Physics 101: Learning Physical Object Properties from Unlabeled Videos

Jiajun Wu[1], Joseph J. Lim[2], Hongyi Zhang[1], Joshua B. Tenenbaum[1], William T. Freeman[13]
[1] {jiajunwu,hongyiz,jbt,billf}@mit.edu
[2] lim@csail.mit.edu

[1] Massachusetts Institute of Technology
[2] Stanford University
[3] Google Research

Introduction We study the problem of learning physical properties of objects from unlabeled videos. Humans can learn basic physical laws when they are very young [1], which suggests that such tasks may be important goals for computational vision systems.

There have been early efforts to build computer vision systems with the physical knowledge of an early child. Recently, researchers started to tackle concrete scenarios for understanding physics from vision [2], some involving deep learning. Different from these, we aim to develop a system that can infer physical properties, *e.g.* mass and density, directly from visual input. Our method is general and easily adaptive to new scenarios, and is more efficient compared to analysis-by-synthesis approaches [3].

Physical World Model There exist highly involved physical processes in daily events in our physical world. We can divide all involved physical properties into two groups: the first is the intrinsic physical properties of objects like mass, many of which we cannot directly measure from the visual input; the second is the descriptive physical properties, *e.g.* velocity of objects, which characterize the scenario in the video. The second group of parameters are observable, and are determined by the first group, while both of them determine the content in videos.

Physics 101 Dataset We collected a dataset of 101 objects made of different materials and with various masses and volumes. We started by collecting videos of them from multiple viewpoints in four scenarios: objects slide down an inclined surface and possibly collide with another object; objects fall onto surfaces made of different materials; objects splash in water; and objects hang on a spring. These seemingly straightforward setups require understanding multiple physical properties, *e.g.*, material, mass, volume, density, coefficient of friction, and coefficient of restitution. We called this dataset Physics 101.

Method Our method is a CNN consisting of three components. The bottom component is a *visual property discoverer*, which aims to discover physical properties like material or volume which could at least partially be observed from visual input; the middle component is a *physics interpreter*, which explicitly encodes physical laws into the network structure and models latent physical properties like density and mass; the top component is a *physical world simulator*, which characterizes descriptive physical properties like distances that objects traveled, all of which we may directly observe from videos. Our network corresponds to our physical world model, and learns object properties from unlabeled data.

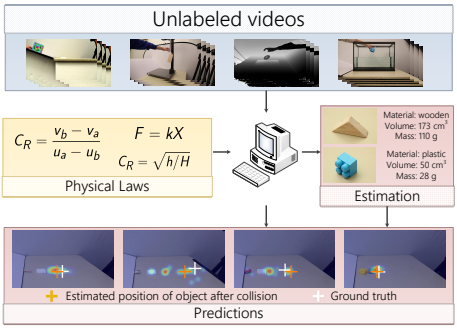

Figure 1: Overview of our model, which learns directly from unlabeled videos, produces estimates of physical properties of objects based on the encoded physical laws, and generalizes to tasks like outcome prediction

Evaluation We demonstrate that our framework develops some physics competency by observing videos. We also show that our generative model can transfer learned physical knowledge from one scenario to the other, and generalize to other tasks like predicting the outcome of a collision.

[1] Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004.

[2] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *PNAS*, 110(45):18327–18332, 2013.

[3] 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. In *NIPS*, 2015.