Modeling the Intuitive Physics of Stability Judgments Using Deep Hierarchical Convolutional Neural Networks

Colin Conwell (conwell@g.harvard.edu)
Department of Psychology, 33 Kirkland Street,
Cambridge, Massachusetts 02139

George A. Alvarez (alvarez@wjh.harvard.edu)
Department of Psychology, 33 Kirkland Street,
Cambridge, Massachusetts 02139

Abstract:

To gauge whether a tower of objects will fall, are visual heuristics sufficient? In this study, we explore the potential of pattern recognition as a viable model of intuitive physical inference in a task that requires observers to estimate the stability of stacked building blocks, comparing the performance of a deep feedforward convolutional neural network to human performance using psychophysics. In analyzing human and machine behavior alike, we identify a pair of image-based visual features that strongly predict both human and machine performance and differ only in the summary statistic used to compute them. Our results suggest that a system trained only to recognize patterns in visual input and given no explicit physical knowledge (e.g. mass, gravity, friction or elasticity) is nonetheless capable of approximating human judgments in a paradigmatic intuitive physics task.

Keywords: Intuitive physics; Deep neural networks; Psychophysics; Human-in-the-loop machine learning

Introduction

If ever you've played Jenga—or the more quotidian versions of Jenga that involve stacking dishes in an already overflowing sink and wrestling with seasonal decorations atop a teetering ladder—you've relied on a competence in physics you most likely weren't even aware you possessed. Navigating a world of complex objects in complex interactions requires a finely calibrated sense of the physical contingencies that define these interactions, a psychological savvy often referred to as 'intuitive physics'. While our capacity for physical inference is well-established by empiricism and Jenga alike, the computational architecture undergirding this capacity remains a matter of debate.

One class of theory proposes a system of structured knowledge and approximation, analogized as the sort of computer graphical game engine deployed in the realistic simulation of physical events in real time (Battaglia and Colleagues, 2013; Ullman and Colleagues, 2017), wherein explicitly physical knowledge is leveraged to explicitly model a diversity of physical scenarios.

An alternative to this "intuitive physics *engine*" is the sort of general purpose pattern matching model exhibited most conspicuously in deep neural networks (Lerer & Colleagues, 2016). These models reformulate physical inference as statistical inference: the extraction of trends in visual features that correlate with physical outcomes.

Here, we explore the potential of pattern matching in a paradigmatic intuitive physical inference task (the block towers task), supplementing previous comparisons (Zhang & Colleagues, 2016) with a more controlled set of stimuli that afford us greater insight into the representations undergirding both human and machine performance.

Methods

Stimulus Set Adapting a technique specified by Zhang and colleagues (2016), we generated an image dataset of stacked blocks, all of the same size ($1m^3$), with enough horizontal jitter in each block's position that towers have a 50/50 chance of falling. We varied the number of blocks from 2-5 with 200,000 images per tower size. The groundtruth for whether a tower will fall can be determined by computing for each junction of blocks the mean position (centroid) of all the blocks above the junction and comparing it to the centroid of the block beneath. If the centroid of the blocks above extends beyond the edge of the block beneath (at any junction), the tower will fall. Importantly, for both training and test sets we apply jitter only along one dimension (such that the variance is fully visible when facing the towers directly). In the training set, we allow some variance in the camera (see Zhang & colleagues (2016) for details), but for the test set situate the camera directly in front of the blocks, with the camera focused at the tower's center.

Hypothesized Features For each stimulus in the test set, we considered 12 hypothetical features a human or machine might use to accomplish the task. The features differ in whether they emphasize local information, or statistical information aggregated over multiple local measurements. These features are as follows:

- Configural deviation: the mean and max values for the distance of each block from the centroid of *all the blocks* above it — **the most direct approximation of groundtruth** in our set of features.
- Local (pairwise) deviation: the mean and max values for the distance of each block in the tower from the block above it, irrespective of other blocks.
- Global deviation: the mean and max values for the distance of each block from the centroid of the tower.
- Number of instabilities: the number of junctions in the tower shown by groundtruth calculations to be unstable.
- Percent unstable: the number of unstable junctions in the tower as a proportion of the total number of junctions.
- Horizontal extent: The horizontal distance from the right edge of the rightmost block in the tower to the left edge of the leftmost block in the tower: the tower's width.
- Vertical extent: The vertical distance from the bottom edge of the bottommost block to the upper edge of the uppermost block: the tower's height.
- Alignment distance: the numerically determined minimum distance each block must be moved to return the tower to a perfectly stable configuration, wherein each block is perfectly aligned with the others.
- Minimum distance to stability: the minimum each block must be moved to return the tower to a minimally stable configuration, wherein there are no unstable junctions.

Each feature is quantifiable in the sense that it constitutes some property of the visual array and differs across exemplars but contains no explicitly physical information (e.g. the mass of the blocks, or the force of gravity).

Results

We first benchmarked our test set on humans (via Amazon Mechanical Turk), with a two-choice forced alternative, asking whether the tower in a given image was stable or unstable. Human performance was generally high for any number of blocks in the range we tested (from 87% with 2 blocks, to 82% with 5 blocks).

Next, we tested the performance of a neural network (VGG16) pretrained on ImageNet, refitted with a binary classifier ('stable' or 'unstable') and finetuned with 100,000 labeled exemplars of three block configurations from the training set. Neural network performance was comparable to human performance; trained only on 3 block configurations, the machine performed well with test towers of the same size (91%), and generalized successfully to 2, 4 and 5 block configurations with only a minor decrement in performance (in chi-square tests comparing proportion of successes to chance, all p-values < .0001; 2 Blocks: $\chi^2(1) = 78.12$, 4 Blocks: $\chi^2(1) = 91.25$; 5 Blocks: $\chi^2(1) = 47.05$).

To assess the degree to which humans and machines agreed on which towers were stable and which were unstable, we compared the pattern of responses to each individual display (200 per tower size). Human agreement was quantified as the mean of a correlation computed separately for each individual subject (against the average response of the other subjects). A similar agreement humans and machines was computed by iteratively removing one subject from the pool and correlating the machine's results with the average of the pool remaining. The results manifest a high degree of agreement between human and machine across the individual displays, with an average intersubject correlation of .76 and average human to machine correlation of .66. The network's predictions correlate with human judgments as well as would a human's judgments if that human had the same overall performance as the network (see Figure 1B): In general, as an individual human's performance increased, their performance correlates more closely with the rest of the humans; the network's performance exhibits a similar trend.

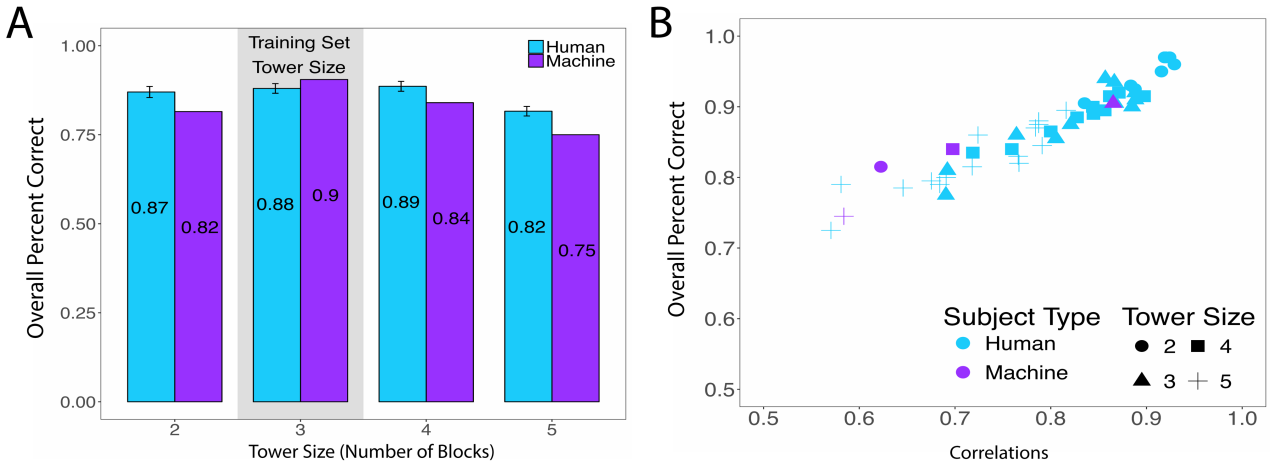

Figure 1 (A). *Human and Network Performance on the Block Towers Test Set. Error bars are 95% confidence intervals.* **(B)** *Correlations between humans and machine across overall percent correct.*

To gauge the features most predictive of performance for humans and machine alike, we used random forest variable importance metrics (Archer & Chimes, 2009), a method of analysis that allows us to highlight each feature's influence on behavior. The most important variable in terms of predicting human performance was a feature we labeled the 'maximum configural deviation'—the *maximum* horizontal distance between the centroid of any one block in the tower and the centroid of all blocks above it (mean decrease in accuracy of 19.49%; mean decrease in Gini of 814.96). Critically, this feature is a direct approximation of the groundtruth (see Methods), suggesting that humans place the most emphasis on the optimal information for judging tower stability. In contrast, the most important variable in terms of predicting machine performance was the 'mean configural deviation'—the *mean* horizontal distance between the centroid of any one block and in the tower and the centroid of all blocks it (mean decrease in accuracy of 40.63%; mean decrease in Gini of 106.14).

To illustrate that humans place greater emphasis on the optimal feature (the maximum configural deviation), we plot both the human and network responses as a function of the maximum configural deviation (see Figure 2). Each point on

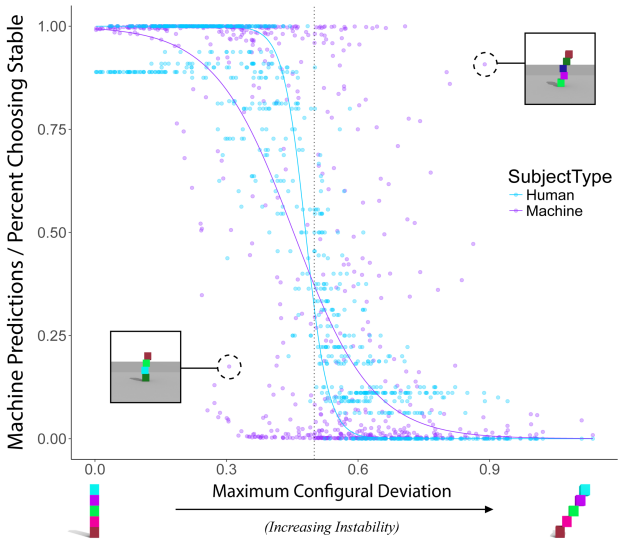

Figure 2. *Psychophysical curves of human and machine performance in the block towers task predicted by the maximum configural deviation. On the x axis is the maximum configural deviation, ranging from perfectly stable at the lower end to very unstable at the upper end. On the y axis is the machine's prediction or the proportion of subjects choosing 'stable'.*

the plot represents the response to an individual stimulus. For humans, we plot the average response across participants to each individual test image. For the neural network, we plot the neural network output, which corresponds to the probability that the tower is stable. Because the maximum configural deviation is an optimal feature for the block towers task, we are able to draw the groundtruth value directly on the x axis, with all stimuli to the left definitively stable and all values to the right definitively unstable. A perfect observer would produce a step function at this point. The shallower curve from the model fit on the output from the machine suggests the machine's responses were less acutely tuned to the maximum configural deviation overall.

Finally, as confirmation of the random forest variable importance measures, we fit binomial logistic regressions for each of the features calculated, regressing the human prediction (0 or 1) or machine's prediction (a value bounded at 0 or 1 based on the machine's confidence, and binarized by rounding) over the feature value, and computed the area under the curve of the receiver operating characteristic (abbreviated as AUROC; Hanley & McNeil, 1982). The mixed effects logistic regression of human judgments ('stable' or 'unstable') on the maximum configural deviation produced an AUROC of .944. The standard logistic regression of the machine's predictions on the mean configural deviation produced an AUROC of .945. These serve as confirmation that the most important feature designated by the random forest algorithm is also the best performing (the most 'sensitive') in terms of diagnosing whether the subject will label a given configuration as 'stable' or 'unstable'.

Discussion

Convolutional neural networks are quintessential pattern recognition systems—designed to find regularities in visual input that reliably associate with some prespecified output. That such a system is able to perform at or above human levels on the same block towers task suggests that the information inherent to the visual array is sufficient for categorizing various configurations of objects as stable or unstable—and that explicit physical knowledge (e.g. mass, gravity, friction and elasticity) are not obligatory when performing feats of physical inference. To say that the sort of pattern recognition engendered by neural networks is a viable model of intuitive physics (and not just a universal approximator), however, we must also evince some measure of similarity to human behavior, in cases of both error and accuracy. Here we have shown that human and machine make similar judgments of stability across exemplars, and that those judgments are predicted by similar features of the stimulus—features that could in principle be computed directly from the visual input. And crucially, though our task is markedly constrained, those *singular* parameters (mean and max configural deviation, respectively) account almost perfectly for the variance in performance across both our subject types. No explicitly physical model is required.

Equally crucial is this: When it comes to performing a task like the block towers task, not all features are created equal.

Attention to irrelevant dimensions of a stimulus (the color of the blocks, for example) will inevitably degrade performance unless that dimension serves as some indirect proxy of the groundtruth value. (Color in this case was almost certainly irrelevant—assigned randomly at rendering from a prespecified pallet). Out of a dozen features computed on every exemplar, from the relatively simple (local deviations) to the more sophisticated (numerically minimized alignment optima), the feature most predictive of human performance appears to be a feature that actually gives *direct* groundtruth access to the stability of the tower: The maximum configural deviation is optimal in the sense that it underscores the most unstable junction in the tower, which (if unstable past a certain threshold) determines the stability of the tower as a whole in terms of whether or not it will fall.

This is particularly consequential when we consider the performance of the machine. Unlike the human subjects, the machine's answers seem best predicted by the *mean* configural deviation, which—unlike the *maximum* configural deviation—does not give direct access to the groundtruth stability of the tower, since two or more deviations in the opposite direction are sometimes averaged in a way that wrongly suggests stability.

Nevertheless, that the machine's answers are indeed predicted by configural information suggests they are able to extract at least a portion of the overall pattern necessary to perform to par in the block towers task.

The discrepancy between human and machine provides a target for future improvement. By further investigating how people perform the task, we can further refine the network. For instance, one speculative possibility (if we interpret the maximum configural deviation a bit more intuitively as the most prominently disjointed block in the tower) is that the optimal feature is also the most salient feature in terms of capturing a human observer's attention. Humans may be benefiting from the ability to selectively attend to the optimal features of the visual array, filtering out irrelevant (and potentially) contradictory information. If this were the case, then equipping the machine to attend more specifically to various features of the tower may benefit the machine in equal measure, both in this task and in others.

Acknowledgments

Many thanks to Tim Menke for assistance in computing features, and to Talia Konkle for feedback.

References

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.

Ullman, T.D., Spelke, E., Battaglia, P., & Tenenbaum, J.B. (2017). Mind Games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9), 649

Zhang, R., Wu, J., Zhang, C., Freeman, W. T., & Tenenbaum, J. B. (2016). A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. arXiv preprint arXiv:1605.01138.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52(4), 2249-2260.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.