Human-Like Judgments of Stability Emerge from Purely Perceptual Features: Evidence from Supervised and Unsupervised Deep Neural Networks

Colin Conwell (conwell@g.harvard.edu)
Fenil Doshi (fenil_doshi@fas.harvard.edu)
George A. Alvarez (alvarez@wjh.harvard.edu)
Department of Psychology, 33 Kirkland Street,
Cambridge, Massachusetts 02139

Abstract

At a glance, the human visual system transforms complex retinal images into generic feature representations useful for guiding a wide range of flexible, efficient behaviors. In this report, we provide evidence that the feature representations that arise from purely feedforward neural networks are sufficient to explain seemingly high-level human judgments, such as how stable a tower of blocks appears to be. Using this now paradigmatic intuitive physics task as a case study, we attempt to linearly decode stability from the features of two deep neural networks – a supervised network trained on ImageNet, and a variational autoencoder trained only to reconstruct images of block towers from various perspectives – neither of which were ever taught stability per se. Decoding almost exclusively above chance in both cases, and with a classifier that produces responses virtually indistinguishable from human responses when trained on ImageNet features, our results demonstrate that systems designed mainly for pattern recognition, entirely void of explicit physical parameters and never trained on physics, nevertheless learn visual features that reliably undergird physical inference in the judgment of stability. More generally, these findings suggest that even seemingly high-level human physical reasoning may be grounded in a direct readout of basic perceptual feature representations.

Keywords: Intuitive physics; Deep neural networks; Unsupervised learning; Psychophysics

Introduction

Exposed for a fraction of a second to various visual stimuli, human subjects extract relatively massive amounts of surprisingly sophisticated information: the trustworthiness of a face, the centroid of complex shapes, and – directly relevant to our purposes in this paper – even the stability of a tower of randomly arranged blocks (Firestone & Scholl, 2016). In some cases, this extraction takes less than a 20th of a second. The speed at which we can decode these bits of meaning in general is taken as evidence that the decoding happens not at the level of some abstract cognitive process, but directly in the rapid cascade of perceptual processing that occurs immediately after the presentation of a stimulus. The modeling of this 'feedforward' processing – predominantly encapsulated from computations performed elsewhere in the brain – we increasingly entrust to deep neural networks trained to do one thing and one thing only: the nonlinear regression of raw inputs (pixels, waveforms, words and numbers) onto various predictors. This is a computation we might also call statistical pattern recognition.

A growing wealth of data suggests our trust in these models is not misplaced, and that the feature representations they learn correspond strikingly with biological reality (Yamins & DiCarlo, 2016). But are they useful as models of judgments that extend beyond the typical purview of sense-percepts? How do they fare in purportedly more complex domains?

Whether you're mastering levitation with a hydroflight jetpack or simply putting one foot in front of the other, 'intuitive physics' is your common sense of a physical world defined by contingencies, multidimensionality and often inscrutable latent causality – and it is precisely a domain in which one might expect the feature-based pattern recognition of neural networks to break down.

Given the complexity and latent structure of the physical world, predominant models of intuitive physics posit that our intuitive physical capabilities are at their core the product of a cognitive architecture that includes a more or less complete 'physics engine', akin to the kind deployed in video games and computer graphical animation (Battaglia and Colleagues, 2013; Ullman and Colleagues, 2017), innately equipped with effectively all the parameters necessary to perform complex simulations of physical scenarios in real time. Inference in this formulation is accomplished by iteratively and repeatedly sampling this simulator – primed by perception but powered exclusively by cognition.

In the current report, we explore an alternative model of intuitive physics based on the pattern recognition capabilities of deep neural networks, wherein physical inference is reformulated as a problem of identifying those perceptual features that serve as optimal proxies for the real physical properties that produce them. Previous work including our own (Conwell & Alvarez, CCN2018, Zhang & Colleagues, 2016; Lerer & Colleagues, 2016) has mainly focused on training networks end to end in a fully supervised fashion, developing features directly for physical targets. Here, we explore another possibility: that features learned by deep

neural networks trained for other tasks may nevertheless encode physically relevant properties that serve as the basis for physical inference.

Methods

To test this hypothesis, we fit linear classifiers of stability on the learned features of two model classes: a supervised neural network trained only to recognize object categories, and an unsupervised neural network trained only to reconstruct images of block towers (never with provision of the groundtruth stability). We compare these results to those of human observers to demonstrate that perceptual features are not only sufficient to mimic performance but may in fact serve as the foundation for human judgments of stability.

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 2-6. The groundtruth for whether a tower will fall can be determined by computing at 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. In one of two datasets we generated with this method (called 'Perspective'), we allow some variance in the camera. In the other (called 'Direct'), we situate the camera directly in front of the blocks, with the camera focused at the tower's center.

Behavioral Tasks Human subjects (from Amazon Mechanical Turk) performed two behavioral tasks: in the first, a benchmark, subjects were shown a series of towers and given a two-choice forced alternative task, designating each tower as stable or unstable. Tower sizes varied across subjects, but each subject only rated one size. The second task was identical to the first, but for two additional constraints: each stimulus was presented for only 250 milliseconds before being covered by a mask (a wall of blocks), and subjects were given only 1.5 seconds to respond. This design was meant to induce in subjects their instinctive 'gut' response, which in other domains has been analogized as the human equivalent of a 'feedforward' processing pass (Elsayed & Co., 2018).

Models & Modeling Tasks: For our supervised neural network, we used Resnet18 pretrained on ImageNet as a fixed feature extractor, freezing all the layers of the network but the batch normalization layers – a technique that maintains the integrity of the features learned by the convolutional and nonlinear filters of the network, but accounts for vacillations in the statistics of the image set currently being processed (Ioffe & Szegedy, 2015). For our unsupervised neural network, we used a variational autoencoder with a latent space of 128 dimensions, trained on the full range of block tower sizes rendered at various perspectives using a mean squared error reconstruction loss and a generative adversarial loss function (Makhzani & Co., 2015) in place of the standard Küllback-Liebler divergence, allowing the model to 'learn' the variational prior (a Gaussian) imposed on it. Importantly, and in contrast to other approaches that attempt to disentangle certain properties in the latent space using techniques like minibatch discrimination (Kulkarni & Co., 2015), we leave the latent space of our autoencoder fully entangled.

For both our supervised and unsupervised encoder models, we decode stability from features using a multilayer linear perceptron, trained with Adam optimization, and in the case of Resnet18 a cyclical learning rate deduced from search (Leslie, 2015). For any given size of tower, we held the process of feature extraction constant, but varied the process of linear decoding such that the classifier was always trained and tested on the same size of tower. Classifiers fit on both Resnet18 and the variational autoencoder were trained using features from 25,000 towers per tower size and tested on the benchmark towers of the same size rated by human subjects.

Results

Humans versus Resnet18 Features: Human performance was generally high for the full range of blocks in the range we tested (from 94% accuracy on 2 blocks to 79.8% accuracy on 6 blocks). The performance of the linear classifier trained on features from pretrained Resnet18 (henceforth 'Resnet Head') produced directly comparable performance (from 92.3% accuracy on 2 blocks to 78.6% on 6 blocks; see Figure 1A). A linear regression of performance on tower size unveiled a slight, but significant negative slope for both humans and Resnet Head (b = -0.023, $p < 0.01$ & b = -0.038, $p < 0.001$, respectively) – suggesting the difficulty of classifying increased with the size of the tower for both human and machine.

To assess the degree to which humans and the feature classifiers agreed on which towers were stable and which were unstable, we compared the pattern of responses to each individual display. 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 between 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 .726 and average human to machine correlation of .723, an insignificant difference. A graphical representation of these correlations (see Figure 1B) show the idiosyncrasies of comparison across individual subjects and individual models.

Variational Autoencoder Features: Overall performance was markedly lower for the linear classifier trained to decode stability from the features of the variational autoencoder. The linear classifier failed completely in the case of two blocks, scoring no greater than chance, and between 63.5% and 59% for the rest of the tower sizes. The failure of the classifier on this portion of the human benchmarks data, despite a relatively high validation score (80%) on a set of 1000 held-out images of two block towers reveals the heightened susceptibility of these classifiers to even slight differences in the distribution of features in the training and testing image set, perturbations it seems were amplified by representational divergences in the latent space of the autoencoder.

The lower performance of the classifier came as a surprise for an encoder model that seemingly had developed strong implicit representations of stability, as evidenced by interpolations in the model's latent space that produced smooth generative samples in the transition from an unstable tower to perfectly stable tower – an 'idealized' tower the network had never directly seen (see figure 2).

Inspired by these interpolations, we launched a set of exploratory analyses to determine whether or not we could close the gap between human subjects and the classifier trained on the features learned by the autoencoder.

The first analysis was predicated on the following hypothesis: in order to interpolate smoothly between tower exemplars, a model must develop some representation of each block's position in absolute space. When groundtruth stability can be linearly calculated from these positions (in a combination of averaging and thresholding operations), it's possible that the failure of the classifier to detect stability is not necessarily due to a paucity of discriminant features in the encoder model, but to the instability of the classifier itself, as may have been the case for the aforementioned failure on the two block subset of the human benchmark data. To test this, we

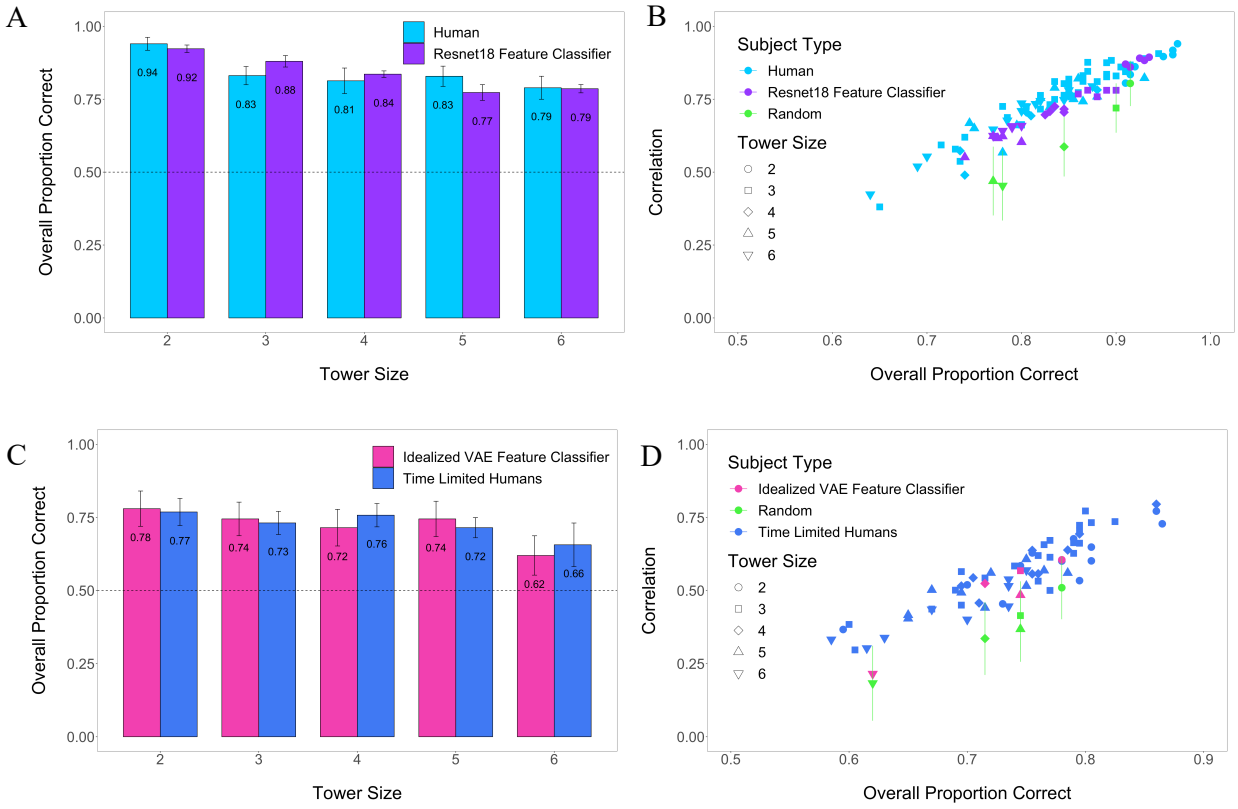

Figure 1 (A) & (C). *Human and Machine Performance on the Block Towers Human Benchmarks Dataset. Error bars are 95% confidence intervals for both humans & machines in (A) and bootstrapped confidence intervals for machines in (C).* **(B) & (D)** *Correlations between humans and machine across overall percent correct. This figure gives a sense for how well we might expect an individual classifier to correlate with human subjects given its accuracy. The green points are made with a random response generator pinned to the accuracy of the classifier, and the lines represent the bootstrapped confidence intervals on this point from many simulations. Notice that the discriminability between the random response generator and both human and machine subjects is best at low to intermediate ranges of performance.*

decided to incorporate the autoencoder into a modified ideal observer analysis, asking: what is the maximal classification accuracy a classifier trained on this network's features could achieve when asked to decode positions directly?

To do this, we modified the accuracy metric of the linear classifier to include the groundtruth computation of stability based on block position and equipped the classifier itself with a least absolute deviations loss on the predicted versus actual positions of blocks in the image. What this produced, in effect, was an ideal observer model of a classifier that told us how accurate the classifier could be given the representations of block position embedded in the latent feature space. Performance in our idealized 'position' classifier exceeded the performance of the classifier trained exclusively to decode stability at all tower sizes, scoring between 78% correct and 62% correct across the range of tower sizes. This discrepancy suggests that while the information for decoding stability may be present in the latent space of our autoencoder, the classifier trained only with stability labels failed to converge on this solution.

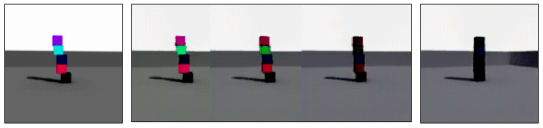

Figure 2. *Interpolations between an unstable tower and a perfectly stable tower in the latent space of a variational autoencoder never given stability labels. The image on the left was the only image not generated by the decoder.*

Time – Limited Humans versus Autoencoder Features: Human subjects limited to a 'feedforward' pass by time pressure (250 milliseconds worth of exposure to the stimulus and 1500 milliseconds of response time) still perform far above chance in the block towers task, scoring between 77% and 66% correct across the range of tower sizes.

Noticing that this decrease put humans roughly in the same performance range as the classifier trained on features in the autoencoder's latent space, we decided to test the correspondence of our time – limited human subjects to the performance of our classifier trained on the features of the autoencoder's latent space (see Figure 1C). The same average correlation analysis we performed with the classifier trained on Resnet18's features produced a notable correspondence between human and machine with some instances in which the machine correlated better with the average human subject than individual human subjects performing at the same level (see Figure 1D). The correspondence in this case may mean that human subjects limited by computational pressures may avail themselves of more specialized features (such as those learned by our autoencoder trained exclusively on block world) – but without further exploration, any such interpretation remains highly speculative.

Discussion

Our ease in navigating the physical world is a testament to a system that has learned to manage the inscrutability of latent physical structure, identifying reliable perceptual proxies to that structure even when that structure evades cognitive conceptualization. For a brain limited by sugar intake and five (variably reliable) senses, statistical regularity is often the most easily available route to reliable inference – a reality that explains perhaps the relatively late invention of classical mechanics by an otherwise very inventive species.

The success of our linear classifiers in decoding the stability of block towers from the features of supervised and unsupervised deep neural networks is not evidence that they have learned the same representations present in human perceptual systems: decades of cognitive science have shown those representations to be more rich, more flexible and more robust than the representations we have explored here. What the success of our linear classifiers does mean is that there exists some linear mapping between the purely perceptual representations learned by a deep neural network and the representations powering the inferences of human subjects in a task traditionally conceptualized as requiring a heavy dose of higher-order abstraction. While this work does not arbitrate on the capacity for such abstraction, it does suggest we may not always need it – and that statistical shortcuts via perceptual features may well trump fully fledged simulation in the pinch of computational pressure. All this to say, we may not always need physics to make physical inferences.

Future work will attempt to further complete the cartography of correspondence between human and machine by pushing and plying how we learn the representations we do, and probing why, despite immense divergences in the material substrate on which these algorithms are instantiated, the correspondences persist. The autoencoder we have included here – though in many ways an undercomplete example precisely because of its highly constrained, synthetic input space – is a nod to the necessity of rethinking how our perceptual systems are tuned, and what features they might develop in the process of the tuning. The more kaleidoscopic our representational palette, the more robust it is to the uncertainties and perturbations we invariably encounter, and the more conducive to a properly calibrated response.

References

Firestone, C., & Scholl, B. (2016). *JOV*, *16*(12), 689-689.
Battaglia, P.W., Hamrick, J.B., & Tenenbaum, J.B. (2013). PNAS *110*(45), 18327-18332.
Ullman, T.D., Spelke, E., Battaglia, P., & Tenenbaum, J.B. (2017). *TICS*, 21(9), 649
Yamins, D. L., & DiCarlo, J. J. (2016). *NatureNeuro. 19*(3), 356.
Zhang, R., Wu, J., Zhang, C., Freeman, W. T., & Tenenbaum, J. B. (2016). arXiv preprint arXiv:1605.01138.
Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). *NeurIPS* (3910-3920).
Kulkarni, T. D., Whitney, W. F., Kohli, P., & Tenenbaum, J. (2015). *NeurIPS*. (pp. 2539-2547).
Ioffe, S., & Szegedy, C. (2015). *arXiv preprint arXiv:1502.03167*.
Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). *arXiv preprint arXiv:1511.05644*.
Smith, L. N. (2015). *arXiv preprint arxiv:1506.01186*.