Learning Intuitive Physics by Explaining Surprise

Hung Nguyen* Jay Patravali* Fuxin Li Alan Fern
Oregon State University
(nguyehu5, patravaj, fuxin.li, alan.fern) @oregonstate.edu

Abstract

The IntPhys Challenge aims to evaluate how well algorithms capture "common sense" about the physical world by measuring the ability to detect violations of intuitive physics in dynamic multi-object visual scenes. One approach to this problem is to define or learn a detailed model of the observations and dynamics and to then detect violations of that model. While viable, this approach poses challenges in acquiring an accurate enough model that can handle detailed non-linear object interactions, such as visual occlusion and collisions. In this work, we consider an alternative approach, the Surprise and Explain (SnE) framework, which aims for simplicity while remaining highly flexible. The key idea is to exploit the assumption that, for the vast majority of time, objects follow simple dynamic models, e.g. linear dynamics. Further, when the simple dynamics are occasionally violated ("surprises") due to non-linear interactions, e.g. collisions and occlusion, it is assumed that there is a small set of detectable explanations for the surprise. Violations of intuitive physics then correspond to surprises for which an explanation cannot be inferred. This paper develops an instantiation of the SnE framework and demonstrates its potential in the IntPhys Challenge by placing 2^{nd} at the time of this paper's submission. [1]

1. Introduction

It is intuitive for human beings to distinguish between real and unreal physical scenarios. Studies on human infants [12] have shown that key to human-like intelligence is the capability to identify objects, reason about shape and physical dynamics – fundamental building blocks towards more complex human tasks. For machines this common sense has yet to be achieved, despite rapid progress in artificial intelligence over the last decade in fundamental tasks such as object classification, detection [9, 6] or neural control [8].

In this paper, we present a novel and efficient 'Surprise and Explain' framework for parsing natural physical object-object interactions that comprises of three modules: perception, dynamics and explanation. The perception module identifies and localizes the objects in a given visual scene. The dynamics module encodes the object state and velocity and predicts future states. The explanation module uses the current observation and past states to provide an explanation for plausibility or implausibility of the current state.

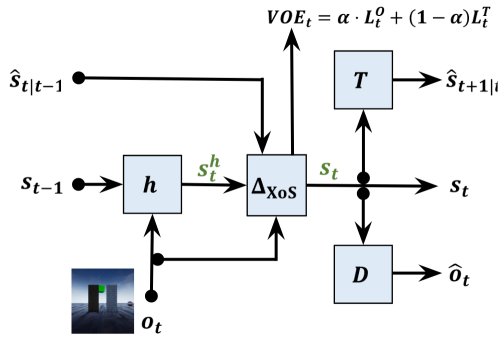

Figure 1. The Surprise and Explain (SnE) framework

Intphys challenge 2019 [10] examined systems for three basic concepts of the physics of macroscopic solid objects: object permanence, shape constancy, spatio-temporal continuity. Each of these concepts are tested in a series of controlled possible and impossible clips, which are presented without labels, and for which models have to return a plausibility score. We apply simple models to this challenge and achieved top performance in it.

2. Related Work

Some of the early works model physics of natural scenes through simulation engines [2, 15]. Later, it was shown that these engines could be made 'learnable' through deep neural networks [1]. Object-centric representation learning for scene understanding is quite prominent in literature: [14] builds on [1] to jointly learning perception with physics, whereas [3, 13] look at learning dynamics of interactions through decomposition of visual scene into object's repre-

[1] Intphys Challenge 2019 Leaderboard. * Equal Contribution

sentation and its state properties.

Another line of work aims at extending the goals of perception modules from knowing where an object is, to also infer and predicting its future states. Visual imagination introduced in [5] equips an agent with the ability to generate potential future states of the world in response to an action without actually performing that action. Two unsupervised learning methods are presented in [10] for future frame predictions: a GAN-based approach and encoder-decoder network. Probabilistic modelling provides robustness to uncertainty in visual inputs such as occlusions or disappearance [7]. Therefore, [4] obtains reliable long-term predictions by outputting a distribution of outcomes while letting the model implicitly learn the underlying physics. Likewise, [11] arrive at robust predictions of future through probabilistic simulations and particle filtering.

3. Method

The goal of our system is to solve the anomaly detection problem where the model has to not only detect impossible scenarios but also localize and identify the objects that are involved in the implausible event.

- One approach is to learn a discriminative model to classify whether an event is possible or not given the current state of the environment. Learning such a model requires a large amount of data of both classes (i.e possible vs impossible), the fact that is not suitable in the context of this challenge where we are only provided videos of plausible events.

- Another approach is to build a generative model whose distribution ideally represents all possible events. We can generate a set of possible events and then match the current observation with those events. If there is a good match, this observation can be considered plausible. Otherwise, we regard the observation as an anomaly. A drawback is that the number of possible events can grow exponentially, depending on the complexity of the environment.

Intuitively, when observing the world, we humans do not think of every possible scenarios that can happen. Instead, we implicitly form some simple events or so-called theories in our minds. If there is something surprising that does not conform with our thought or theories, we then try to find some explanations based on our prior knowledge for the current observation. This idea motivates us to build a framework coined as Surprise and Explain (SnE). This framework is especially efficient when the number of possible violations is small, comparing to the number of possible events. Our current framework deals with 4 types of violation including object permanence, energy conservation, space-time discontinuity, and shape consistency.

Figure 1 describes our SnE framework. o_t and $\hat{o}_t$ are the observation at time t and the observation prediction. s_t^h, s_t, and $\hat{s}_{t+1|t}$ are the nominal latent state, latent state, and the prediction of s_{t+1} when at time t respectively. L_t^T and L_t^O are the transition and the observation likelihood. **h**, **T**, and Δ_{XoS} are the state encoder, transition function, and correction function respectively.

Assume that a video includes N objects. Each object O^n, where $n = 1...N$ is the object's index, maintains a track, which is a set of states up to frame t denoted by $S^n(t) = (s_1^n, s_2^n, ...s_t^n)$. From now on, we use s_t instead of s_t^n for short. The state at each frame s_t includes the current position p_t (i.e bounding box), the temporal velocity v_t, the current observation o_t, the occlusion status z_t, and the average depth d_t of the object.

3.1. Object Tracking

Because violations in a video of this challenge is always related to objects, it is important that our framework is capable of tracking objects by predicting their precise trajectories. However, we will show in later parts that our framework can work well with a simple tracker by adding a series of surprise explanations. Let G denote our tracker. G is implemented using Hungarian algorithm based on the proximity between the detected objects in frame t with the set of object tracks $\{S^n(t-1)|n = 1...N\}$. We only use the proximity and defer the appearance information for later violation detection in section 3.4.

3.2. State Encoding

For s_t^h, given the observation o_t and s_{t-1}, we perform the tracking algorithm described above. Given the newly associated object, we can directly extract the position p_t^h, i.e the bounding box itself, and temporal velocity v_t^h by subtracting the current bounding box's center with the previous one's. For those tracks that are associated with a proposal in current frame, then their occlusion status $z_t^h = 0$. Otherwise, for unassociated tracks, $z_t^h = 1$.

3.3. Transition Prediction

Let $\hat{s}_{t|t-1} = (\hat{p}_t, \hat{v}_t, \hat{d}_t)$. The predicted temporal velocity of the object is calculated by averaging the temporal velocity v_{t-1}^h, difference of position between consecutive frames, and the past velocity. Briefly, $\hat{v}_t = \beta v_{t-1} + (1 - \beta)v_{t-1}^h$. Therefore, the new position of the object can be specified by $\hat{v}_t$ and p_{t-1}, i.e. $\hat{p}_t = p_{t-1} + \hat{v}_t$. $\hat{d}_t$ is updated using a linear model.

3.4. Explanation of Surprise (XoS)

Surprises can come from many sources including but not limited to violations of physics (e.g discontinuity in movement or inconsistent appearance of an object along the video) and the inability of the model to detect rare events

(e.g collision or appearance deformation caused by partial occlusion). Based on types of violations, we classify surprises into 2 classes, motion-based and appearance-based surprises which are captured by L_t^T and L_t^O respectively in equation (1). Given s_t^h, $\hat{s}_{t|t-1}$, and o_t, we compute the Violation of Expectation (VOE) signal for a particular object u at timestep t as follow:

$$VOE_t^u = \alpha L_t^O + (1-\alpha) L_t^T \tag{1}$$

in which L_t^O is proportional to the similarity between the current appearance with the appearance at the nearest timestep t_0 when the object is not occluded:

$$L_t^O \propto F(o_t, o_{t_0}) \tag{2}$$

where F is function that outputs the similarity of 2 objects. We will discuss this function in section 3.5. When L_t^O is too small (i.e $L_t^O < \theta_O$), it is not necessarily that a violation would takes place. Potential reasons could be either partial occlusion or drastic change in object's size. If this is the case, we need to reset L_t^O to 1. This event can be checked by comparing the current object's size with the average size of the object in the last k frames.

When the object is visible $z_t^h = 0$, we assume that the observed velocity v_t^h follows the normal distribution with $\mu = \hat{v}_t$ and σ^2 which is manually tuned by using the validation set. Therefore, the transition likelihood is obtained by:

$$L_t^T \propto \exp\left(-(v_t^h - \hat{v}_t)^2/\sigma^2\right) \tag{3}$$

If $L_t^T < \theta_T$ (i.e potential violation), then we check the region in the moving direction. If there are some potential objects, e.g. floor, other objects, occluder, etc., that can collide with the object, we reset to $L_t^T = 1$, otherwise keep L_t^T the same. When the object disappears $z_t^h = 1$, assume that the object is occluded if $d_t^o < \hat{d}_t$, then:

$$L_t^T \propto \exp\left(-max(d_t^o - \hat{d}_t, 0)\right) \tag{4}$$

We made an assumption that our tracker can track the object reasonably well. Therefore, $z_t^h = 1$, which happens only if we cannot associate the track of this object with any proposal in the current frame, indicates that the object either is occluded by some occluders or simply disappeared (i.e violation of the spatial-temporal continuity rule). In this situation, L_t^T computed in equation (4) is the likelilhood of the object being actually occluded.

3.5. Appearance Matching

We implement appearance matching function **F** using the Siamese network described in figure 2. To extract objects' features, we first extract objects from video frames using bounding boxes given by the tracker. The images are then resized to 32×32, normalized between -1 and 1, and fed

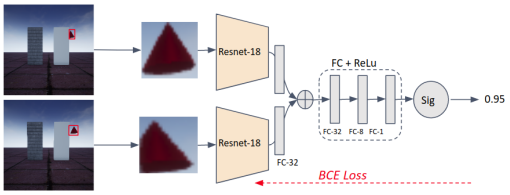

Figure 2. Appearance Matching network

to the Resnet-18. The resulting feature vectors are concatenated and fed through a neural net including 3 fully connected layers with 32, 8, and 1 hidden nodes respectively. A sigmoid activation is applied to squash the similarity value between 0 and 1. Training the Siamese requires triples of images, each including an anchor, a positive, and a negative image. We build a triple by first randomly choosing an object from a arbitrary video. The positive one is chosen by picking an object whose properties such as material, color, and shape match those of the anchor. An object which has at least 1 different property is chosen as the negative object.

4. Evaluation on IntPhys

4.1. Dataset

The Intphys dataset [10] includes training, validation, and test sets organized as follow:

- Training set includes 11250 videos at 15fps. Each video has 100 frames of size 284×284. Besides, the segmentation masks and depth maps are also provided. All the videos are possible; there are no violations.
- Validation and test's videos are structured in groups of 4 or so-called k-uplet with $k = 4$. Each group contains 4 videos, 2 possible and 2 impossible scenarios. The test set includes 3240 groups split equally into 3 blocks. Each block contains videos to exclusively test a concept.

Each video has several components that can be categorized into 5 classes: sky, floor, wall, occluder, and object. The object class can be further classified as cube, cone, and sphere. The scene in a video is viewed from a fixed-position camera. The difficult levels are controlled by the number of objects, type of occlusion, and the dynamics.

4.2. Validation

To evaluate the results, the authors of [10] propose using 2 types of metric, absolute and relative error rate. Consider a k-uplet $S_{i=1..N} = \{Pos_i^1..Pos_i^{k/2}, Imp_i^1..Imp_i^{k/2}\}$. Then we define relative error and the absolute error as:

$$L_{Rel} = \frac{1}{N} I\{\Sigma_j P(Pos_i^j) < \Sigma_j P(Imp_i^j)\} \tag{5}$$

	block O1		block O2		block O3	
	rel	**abs**	**rel**	**abs**	**rel**	**abs**
dpathak	**0.07**	**0.15**	**0.18**	**0.22**	**0.16**	**0.19**
dannygut	0.17	0.27	0.43	0.41	0.38	0.38
jcronsenb	0.34	0.32	0.43	0.43	0.41	0.4
cassini	0.12	0.42	0.21	0.37	0.36	0.47
Ours	**0.11**	**0.24**	**0.08**	**0.17**	**0.21**	**0.28**

Table 1. Intphys 2019 Leaderboard. The best and second best results are colored in red and blue respectively.

$$L_{Abs} = 1 - AUC(\{i, j; P(Pos_i^j)\}, \{i, j; P(Imp_i^j)\}) \quad (6)$$

The relative error encourages that within a set the score of positive videos should be higher than that of negative videos. On the other hand, the absolute error rate encourage that globally the average score for positive videos should be higher.

4.3. Results

According to the leaderboard as shown in 1, our framework achieves 2nd place in block O1 and O3 in term of relative and absolute errors. For block O2, we achieved the highest score. Unfortunately, detailed descriptions of the competing systems are not yet available for technical comparison.

5. Summary and Future work

We proposed a Surprise and Explain (SnE) framework which was shown to be competitive in the IntPhys Challenge at the time of this writing. The key idea was to assume that objects in the world typically follow simple dynamics that are easy to predict. However, when the simple dynamics are violated, explanations for those violations can be generated for typical situations that satisfy normal intuitive physics. In this paper, the simple dynamic model and explanation module were engineered components. In future work, it will be interesting to learn both of these components. Further, there are a number of straightforward and more fundamental improvements that can be made to the perception module of the current system. It will also be interesting to consider tighter integrations of perceptions and the rest of the framework. In particular, a key advance will be to learn latent state representations grounded in visual data that support the assumptions of the SnE framework.

Acknowledgement

This work is supported by DARPA under grant N66001-19-2-4035.

References

[1] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016. 1

[2] 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. 1

[3] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016. 1

[4] Sebastien Ehrhardt, Aron Monszpart, Niloy J Mitra, and Andrea Vedaldi. Learning a physical long-term predictor. *arXiv preprint arXiv:1703.00247*, 2017. 2

[5] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR*, pages 770–778, 2016. 1

[7] Jogendra Nath Kundu, Rahul M V, Jay Patravali, and Venkatesh Babu R. Unsupervised cross-dataset adaptation via probabilistic amodal 3d human pose completion. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[8] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 1

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[10] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *CoRR*, abs/1803.07616, 2018. 1, 2, 3

[11] Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *NeurIPS*, 2019. 2

[12] Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992. 1

[13] Sjoerd Van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. *arXiv preprint arXiv:1802.10353*, 2018. 1

[14] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *NeurIPS*, 2017. 1

[15] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh 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. 1