Rapid Physical Predictions from Convolutional Neural Networks

Filipe de A B Peres*
Columbia University
New York, NY 10027
filipe.peres@columbia.edu

Kevin A Smith
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
k2smith@mit.edu

Joshua B Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
jbt@mit.edu

1 Introduction

Every day we use intuitive reasoning to make and update predictions about the world. For instance, when playing soccer we must predict the trajectories of the ball but also update those predictions in light of new information. But these predictions must also be rapid – if we turn to see a soccer ball flying at us, we must quickly decide whether or not to duck. What prediction mechanism would allow for such fast and flexible predictions?

Previous research has suggested that human physical intuition can be explained by probabilistic simulation using an "intuitive physics engine" [1,2], and that updating predictions can be explained as accumulating evidence from these simulations over time [3,4]. However, these simulations take time to perform [5], which may make it difficult to use for fast decisions. Since physical reasoning can influence rapid human judgments (e.g., object detection [6]), this type of capability would suggest a faster mechanism might be employed concurrently with physical simulation.

Convolutional neural networks (CNNs) were recently suggested as another class of models that could account for human physics judgments [7]. Although it is not clear whether these networks learn to make physical predictions in the same way as humans in situations where deliberation is possible [8], the faster nature of their feedforward architectures makes this class of models an interesting candidate for generating rapid physical judgments.

In this paper we used a CNN to learn and perform a continuous physics prediction task, and compared its performance both to human responses and to simulation-based physics models. We aimed to understand if CNNs can help explain human performance beyond the capabilities of probabilistic physics models, and at what time in the prediction process they might excel over physical simulation.

2 Methods

2.1 Task

For our experiments, we employed the same experimental task used by Smith et. al. [3,4]. In this task, participants predicted the path of a ball that bounces in a virtual environment such as in Figure 1.

*Correspondence should be sent to: Filipe de A. B. Peres, Columbia University, 6155 Lerner Hall, 2920 Broadway, New York, NY 10027

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

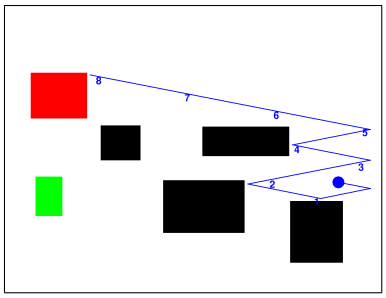

Figure 1: Illustration of a trial (the blue line was not visible but tracks the path of the center of the ball).

	CNN	Simulation
Correlation	0.84	0.96
Part. Correlation	0.19	0.85
Part. Cor (first 2 s)	0.36	0.74
Part. Cor (last 2 s)	0.02	0.89

Table 1: Correlations and partial correlations (partialling out the other model) between models and human responses.

After some time, this ball will reach one of two goals – red or green – and participants held down a button indicating which goal they predicted the ball would reach first. At any point participants could change the current prediction or choose to make no prediction at all, if uncertain. Predictions were measured every 100 ms.

2.2 Models

As a baseline for comparisons with the CNN, we used the same physics model and human response data as Smith [4]. This physics model runs multiple noisy simulations forward and bases its prediction on evidence accumulated from those simulations over time. The physics model was designed to also explain both when people made a prediction and which goal they predicted, but the CNN only predicts which goal the ball will reach. We therefore ignore the choice of whether to make a decision in the human and physics model data and focus on which goal was chosen given that a decision was made.

We trained a CNN adapted from a standard GoogLeNet [9] in two ways. First, to allow the network to receive motion information, we provided two consecutive frames (100 ms apart) to the input layer, which required doubling the depth of this layer. This also required doubling the depth of the first convolutional kernel to account for this modification. Second, the output layer was a single node that produced the probability of choosing the red vs. green goal.

The network was initialized with random weights, and we used supervised training based on the actual goal the ball reached, as calculated by a 2D physics engine [10]. The training set consisted of all frames from 10,000 different trials which were randomly created with a varying number of obstacles (1 to 5). The network was trained for 6400 steps, each with 64 training examples (consisting of a pair of consecutive frames). After training, the network was tested on an independent test set: the trials used in Smith [4].

3 Results

3.1 Does the model learn to predict physics outcomes?

The CNN was able to learn enough to predict the outcomes of trials better than chance both initially and throughout the trial, but did not reach human levels of performance. We first investigated the accuracy of initial predictions – the prediction from the first two frames of the CNN and the proportion of participants who chose the correct goal on their initial decision (typically chosen within the first second). The models first predictions were not as accurate as the average human accuracy (CNN accuracy = 0.59, human accuracy = 0.68).

As the trial progressed, both the CNN and participants became more accurate, but the increase in accuracy was greater for human decisions. The CNNs weighted average prediction for the correct goal over the entire trial was 0.66, as compared to an average human weighted accuracy of 0.82.[1]

[1] Predictions over the trial were weighted by the number of participants making a decision at each timestep to correct for the reliability of measurement at that point.

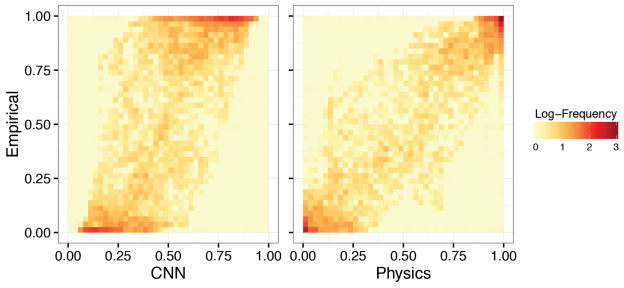

Figure 2: Joint histogram of the CNN (*left*) and physics model (*right*) predictions versus proportion of participants indicating the ball would reach the red goal, bucketed at each measured time step across all trials and weighted by the number of participants that made a decision. Redder areas represent higher frequency, and buckets along the diagonal represent observations where the model nearly perfectly matches empirical observations, suggesting the physics model explains human predictions with less bias and variability than the CNN.

3.2 How does the CNN compare to physical simulation?

While the CNN's predictions were well correlated to human judgments ($r = 0.84$, Figure 2 left), it does not explain these judgments as well as a physics simulation model ($r = 0.96$, Figure 2 right). Nonetheless, the CNN does explain some part of participants' judgments that the physics model cannot, as demonstrated by the partial correlation of the CNN's predictions to human judgments controlling for the predictions of the physics model ($r_p = 0.19$).

We also investigated *when* in the course of the trial the CNN explains participants' decisions better than a physical simulation model. Here we calculated the partial correlations of each model against human judgment (partialling out the other model) at each time step in the trial. The CNN has explanatory power not captured by the physics model in the first few seconds of each trial, but does not explain judgments beyond physical simulation towards the end of the trial (Table 1, Figure 3). This suggests that the CNN is capturing part of immediate human judgments, but participants' later predictions are more consistent with physical simulation.

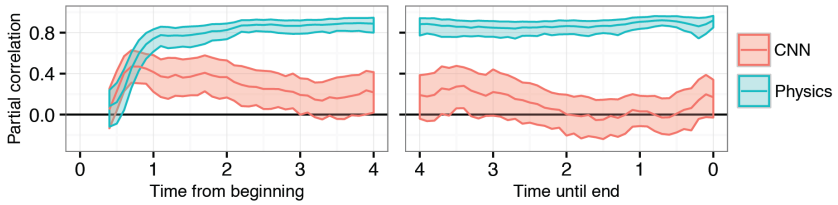

Figure 3: Partial correlations between both the CNN and the physics model to the human responses controlling for the other model at each time step over time: the first 4 seconds from the beginning of trials (*left*) and the last 4 seconds to the end of trials (*right*). 95% CIs calculated from 500 bootstrapped samples.

3.3 What is the CNN learning?

It is frequently hard to characterize what a neural network has learned. However, to investigate what heuristics it might be learning, we can test whether its predictions for the correct goal correlate with certain features of the trial: the distance to each goal, the angle between the ball's heading and the shortest straight line to each goal, the area of each goal, and whether there is a wall along the straight line between the ball and each goal.

From this analysis, we see that the CNN is learning about surface features of the trial such as how close the ball is to the goal, and whether there is a clear path to that goal (Table 2, columns 1 & 3). However, it does not use motion information about whether the ball is headed towards the goal, nor does it differentiate predictions based on other features such as the goal area (columns 2 & 4).

Heuristic	Distance To Goal	Heading Offset	Goal Area	Wall Between
Correct Goal	-0.650	0.001	0.109	-0.323
Other Goal	0.503	-0.042	0.079	0.110

Table 2: Correlations between the CNN's belief in the correct goal based on features of the trial relating to both goals.

4 Discussion

Our results indicate the CNNs are able to perform well on this task, achieving a near human performance level, but that a simple CNN cannot in general explain human predictions as well as noisy physical simulation can. However, this CNN still explains part of human judgments beyond physical simulation in the first couple seconds of prediction, suggesting that it is capturing something about fast "at-a-glance" predictions that simulation might be too slow to produce. These predictions do not involve physical simulation, instead relying on heuristics and statistical regularities of the trials.

This suggests that the CNN can learn fast heuristics (e.g., pick the closest goal) that people might rely on before simulation provides enough information to make a confident physical prediction; conversely, later in the trial, people will have had time to produce a set of noisy simulations that will be more concentrated on one of the goals, so people are therefore more likely to rely on physical simulations for their predictions.

In sum, we have demonstrated that a simple CNN can account specifically for the fast decisions humans make when faced with an ongoing process, whereas simulation explains later decisions. As a future direction, it seems promising to combine these feedforward models with the currently existing simulation-based physics ones, in order not only to harness the advantages of both approaches, but also to build a more realistic aggregate explanation of the diverse types of inferences humans employ while performing physics prediction tasks.

Acknowledgments

This work was supported by the Center for Brains, Minds & Machines, funded by NSF STC award CCF-1231216.

References

[1] 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.

[2] Adam N. Sanborn, Vikash K. Mansinghka, and Thomas L. Griffiths. Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411–437, 2013.

[3] Kevin A Smith, Eyal Dechter, Tenenbaum Joshua B, and Edward Vul. Physical predictions over time. In *Proceedings of the 35th annual meeting of the cognitive science society*, pages 1–6, 2013.

[4] Kevin Smith. Principles underlying human physical prediction. *eScholarship*, 2015. Chapter 4.

[5] Jessica Hamrick, Kevin A Smith, Thomas L Griffiths, and Edward Vul. Think again? Optimal mental simulation tracks problem difficulty. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX, 2015. Cognitive Science Society.

[6] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, 1982.

[7] Adam Lerer, Sam Gross, and Rob Fergus. Learning Physical Intuition of Block Towers by Example. *arXiv:1603.01312*, 2016.

[8] Renqiao Zhang, Jiajun Wu, Chengkai Zhang, William T. Freeman, and Joshua B. Tenenbaum. A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding. *arXiv:1605.01138*, 2016.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper With Convolutions. pages 1–9, 2015.

[10] Scott Lembcke. Chipmunk 2D Physics Engine. `https://chipmunk-physics.net/`.