How can a Surprised AI Learn Commen Sense and Reasoning

Feiyang Xie
Yuanpei College
Peking University
2100017837@stu.pku.edu.cn

Abstract

As a field that has been widely studied in recent years, intuitive physics is indispensable for General Artificial Intelligence. There are already many methods and models, such as heuristics, probabilistic simulation, and learning-based models like PhysNet. However, modeling intuitive physics in complex situations remains very difficult for current AI systems. In this essay, we begin by analyzing some magic shows and discussing which commonsense principles they violate. We then propose a method to learn intuitive physics and common sense based on a unique AI that can be surprised when its expectations based on common sense are violated in a magic trick. Finally, we explore future research directions toward more general AI systems that understand human intuitive physics.

1 Introduction

Intuitive physics is used constantly in daily life. We frequently, rapidly, effortlessly and automatically make intuitive inferences about the physical world. Although we have common misconceptions and biases when predicting, judging, and explaining phenomena in the physical world [1], we rely on intuitive physics to understand our surroundings and predict what will happen next. Clearly, this ability will also be crucial for future robots and artificial intelligence systems. Magic shows often surprise us because they violate our intuitive understanding of physics and common sense. If we develop an AI system that can be surprised by magic tricks in the same way humans are, we could potentially combine existing methods (such as the Intuitive Physics Engine) with contrastive learning to enable the system to acquire common sense knowledge and use it for reasoning.

2 Some magic show examples

2.1 Coin passes through membrane:object solidity

In this magic trick, the magician appears to have a coin outside a cup sealed by an intact membrane. He passes the coin through the membrane into the cup, without breaking the membrane. According to our common sense, two solid objects cannot pass through each other. If they do, one of the solids must have a "channel" allowing the other solid to pass through. Only liquids can interpenetrate without being affected. Clearly, this magic trick violates our intuitive understanding of physics.

The secret behind this illusion is that the coin is actually inside the membrane from the start. The coin only appears to be outside the cup before the magician "pushes" it through the intact membrane.

2.2 Floating banknote: gravity and Newtonian physics

In this magic trick, a banknote is initially placed on a table to show it is an ordinary banknote. The magician then picks it up in his hand and after some manipulations makes the banknote levitate in the air. According to our common sense, gravity acts everywhere on Earth. Newton's second law states

(a) **Subfigure 1**. Initial state of the magic show.

(b) **Subfigure 2**. End state of the magic show.

Figure 1: **Coin Passes Through Membrane**

(a) **Subfigure 1**. Initial state of the magic show.

(b) **Subfigure 2**. End state of the magic show.

Figure 2: **Floating Banknote**

that objects accelerate when an external force is applied. Thus, we know that unsupported objects affected by gravity will fall. However, the banknote in this trick defies gravity and remains suspended in mid-air, violating our intuitive concept of forces.

The secret to the illusion is that nearly invisible threads are attached to the banknote in the magician's hand, allowing him to make it appear to float.

2.3 Disappearing arm:object permanence and contact

(a) **Subfigure 1**. Initial state of the magic show.

(b) **Subfigure 2**. End state of the magic show.

Figure 3: **Disappearing Arm**

In this magic trick, the magician initially reaches into an opaque box. He then removes the middle section of the box, revealing that the middle part of his arm is missing, while the front part can still move. This illusion violates two principles of common sense. First, it defies object permanence - the idea that objects continue to exist even when no longer perceivable. We infer from this that objects cannot vanish or appear from nothing. Thus, the arm should not disappear. Second, it violates the

notion that forces cannot be transmitted over distance, excluding electromagnetic fields. So after losing the middle section, the fingers should not move.

The secret behind the illusion is that the magician hides his real arm under the box the whole time.

3 Related works

3.1 Contrastive Learning

Contrastive learning is an unsupervised deep learning technique for learning data representations. The goal is to learn representations in which similar data instances are close together in the representation space, while dissimilar instances are far apart. Contrastive learning has been shown effective for various computer vision and natural language processing tasks, including image retrieval, zero-shot learning, and cross-modal retrieval. For these tasks, the learned representations can be used as features for downstream tasks like classification and clustering [2].

3.2 Intuitive Physics Engine

Physics Engine is computer software that provides an approximate simulation of certain physical systems [3]. Whether humans have an intuitive physics engine is a matter of debate. In my opinion, The function of Intuitive Physics Engine is to simulate the state of the objects and predict the status and movement of objects over time by the Intuitive Physics law. It can also refer to the core part of AI used to learn physical knowledge.In the noisy Newton framework, physics engine is defined by the principle of some normative physics laws and we can obtain inference by passing noisy information through it[4].In Probabilistic simulation models, physical variables provided as the input to a physics engine, and physical principles encoded in the engine.Some deep learning methods show that learning-based methods can be effectively integrated with a knowledge-based physics engine to infer the attributes and dynamics of objects in the environment [5].

4 Method

Considering our method relies on a unique AI that can be surprised by magic tricks, we can use contrastive learning to rapidly learn physics knowledge. Inspired by RND [6], when encountering novel scenarios, RND gives high intrinsic rewards. Given a dataset D of various scenarios, we input the scene S into the AI and get output R, which is high if S violates common sense (showing the AI's surprise). From R's level, we roughly classify D into positive samples S^+ and negative samples S^- for contrastive learning. Referring to GANs [7], we build a basic generative model G and use the unique AI as the discriminant model D. With embeddings z as input, G tries to predict objects' positions and motions over time steps. The goal is minimizing D's reward - making G's generated scenes match real physics. Finally, G serves as a physics engine, having learned common sense physics and gained reasoning skills by predicting scenario developments.

Pros This contrastive learning approach enables self-supervised learning of data embeddings, allowing us to leverage large volumes of unlabeled online video data depicting physical scenes. Obtaining diverse, abundant data in this manner facilitates the physics engine's learning across more scenarios. With an AI surprised by magic tricks providing supervision, generative models can acquire more accurate and reasonable physical knowledge.

Cons Training the contrastive learning component will be challenging since effectively selecting positive and negative samples for learning common sense is difficult. Inputting two random frames from a time series into the unique AI may also produce a high reward output, even though no physics common sense is contained.

5 Conclusion

Though magic tricks defy physics intuition, they present an opportunity to advance AI's intuitive physics capabilities. With an AI that reacts with surprise to magic, we can label physical scenes as positive or negative examples for contrastive learning to represent common sense. A generative model can then perform reasoning, yielding an improved physics engine. However, this model

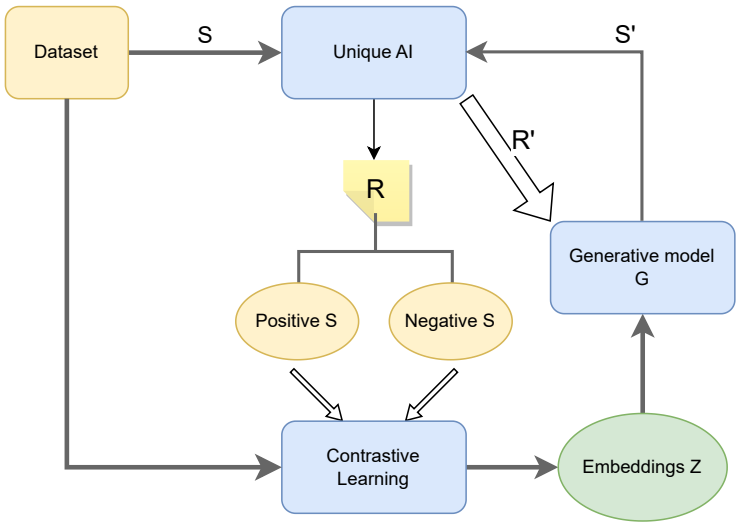

Figure 4: **Framework**

may generalize poorly and encounter training obstacles. Effective representation, acquisition, and application of physical common sense for reasoning remain challenges.

References

[1] James R. Kubricht, Keith J. Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10):749–759, 2017. 1

[2] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases, 2022. 3

[3] Robert D. Christ and Robert L. Wernli. Chapter 4 - vehicle control and simulation. In Robert D. Christ and Robert L. Wernli, editors, *The ROV Manual (Second Edition)*, pages 93–106. Butterworth-Heinemann, Oxford, second edition edition, 2014. 3

[4] Adam N. Sanborn, Vikash K. Mansinghka, and Thomas L. Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, 120 2:411–37, 2013. 3

[5] Jiajun Wu, Ilker Yildirim, Joseph J. Lim, William T. Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. volume 2015-January, page 127 – 135, 2015. Cited by: 198. 3

[6] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation, 2018. 3

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3