2016.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces.

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

[4] L. Brummitt. http://brm.io/matter-js. URL `http://brm.io/matter-js`.

[5] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik. Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*, 2015.

[6] T. Gerstenberg, N. Goodman, D. A. Lagnado, and J. B. Tenenbaum. Noisy newtons: Unifying process and dependency accounts of causal attribution. In *In proceedings of the 34th*. Citeseer, 2012.

[7] N. D. Goodman and J. B. Tenenbaum. Probabilistic models of cognition, 2016. URL `http://probmods.org`.

[8] A. Lerer, S. Gross, R. Fergus, and J. Malik. Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*, 2016.

[9] W. Li, S. Azimi, A. Leonardis, and M. Fritz. To fall or not to fall: A visual approach to physical stability prediction. *arXiv preprint arXiv:1604.00066*, 2016.

[10] R. Mottaghi, H. Bagherinezhad, M. Rastegari, and A. Farhadi. Newtonian image understanding: Unfolding the dynamics of objects in static images. *arXiv preprint arXiv:1511.04048*, 2015.

[11] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi. " what happens if..." learning to predict the effect of forces in images. *arXiv preprint arXiv:1603.05600*, 2016.

[12] A. Solar-Lezama. *Program synthesis by sketching*. ProQuest, 2008.

[13] I. Sutskever, G. E. Hinton, and G. W. Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2009.

[14] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[15] T. Tieleman and G. Hinton. Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. 2012.

[16] T. Ullman, A. Stuhlmüller, and N. Goodman. Learning physics from dynamical scenes. 2014.

[17] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*, pages 127–135, 2015.