Physion++: Evaluating Physical Scene Understanding with Objects Consisting of Different Physical Attributes in Humans and Machines

Hsiao-Yu Tung
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Mingyu Ding
University of California, Berkeley, Berkeley, California, United States

Zhenfang Chen
MIT-IBM Watson AI Lab, Cambridge, Massachusetts, United States

Sirui Tao
University of California San Diego, La Jolla, California, United States

Vedang Lad
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Daniel Bear
Stanford University, Stanford, California, United States

Chuang Gan
MIT-IBM Watson AI Lab, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Daniel Yamins
Stanford University, Stanford, California, United States

Judith Fan
Stanford University, Stanford, California, United States

Kevin Smith
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

Human physical scene understanding requires more than simply localizing and recognizing objects — we can quickly adapt our predictions about how a scene will unfold by incorporating objects' latent physics properties, such as the masses of the objects in the scene. What are the underlying computational mechanisms that allow humans to infer these physical properties and adapt their physical predictions so efficiently from visual inputs? One hypothesis is that general intuitive physics knowledge can be learned from enough raw data, instantiated as computational models that predict future video frames in large datasets of complex scenes. To test this hypothesis, we evaluate existing state-of-the-art video models. We measured both model and human performance on Physion++, a novel dataset and benchmark that rigorously evaluates visual physical prediction in humans and machines, under circumstances where accurate physical prediction relies on accurate estimates of the latent physical properties of objects in the scene. Specifically, we tested scenarios where accurate prediction relied on accurate estimates of objects' mechanical properties, including masses, friction, elasticity and deformability, and the values of these mechanical properties could only be inferred by observing how these objects moved and interacted with other objects and/or fluids. We found that models that encode objectness and physical states tend to perform better, yet there is still a huge gap compared to human performance. We also found most models' predictions correlate poorly with that made by humans. These results show that current deep learning models that succeed in some settings nevertheless fail to achieve human-level physical prediction in other cases, especially those where latent property inference is required.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*