Unlike Brains, Pre-trained CNNs do not Encode Current or Predicted Object Contact

RT Pramod (pramodrt@mit.edu)
McGovern Institute for Brain Research, MIT
Cambridge, MA - 02139, United States

Josh Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA - 02139, United States

Nancy Kanwisher (ngk@mit.edu)
McGovern Institute for Brain Research, MIT
Cambridge, MA - 02139, United States

Abstract

Interacting with the physical world requires predicting what will happen next, from catching a ball to stacking dishes to changing lanes in traffic. This ability in turn often hinges on representing contact relationships among objects, as the fate of two objects is intertwined when they are in contact. We found recently that the brain's hypothesized "physics network" both represents whether two objects are in contact and predicts future contact in simple scenarios. What computations underlie this ability? Might fast pattern recognition mechanisms like those found in convolutional neural networks (CNNs) suffice? To find out, we presented our same stimuli to CNNs pre-trained on object recognition (VGG-16) and action recognition (3D-ResNeXT-101). The scenario-invariant current and predicted object contact information we found in the brain could not be linearly extracted from these networks. Future work will test whether training on our scenarios and tasks may enable these networks to represent this information. Alternatively, the brain's ability to extract current and future contact information may depend on different computational mechanisms better captured by a structured generative model that runs approximate probabilistic simulations based on knowledge of physics and of the physical properties of the current scene, akin to those in video game engines.

Keywords: Intuitive Physics; Object Relations; CNN

Introduction

Planning actions entails predicting future world states, which in turn requires knowledge of how the world works (e.g., gravity) and information about the current scene such as object contact relationships (e.g., containment) which constrain future world states. If a container moves, so does its containee, but the same is not true for an object that only occludes another object without touching it. Underscoring their importance for physical scene understanding, object contact relationships emerge early in development (Hespos & Baillergeon, 2001; Baillergeon, Needham & DeVos, 1992; Spelke, Philips & Woodward, 1996), are privileged in language (e.g., *in* vs. *behind*; Hafri, Green & Firestone, 2023), are extracted quickly and automatically (Hafri & Firestone, 2021; Hafri et al. 2024), and are represented in brain regions implicated in physical reasoning (Pramod et al., *in prep*).

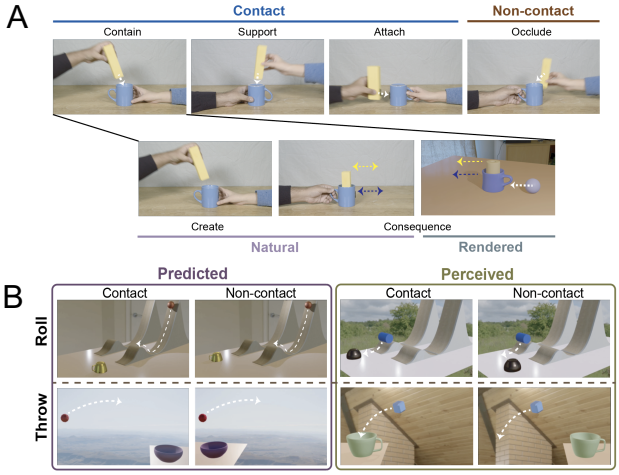

Figure 1: Example stimuli for (A) *Contact Detection* and (B) *Contact Prediction.* Arrows indicate motion trajectories of the respective objects.

What kind of computations underlie our abilities to perceive object contact and to predict what will happen next in a physical scene? Prior work has shown that CNNs can learn powerful representations that support not only object recognition, detection, segmentation, and retrieval but also accurate prediction of scene stability (Lerer, Gross & Fergus, 2016; Conwell, Doshi & Alvarez, 2019) and infant-like physical prediction behaviors (Piloto et. al., 2022). However, these studies have tested CNNs only within narrow domains (block towers) or by providing ground truth object masks and motion trajectories, so it is unclear if CNNs are useful

for physical reasoning in complex natural scenarios (see Pramod et al., 2022). On the other hand, human behavior in various physical reasoning tasks is well modeled by approximate probabilistic simulation in game-style physics engines (Battaglia, Hamrick & Tenenbaum, 2013; Zhang et al., 2016).

Here we test whether the representations learned when CNNs are trained on object classification (ImageNet) and action recognition (Kinetics) contain the scenario-invariant information about current and predicted object contact that we found recently in the brain (Pramod et al., in prep).

Methods

Stimuli: We used two sets of 1.5-s video clip stimuli, developed for our recent brain imaging study, which varied the presence and predictability of contact, orthogonally from scenario, objects, and motion (Fig 1). *Contact Detection:* This set had 4 object relationship types (Contain, Support, Attach and Occlude), each embedded in 3 different scenarios (Figure 1A). In the *Natural-Create* scenario, a hand placed an object relative to a base object to create the relationship. The *Natural-Consequence* scenario showed a hand moving the base object back and forth with the second object already in position to reveal the contingency between the two objects. The *Rendered* scenario revealed the contingency through a collision between the base object and a ball (with no hand visible). The Natural scenarios were filmed in-house with human actors, and the Rendered scenario was created using Blender. We created a total of 768 video clips across all conditions. *Contact Prediction:* 96 Blender-created video clips showed an agent and patient object varying orthogonally in a 2x2 design with event type (Contact or Non-contact) and condition (Perceived or Predicted) as factors (Figure 1B). The Perceived condition showed actual collision and non-collision events whereas the Predicted condition only showed partial trajectory of the agent object from which an imminent collision or a non-collision event could be predicted. These four conditions were each shown in three other orthogonally crossed dimensions: 2 scenarios (Roll or Throw), 6 background scenes (3 indoor and 3 outdoor), and two motion trajectories of the agent object (left or right).

Models: We extracted feature representations from the penultimate layer of the two pre-trained CNNs: VGG-16 trained on ImageNet object recognition, and 3D-ResNeXT-101 trained on Kinetics videos for action recognition. For VGG-16 we either averaged features across all frames of the video or used features for a single selected frame in each video that most clearly revealed the object-object relationship.

Contact Detection: We trained a linear SVM classifier on CNN features for contact vs. non-contact decoding on one scenario (e.g., Natural-Create) and tested on the remaining two scenarios. We randomly sampled stimuli uniformly from the 3 contact relations to match the sample size of the non-contact (i.e., occlude) relationship. We also tested contain vs. occlude decoding, as the two most visually similar conditions.
Contact Prediction: We trained a linear SVM classifier on CNN features for contact vs non-contact decoding on one of the scenarios (say, Roll) in the Perceived condition and tested it on the held-out scenario (say, Throw) in the Perceived condition (to test contact detection), and the Roll and Throw scenarios in the Predicted condition (to test contact prediction).

Table 1. Decoding accuracy for contact detection and prediction in both CNNs across scenarios.

	Contact Detection (Natural-Rendered)		Contact Prediction	
	Contact Vs. Noncontact	Contain Vs. Occlude	Perceived Vs. Perceived	Perceived Vs. Predicted
VGG-16	55.9 (54.2)	54.5 (54.5)	50	51.04
3D-Res NeXT101	59.4	48.2	45.8	53.1

Results

The CNN decoding results are summarized in Table 1. In both pre-trained VGG-16 and 3D-ResNeXT-101, contact detection accuracy was close to chance when trained on either of the Natural scenarios and tested on the Rendered scenario (and vice versa). We observed similar results for the more visually matched Contain vs. Occlude decoding. Since ImageNet pre-trained VGG-16 cannot take video inputs, we averaged features across all frames to obtain a single feature vector for each stimulus. However, because this procedure could blur critical features, we repeated the train and test procedure using feature representations for a single frame in each video clearly showing the underlying object relationship. Here too, we found near chance contact decoding accuracy (Table 1, *top row in parentheses*).

Contact decoding was at chance for both CNNs, both when trained and tested on Perceived conditions, and when trained on Perceived and tested on Predicted, indicating that they lack both generalizable representations of current contact, and the ability to predict future object contact.

Conclusions

Unlike human minds and brains, pretrained CNNs apparently do not extract scenario-invariant information about current or predicted object contact. Future work will test whether CNNs trained on this task, or generative models, better fit human data.

Acknowledgments

This research was funded through NIH EY033843 and NSF 2124136 (both to NK) and NSF STC Award CC-1231216 for the Center for Brains, Minds and Machines (CBMM). The authors would like to thank Elizabeth Mieczkowski and Cyn Fang for help with stimulus creation and fMRI data collection.

References

Baillargeon, R., Needham, A., & Devos, J. (1992). The development of young infants' intuitions about support. *Early Development and Parenting*, *1*(2), 69–78.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Conwell, C., Doshi, F., & Alvarez, G. (2019). Human-Like Judgments of Stability Emerge from Purely Perceptual Features: Evidence from Supervised and Unsupervised Deep Neural Networks. *Conference on Cognitive Computational Neuroscience*.

Hafri, A., & Firestone, C. (2021). The Perception of Relations. *Trends in Cognitive Sciences*, *25*(6), 475–492.

Hafri, A., Green, E. J., & Firestone, C. (2023). *Compositionality in visual perception*. Center for Open Science.

Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2024). A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing.

Hespos, S. J., & Baillargeon, R. (2001). Reasoning about containment events in very young infants. *Cognition*, *78*(3), 207–245.

Lerer, A., Gross, S., & Fergus, R. (2016). Learning Physical Intuition of Block Towers by Example. *Proceedings of the 33rd International Conference on Machine Learning*.

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267.

Pramod, R., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *eLife*, *11*.

Pramod, R., Mieczkowski, E., Fang, C., Tenenbaum, J. B., & Kanwisher, N. (*in prep*). Decoding current and future object contact in the human brain.

Spelke, E. S., Phillips, A., & Woodward, A. L. (1996). Infants' knowledge of object motion and human action. In *Causal Cognition* (pp. 44–78). Oxford University Press.

Zhang, R., Wu, J., Zhang, C., Freeman, W., & Tenenbaum, J. (2016). A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding. *arXiv Preprint arXiv:1605.01138*.