A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding

Renqiao Zhang, Jiajun Wu, Chengkai Zhang, William T. Freeman, Joshua B. Tenenbaum

1Massachusetts Institute of Technology  2Google Research

{andy17, jiajunwu, ckzhang, billf, jbt}@mit.edu

Abstract
Humans demonstrate remarkable abilities to predict physical events in complex scenes. Two classes of models for physical scene understanding have recently been proposed: “Intuitive Physics Engines”, or IPEs, which posit that people make predictions by running approximate probabilistic simulations in causal mental models similar in nature to video-game physics engines, and memory-based models, which make judgments based on analogies to stored experiences of previously encountered scenes and physical outcomes. Versions of the latter have recently been instantiated in convolutional neural network (CNN) architectures. Here we report four experiments that, to our knowledge, are the first rigorous comparisons of simulation-based and CNN-based models, where both approaches are concretely instantiated in algorithms that can run on raw image inputs and produce as outputs physical judgments such as whether a stack of blocks will fall. Both approaches can achieve super-human accuracy levels and can quantitatively predict human judgments to a similar degree, but only the simulation-based models generalize to novel situations in ways that people do, and are qualitatively consistent with systematic perceptual illusions and judgment asymmetries that people show.

Keywords: physical scene understanding; neural network; analysis by synthesis; simulation engine; blocks world

Introduction
The outputs of vision include not only the objects in a scene and their spatial relations, but also their physical properties and relations: What is heavy or light? What is balanced or attached, and what isn’t? What is likely to fall? What will happen next? When objects move, their motion can be predicted from these physical inferences; motion can also affect our physical judgments when objects move in unexpected ways.

These capacities for physical scene understanding are basic to how we see the world. Precursors to them can be found in infants as young as 3-5 months old, even before children acquire their first words labeling kinds of objects (Carey, 2009; Baillargeon, 2004). Building computational models of these abilities has been a target for recent work in both cognitive science and computational vision (Battaglia, Hamrick, & Tenenbaum, 2013; Gupta, Efros, & Hebert, 2010; Motzghi, Bagherinezhad, Rastegari, & Farhadi, 2015; Fragkiadaki, Agrawal, Levine, & Malik, 2015; Zheng, Zhao, Yu, Ikeuchi, & Zhu, 2015; Li, Azimi, Leonardis, & Fritz, 2016). In contrast to earlier work on intuitive physics that emphasized explicit reasoning about textbook-style physics problems (McCloskey, 1983), with models focused on people’s qualitative judgments (Forbus, 1984; Siegler, 1976), recent studies of physical scene understanding have looked at more rapid, perceptual inferences, which can be parametrically manipulated and modeled quantitatively, and which could serve as the basis for grounded action planning. Several studies have argued that rapid perceptual inferences about the physics of scenes can be explained by positing an “intuitive physics engine” (IPE), a mental system for approximate probabilistic simulation analogous to those used in video-game physics engines (Sanborn, Mansinghka, & Griffiths, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; K. A. Smith & Vul, 2013). These simulation engines approximate object dynamics interacting under Newtonian or other forms of classical mechanics over short time scales, in ways that are perceptually reasonable (if not necessarily physically accurate) and efficient enough to run in real time for complex scenes.

Other authors have suggested that the simulation-based IPE scheme might be prohibitively expensive for brains to implement (Davis & Marcus, 2016). An alternative class of models has been proposed based on stored memories of experienced scenes and physical outcomes, together with pattern recognition algorithms (such as neural networks) for accessing appropriate memory items to predict outcomes in a new scene context (Sanborn et al., 2013; Sanborn, 2014).

Although cognitive scientists have yet to seriously test memory-based alternatives to simulation in physical scene understanding tasks, AI researchers at Facebook recently demonstrated such a possibility in a working system. Lerer, Gross, and Fergus (2016) trained deep convolutional neural networks (CNNs) to make physical predictions directly from visual images, judging for instance whether a stack of blocks will fall, as Battaglia et al. (2013) studied empirically and modeled using approximate probabilistic simulation. The FAIR neural network, named PhysNet, was partly pretrained on ImageNet (Krizhevsky, Sutskever, & Hinton, 2012) and then trained on after a large dataset of synthetic scenes and outcomes. It achieved a high accuracy (89%) on the stability prediction task, generalized to real images reasonably well (67%), and exhibited positive correlations with human responses. This suggests that memory-based systems for visual intuitive physics may be promising at least in AI applications, and perhaps also as cognitive models.

Motivated by the success of CNNs in machine vision object recognition tasks (Krizhevsky et al., 2012), neuroscientists have proposed analogous architectures as accounts of the fast feedforward aspects of human visual object recognition (Yamins et al., 2014; Serre, Oliva, & Poggio, 2007). If CNNs can be successfully applied to physical scene understanding tasks as well, they could offer a compelling alternative to simulation as an account of how people can predict physical outcomes so well, so quickly.

Our goal in this paper is to conduct the first rigorous em-
empirical comparisons of simulation-based (IPE) and neural-
network-based (CNN) models for physical scene understand-
ing. Although CNNs have many appealing features as models
of visual cortex, they also have features that are less appealing –
and arguably less human-like. They typically require large
amounts of training data, which a human might not have ac-
cess to. Large training sets may be required for any new sce-
nario, even if it is just a simple variation on previously seen
cases. For instance, in order to predict whether a pile of four
blocks is stable, a CNN may have to see at least thousands of
cases that either do or do not fall under gravity. In contrast,
an IPE model, just like humans, is able to make many predic-
tions with reasonable accuracies without training, as the sim-
ulation engine within encodes abstract physical knowledge
that applies to a very wide range of scenes.

Even with a large amount of training data, it is unclear
whether the knowledge learned by CNNs may be transfer-
able to some similar cases. Lerer et al. (2016) showed that
a network trained on images of two and four blocks could
generalize to images of three blocks to some extent, but there
is no clear way for a neural network to answer a different
but related question to those it is trained for, e.g., in which
direction the blocks would fall, unless explicit labels are pro-
vided during training. One of the main points in favor of IPE
models is their ability to explain how people can easily make
many different judgments about very different configurations
of blocks, without specific training (Battaglia et al., 2013).

Perhaps most interestingly, people are prone to systematic
“physics illusions” that IPE models naturally capture. For in-
stance, stacks of blocks often look to people as if they are sure
to fall when they are actually carefully balanced. People do
not, however, make the opposite error: They do not system-
atically mistake unstable stacks for stable ones. Probabilistic
simulation-based models are similarly tempted to make this
asymmetric pattern of errors (Battaglia et al., 2013): Small
amounts of uncertainty in the simulation can make a sta-
bility configuration appear unstable, but are unlikely to make
an unstable one appear stable. It is unclear whether neural-

In this paper, we report four experiments comparing the
behavior of discriminatively trained neural networks and gen-
erative simulation-based models with human judgments on
blocks-world physics tasks, addressing the questions above.
Exp. 1 evaluates the performance of the IPE model and performance-optimized neural networks in predicting block
stability. Exp. 2 explores the role of limiting CNN train-
ing data, to see if performance on smaller training sets looks
more human-like. Exp. 3 evaluates both model classes for
asymmetries in the stability illusions described above. Exp. 4
tests CNNs and IPE models’ ability to generalize to situations
slightly different from those the CNN was trained on.

The Blocks World

For our experiments, we study a set of seemingly simple but
physically rich scenarios: a pile of blocks with one on top of

Figure 1: Sample stimuli used by Battaglia et al. (2013),
Facebook AI Research (Lerer et al., 2016), and us. Our stim-
uli are ordered by increasing visual instability (defined in Ex-
periment 3) another. Our goal is to study how humans and computational
models behave on various tasks given these stimuli, and to
reveal possible correlations between them. We now illustrate
our stimuli in detail.

For each stimulus, there are four blocks with side length 1
meter piled on the ground, each supporting another on top of
it. There is only one block at the same height level. Because
laying blocks at uniform random is likely (p = 75%) to re-
sult in an unstable system, we draw the horizontal position of
a block from a normal distribution with variance 0.29 centered
at the horizontal position of the block under it, to ensure that
there are both stable and unstable piles in the dataset. Later,
we study cases where the number of blocks varies, and
for them we update the variance accordingly.

Whether blocks are stable, i.e., groundtruth labels, can be
derived from the coordinates of blocks. A block will fall if
and only if the center of mass of all blocks above it, including
itself, does not fall on top of the block under it.

For rendering, we generate images of resolution 256 ×
256. We place a pile of blocks in a virtual experiment field
with a size of 30 × 30 meters and a height of 4 meters. We
have one light source, 16 meters high, to simulate real-life
lightening. We also vary the position, focal point, and tilt an-
gle of the camera. We represent its coordinates in cylindrical
coordinates (r, θ, z), with origin on the ground right beneath
the center of the bottommost block. The camera positions
are sampled from $r \sim N(11, 0.3^2)$, $\theta \sim \text{Uniform}(0, \pi/2)$, and
$z \sim N(3, 0.01^2)$. We choose these parameters to ensure all
blocks are within the view of the camera. The focal point of
the camera is set at the center of the pile plus a Gaussian noise
with variance 0.2. We also tilt the camera; its angle from the
We set the friction coefficient to 0. The direction of projection is sampled from \( N(0, 2^2) \). We incorporate these variances for evaluating the generalization ability of the models.

**Computational Models**

We study two classes of computational models. One is the Intuitive Physics Engine (IPE) Model (Battaglia et al., 2013), which aims to simulate humans’ reasoning on physical scenes by an approximate probabilistic simulation engine. The other is convolutional neural networks (CNNs), a class of discriminative recognition models that have gained much popularity in AI fields like computer vision in recent years.

**The Intuitive Physics Engine Model**

The Intuitive Physics Engine (IPE) consists of two components: a Bayesian vision system, which infers the configurations of blocks from given images, and a physical inference system, which calculates the Bayesian posterior probability distribution of physical properties (i.e., stability) by running a number of simulations under perturbation forces and geometric noises. Figure 2 illustrates the IPE model. For more details, please see Battaglia et al. (2013).

For each scene, we render images of the initial state under perspective projection from three fixed viewpoints rotated by 45°. These triplets of images are then fed into the Bayesian vision system, which uses a Metropolis-Hasting (MH) sampling algorithm to infer a Bayesian posterior distribution of the scene’s initial state (position, height, and the number of blocks presented). We run the MH sampling for 5,000 steps, with a 2D Gaussian blurring kernel of width 2 on the observed images, as suggested by Battaglia et al. (2013).

With the inferred initial geometry, we run 20 simulations for each scene using the Open Dynamics Engine (ODE) (R. Smith, 2006). We set the friction coefficient to 0.2, the bounce coefficient to 0.2, and the side-length and density of each block to 1m and 500kg/m³, respectively. Gravity is set to 9.81m/s² pointing downwards. Before each simulation starts, a horizontal zero mean Gaussian noise \( \sigma \) is added to the positions of blocks. Then the simulation runs at a step size of 10ms for 2 seconds. During the first second, a horizontal force with magnitude \( \phi \) is exerted at the center of the bottom face of the bottommost block. The direction of the force is uniformly sampled from \((0, 2\pi)\) and changes at a frequency of 50Hz. We consider a pile unstable if the vertical coordinate of the top block changes by more than 0.2 meters when the simulation ends.

**Convolutional Neural Networks**

CNNs have gained much popularity in computer vision (Krizhevsky et al., 2012). Here we consider two popular CNN frameworks: the small but powerful LeNet (LeCun, Bottou, Bengio, & Haffner, 1998), and the widely used AlexNet (Krizhevsky et al., 2012).

LeNet, originally proposed for digit recognition, has been widely used as a recognition model in vision because of its effectiveness and simplicity (LeCun et al., 1998). LeNet consists of two convolutional layers, each followed by a pooling layer and an activation layer. There are then two fully connected linear layers at the end. We modify the final layer so that instead of ten outputs for digit classification, the model now has two output units — its confidences on whether the blocks will fall or not. Figure 3 shows the structure of LeNet.

The second is the popular AlexNet (Krizhevsky et al., 2012), which achieves impressive performance on ImageNet classification. AlexNet consists of five convolutional, pooling, and activation layers, and three linear layers at the end. We evaluate both AlexNet pretrained on ImageNet, as well as AlexNet trained from scratch.

We use Torch (Collobert, Kavukcuoglu, & Farabet, 2011) for implementation. We set the learning rate to 0.01 for LeNet and for fine-tuning AlexNet, and to 0.2 for training AlexNet from scratch. We use stochastic gradient descent for training.

**Behavioral Experiments**

To collect human responses, we first randomly divide all test images into groups, each consisting of 10 images. We then add four easy cases (two stable, two unstable), whose stability is visually apparent, into the group. For each group, we collect 50 responses on Amazon Mechanical Turk. We only allow workers with an approval rate > 90% to submit responses, and we only accept responses from workers that answered all four easy cases correctly.

**Experiment 1: Predicting Falling Blocks**

In our first experiment, we test the performance of the IPE model and neural networks on images with four blocks, and compare the results with human responses.

**Experimental Setup** For the IPE model, we consider cases with various levels of geometric Gaussian noises \( \sigma \) and external forces \( \phi \) during physical simulations. We then compare their performance with LeNet, AlexNet, and humans.
Table 1: Accuracies (%) of the IPE model with different $\sigma$ and $\phi$, and their correlations with human responses. We use $(\sigma, \phi) = (0.1, 40)$ for following experiments.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\phi$</th>
<th>0</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>94.2</td>
<td>87.2</td>
<td>79.5</td>
<td>71.3</td>
<td>63.8</td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>91.3</td>
<td>83.4</td>
<td>76.1</td>
<td>69.1</td>
<td>61.8</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>83.2</td>
<td>75.7</td>
<td>70.3</td>
<td>62.6</td>
<td>56.4</td>
</tr>
<tr>
<td>0.15</td>
<td></td>
<td>72.2</td>
<td>66.8</td>
<td>59.4</td>
<td>54.2</td>
<td>51.2</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td>58.5</td>
<td>53.8</td>
<td>52.1</td>
<td>51.0</td>
<td>50.9</td>
</tr>
</tbody>
</table>

Corr $\geq 0.45 \geq 0.54 \geq 0.56 \geq 0.58 \geq 0.60$

Table 2: Accuracies (%) of humans, IPE, LeNet, and AlexNet (pretrained and not pretrained), on 200K or 1,000 images. The results on 1,000 images are averaged over five models trained on independently sampled sets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Stable</th>
<th>Unstable</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>38.0</td>
<td>92.9</td>
<td>65.5</td>
</tr>
<tr>
<td>IPE</td>
<td>40.7</td>
<td>99.0</td>
<td>70.3</td>
</tr>
<tr>
<td>LeNet (200K)</td>
<td>91.3</td>
<td>89.0</td>
<td>90.1</td>
</tr>
<tr>
<td>AlexNet (200K)</td>
<td>91.5</td>
<td>92.3</td>
<td>91.9</td>
</tr>
<tr>
<td>AlexNet (Pretrained, 200K)</td>
<td>94.5</td>
<td>94.7</td>
<td>94.6</td>
</tr>
<tr>
<td>LeNet (1,000)</td>
<td>68.0</td>
<td>69.3</td>
<td>68.7</td>
</tr>
<tr>
<td>AlexNet (1,000)</td>
<td>71.8</td>
<td>70.1</td>
<td>70.9</td>
</tr>
<tr>
<td>AlexNet (Pretrained, 1,000)</td>
<td>72.5</td>
<td>74.2</td>
<td>73.4</td>
</tr>
</tbody>
</table>

Figure 4: CNN models with different sizes of training sets the networks trained with 1,000 images. As shown in Table 2, there is still no asymmetric pattern in the responses of the less-trained networks.

We now look into how each model correlates with human responses in more detail. Figure 5 (a) and (b) demonstrate that the IPE model has a stronger correlation with humans, compared to LeNet trained on the full training set. Another interesting finding is that the less-trained LeNet (c) is more human-like. We will discuss this more in the final section.

**Experiment 3: Boundary Cases**

We now systematically study the asymmetry we observed in Experiment 1. In particular, we focus on a few groups with visually unstable piles, i.e., piles that are carefully balanced and therefore stable, but illusory to humans so that they believe these blocks will fall.

**Experimental Setup** We define visual instability, scaling from 0 to 5, to describe how unstable a pile of blocks looks like. A pile with instability value $x$ means there exists at least one block so that the center of mass of the blocks above it lies $x/10$ meters away from its center on x-y plane. As the side-length of blocks is 1 meter, a pile with a visual instability value 4 looks very unstable to humans, significantly different from one with value 1. Figure 6 shows examples with various visual instability values.

For this experiment, we restrict possible camera positions so that the deviations of blocks can be clearly perceived. We generate four datasets of stable blocks with visual instabilities of 1, 2, 3, and 4 respectively, each with 100 images.

**Results and Discussions** As shown in Figure 4, the performance of CNNs decreases as there are fewer training data. Although AlexNet (not pretrained) performs better with 200,000 training images, it also suffers more from the lack of data, while pretrained AlexNet is able to learn better from a small amount of training images. For our task, both models require around 1,000 images for their performance to be comparable to the IPE model and humans. We then evaluate
Figure 5: From left to right: human responses vs (a) responses of IPE (normalized numbers of moving blocks), (b) LeNet trained on the full training set (200,000 images), and (c) LeNet trained on 1,000 images. Results for AlexNet are similar. We list Pearson’s correlation coefficients at the bottom-right corner.

Results and Discussions
As shown in Figure 6, the performance of neural networks are, in general, better than their performance in Experiment 1, probably because images here are easier as the camera positions are restricted. Also, their performance barely changes for groups with different visual instabilities. Even for the most deceptive group (visual instability 4), a LeNet has an accuracy of 93%. We also test AlexNet (both pretrained and not pretrained) on cases where blocks are unstable but visually stable, and the network, again, gives highly accurate results (≥ 93%).

The performance of IPE and humans, on the other hand, changes drastically across groups. Corresponding to results in Experiment 1, both IPE and humans consistently predict that blocks with visual instability 4 will fall. Their accuracies are higher when visual instability is smaller, but still not close to those of neural networks. This confirms our observation of the asymmetry. More discussions follow in the final section.

Experiment 4: Knowledge Transfer
A possible explanation to humans’ one-shot learning ability is based on the concept of transfer learning. In our fourth experiment, we evaluate the behaviors of computational models on tasks involving knowledge transfer.

Experimental Setup
For this experiment, we generate 200 test images with three and five blocks, respectively. Examples are shown in Figure 7. We modify the variance of block positions to ensure there are half stable and half unstable cases.

Our Bayesian vision system is extended to include the number of blocks as one parameter in sampling. Because the number of blocks directly determines the total mass, we also vary the magnitude of the perturbation force according to the inferred number of blocks to keep its effect consistent. For neural networks, we simply test the models previously trained on the 200,000 images with four blocks.

Results and Discussions
Table 3 shows that while CNNs achieve ∼ 90% accuracies on four-block cases, their performance is much worse on cases where the number of blocks is smaller than that in training examples. Specifically, the predictions of models trained on 200,000 images are at chance. For cases with more blocks, CNNs, especially pretrained AlexNet, can learn to generalize to some extent. However, their behaviors are different from human responses. In comparison, humans and the IPE model have relatively consistent performance, with slight decreases in accuracies as the number of blocks goes up and the task becomes more difficult.

These experiments demonstrate that the knowledge learned by neural networks cannot be transferred, at least in a straightforward way, to scenarios outside the training set. The IPE model and humans enjoy more flexibility in reasoning in the complex world and solving more general problems.

General Discussion
Following Facebook AI’s reported results, we found that convolutional neural networks can be trained to achieve superhuman accuracy levels on stability judgment tasks from raw images (Exps. 1 and 2). CNNs also correlate reasonably well with human intuitions about how likely a stack of blocks is to fall, and once trained, they can respond to new images extremely quickly. However, these features do not automatically make CNNs a good model of people’s physical intuitions. They do not capture systematic judgment asymmetries that humans make, which simulation-based IPE models do capture (Exps. 1-3). CNNs also have limited generalization ability across even small scene variations, such as changing
the number of blocks. In contrast, IPE models naturally generalize and capture the ways that human judgment accuracy decreases with the number of blocks in a stack (Exp. 4).

Taken together, these results point to something fundamental about human cognition that neural networks (or at least CNNs) are not currently capturing: the existence of a mental model of the world’s causal processes. Causal mental models can be simulated to predict what will happen in qualitatively novel situations, and they do not require vast and diverse training data to generalize broadly, but they are inherently subject to certain kinds of errors (e.g., propagation of uncertainty due to state and dynamics noise) just in virtue of operating by simulation.

Despite the success of CNNs in accounting for other high-level human perceptual capacities, such as rapid object classification (Yamins et al., 2014), our results suggest that at least some perceptual judgments which people can make in a quick glance are not well explained by current feedforward neural networks. We should not conclude however, that neural networks cannot help to explain how people make intuitive physical judgments. If people do indeed have a “physics engine in the head”, somehow this simulator must be implemented in neural circuits. Recurrent neural networks (RNNs) could provide one model for this (Fragkiadaki et al., 2015). It is also possible that CNNs, if trained on more diverse scenes and physical judgments than those studied here and/or pretrained on large-scale image classification tasks (as in Lerer et al., 2016), could capture more of the qualitative inference behavior people show in our tasks. Lastly, CNNs could be useful for visual intuitive physics by quickly estimating the relevant object properties in images needed to represent the world’s state in a physics engine, which would then support more sophisticated reasoning and prediction by simulation (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). Going forward we are eager to explore these and other productive lines of exchange between simulation-based generative models and memory-based neural network models.

Table 3: Results on the task of transfer learning

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>LeNet (200K)</td>
<td>4</td>
<td>50.5</td>
</tr>
<tr>
<td>AlexNet (200K)</td>
<td>4</td>
<td>52.5</td>
</tr>
<tr>
<td>AlexNet (P, 200K)</td>
<td>4</td>
<td>51.0</td>
</tr>
<tr>
<td>LeNet (1,000)</td>
<td>4</td>
<td>57.0</td>
</tr>
<tr>
<td>AlexNet (1,000)</td>
<td>4</td>
<td>54.0</td>
</tr>
<tr>
<td>AlexNet (P, 1,000)</td>
<td>4</td>
<td>55.0</td>
</tr>
<tr>
<td>IPE (0.1,10x)</td>
<td>N/A</td>
<td>72.0</td>
</tr>
<tr>
<td>Human</td>
<td>N/A</td>
<td>76.5</td>
</tr>
</tbody>
</table>

Figure 7: Images with three or five blocks

Acknowledgement

We are grateful to Tomer Ullman for helpful discussions. This work is in part supported by NSF Robust Intelligence 1212849 Reconstructive Recognition, the Center for Brain, Minds and Machines (NSF STC award CCF-1231216), and MERL.

References


