Title
Psychophysical-Score: A Behavioral Measure for Assessing the Biological Plausibility of Visual Recognition Models

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Richardwebster, Brandon
Dulay, Justin
DiFalco, Anthony
et al.

Publication Date
2023

Peer reviewed

Psychophysical-Score: A Behavioral Measure for Assessing the Biological Plausibility of Visual Recognition Models

Brandon RichardWebster
Kitware, Inc.

Justin Dulay
University of Notre Dame

Anthony DiFalco
SPRX

Elisabetta Caldesi
Microsoft

Walter J. Scheirer
University of Notre Dame

Abstract

For the last decade, convolutional neural networks (CNNs) have vastly superseded their predecessors in nearly all vision tasks in artificial intelligence, including object recognition. However, despite abundant advancements, they continue to pale in comparison to biological vision. This chasm has prompted the development of biologically-inspired models that have attempted to mimic the human visual system, primarily at a neural level, which is evaluated using standard dataset benchmarks. However, more work is needed to understand how these models perceive the visual world. This article proposes a state-of-the-art procedure that generates a new metric, Psychophysical-Score, which is grounded in visual psychophysics and is capable of reliably estimating perceptual responses across numerous models — representing a large range in complexity and biological inspiration. We perform the procedure on twelve models that vary in degree of biological inspiration and complexity, we compare the results against the aggregated results of 2,390 Amazon Mechanical Turk workers who together provided ~ 2.7 million perceptual responses. Each model's Psychophysical-Score is compared against the state-of-the-art neural activity-based metric, Brain-Score. Our study indicates that models with a high correlation to human perceptual behavior also have a high correlation with the corresponding neural activity.

Introduction

Could biologically-inspired artificial vision — that is to say, artificial vision possessing neural-like connections or a model of the dynamics of neural activity — be the next frontier for computer vision scientists? A growing community of neuroscientists, computer vision scientists, funding agencies, and participants in congressional hearings envision this as the next age of artificial vision (IARPA, n.d.; of Health: NIH, 2021; Dickson, 2021; Kasthuri, 2020). Regardless, few experts would be remiss enough to let the current struggles of modern artificial vision systems go unacknowledged. Take convolutional neural networks (CNNs) for example, CNNs require enormous numbers of input examples, easily become overfit, are sensitive to any deviation from the original input, and generally lack robustness to novelty within object classes (Marcus, 2018). In contrast, a young human child can be shown for the first time a cartoon image of a horse, and with a high degree of accuracy, after just one example, can recognize living horses, pictures of horses, and even cartoon horses — all of which can differ in breed, color, or artistic style (Guthrie, 1946). Human vision requires few examples, generalizes quickly, and is extremely tolerant to many of the conditions that negatively impact CNNs.

There are two major questions that computer vision scientists need to ask to move forward with the development of more sophisticated artificial object recognition systems that match or exceed human performance: (1) how are models constructed and (2) how does the evaluation of the model affect its construction? Let us begin with the latter. In the task of object recognition in computer vision, for roughly a decade now, progress has been defined and often constrained to rank-1 or rank-5 accuracy on benchmark datasets such as ImageNet (Russakovsky et al., 2015) (the one most frequently used), COCO (Lin et al., 2014), ShapeNet (Chang et al., 2015), and Open Images V6 (Pont-Tuset, 2020). Taken as one pool of data, the combination of images in ImageNet and COCO represents the largest publicly available collection of data for the study of object recognition in computer vision. But it still has the same underlying problem as all datasets. That is, it is a discrete representation of the real world. Thus, while the goal is to improve recognition rates in real-world applications using these dataset(s), they often do not reflect a one-to-one correspondence from the dataset to the real world. If the objective of a computer scientist is to produce models which demonstrate improvement on real-world tasks, then the pursuit of sophisticated object recognition models can be sidetracked by constraining evaluation to just the dataset.

Coming back to the former point, in the last half-decade, the vast majority of the progress made in computer vision has been made through more complex techniques in deep learning that a) allow for deeper models under the current limits of GPU capacity (with the assumption that deeper models perform better), and b) train more quickly so more input data can be ingested. It is a routine occurrence that with a new generation of GPUs comes a new record performance on traditional metrics. But does the optimization of models on hardware and data equate in the long term to better or more sophisticated artificial object recognition? Not likely. Instead, we need to develop ways to circumvent the constraints of hardware such as small models trained to recognize low-level features like human observers which could improve model generalization.

To help overcome this challenge, we propose a new procedure that generates a new metric, Psychophysical-Score, for examining perceptual similarities between human observers and machine observers. Psychophysical-Score isn't just a metric that helps us quantifiably represent *what* an observer

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

— whether human or machine — perceives but also *why* the observer perceives it that way. Beyond the why, it also could be used to teach a machine observer *how* to perceive like a human observer perceives the *what*. The *how* is particularly possible because Psychophysical-Score doesn't just measure how observers differentiate between objects but also what lies between those objects, the interspace.

We perform the procedure on seven traditional convolutional neural networks (AlexNet, GoogleNet, ResNet, SqueezeNetv1.0, SqueezeNetv1.1, VGG-16, and VGG-19) (Krizhevsky, 2014; Szegedy et al., 2015; He, Zhang, Ren, & Sun, 2016; Iandola et al., 2016; Simonyan & Zisserman, 2015) and five biologically inspired models (Gabor, HMAX, HT-L3, PredNet, and CORnet) (Fogel & Sagi, 1989; Riesenhuber & Poggio, 1999; Cox & Pinto, 2011; Lotter, Kreiman, & Cox, 2017; Kubilius et al., 2019), and compare the behavior to 2,390 Amazon Mechanical Turk workers who contributed ~ 2.7 million perceptual responses. We then perform an analysis that compares Psychophysical-Score results with the published results of the Brain-Score metric (Schrimpf et al., 2020). Our study indicates models which have a high correlation with human perceptual behavior (a high Psychophysical-Score) also have a high correlation with the corresponding neural activity (a high Brain-Score). The positive relationship between neural activity and perceptual behavior indicates a much-needed shift in research from traditional machine learning to machine learning derived from biology — perceptual behavior, neural activity, or both.

Related Work

Recent developments between scientists in machine learning, neuroscience, and psychology often use artificial neural networks as a model of the brain. Increasingly, there is a pertinent need to instill interpretability in neural models (Olah et al., 2018). Lotter et al. (Lotter et al., 2017) published PredNet, a biologically inspired predictive coding model which displays similar perceptual behavior to that of human observers on the visual phenomena of the flash-lag effect and end-stopping. Kubilius et al. (Kubilius et al., 2019) introduced a model called CORnet-S, which they report to have similar activation patterns to the neural activity of monkeys. And the most recent addition to artificial neural network models, though not a model by itself, VOneNet (Dapello et al., 2020), attempts to model the V1 primary visual cortex of primates. When VOneNet is prepended to the input layer of other artificial neural networks, it is shown to improve protection against white-box adversarial attacks. Each of these recent models represents advancements in both the understanding of biological vision and the quantitative approaches used in artificial neural networks.

An alternative to expensive and often difficult-to-obtain physiological recordings which were part of these models is a psychological method called psychophysics. Psychophysics is a well-established century-old set of methods for studying the relationship between stimuli and perceptual behavior. In computer vision, psychophysics has been applied to assess whether or not models with claimed high biological fidelity have perceptual behavior that also corresponds to biology. Rajalingham et al. (Rajalingham et al., 2018a) compare the recognition behavior of monkeys and people with CNNs, and report that CNNs did not represent the perceptual behavioral patterns of primates. Similar studies have observed the same disparities (Gerhard, Wichmann, & Bethge, 2013; Heath, Sarkar, Sanocki, & Bowyer, 1996; Eberhardt, Cader, & Serre, 2016). Concerning psychophysics applied to computer vision algorithms specifically, PsyPhy was introduced by RichardWebster et al. (RichardWebster, Anthony, & Scheirer, 2018; RichardWebster, Kwon, Clarizio, Anthony, & Scheirer, 2018a). PsyPhy facilitates a psychophysical analysis for object recognition and face recognition through the use of item-response theory (Embretson & Reise, 2000). An example finding from RichardWebster et al. (RichardWebster, Anthony, & Scheirer, 2018; RichardWebster, Kwon, Clarizio, Anthony, & Scheirer, 2018b) that is representative of this type of analysis is that there is behavioral consistency between deep learning algorithms and human observers when a blur was applied to input stimuli and in low contrast.

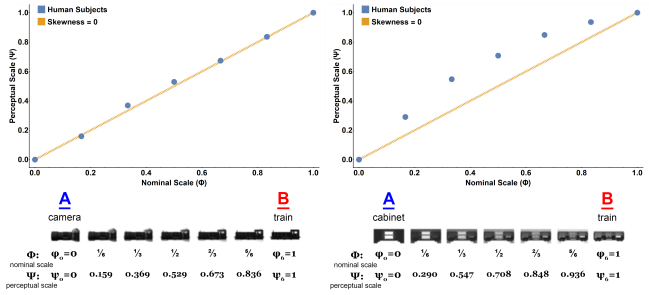
Some other datasets demonstrate low-level embeddings of human behaviors within machine learning models, but they focus on concept maps (Hebart et al., 2019), object structure (Peterson, Abbott, & Griffiths, 2018), and classification loss functions (Roads & Love, 2021). This paper focuses more intently on assessing the behavioral plausibility of machine learning models with respect to human intentions as opposed to observing similar structures to approaches.

Methods

Maximum-likelihood Difference Scaling (MLDS): To understand the interspace between two objects, whether it's the neurological representation in a human observer or the numerical representation in a computer vision observer, we build on a technique called Maximum-likelihood Difference Scaling (MLDS) developed by Maloney et al., 2003 (Maloney & Yang, 2003). MLDS is a technique to estimate the parameters of a stochastic model of perceptual differences based on an error measurement. Let us consider Figs. 1a & 1b.

Figs. 1a & 1b each contain a sequence of 7 images where the leftmost image contains an object A, the rightmost object B, and the inner 5 images each contain a composite of A and B. Each image has a corresponding nominal value denoted as ϕ_i where $\Phi = \phi_0, \phi_1, ..., \phi_6$ represents the nominal scale for the entire sequence. Recognize that the nominal scale, Φ, is not how observers perceive the differences in the images of the sequence — although it can be. Figs. 1a & 1b, as will be the case for all sequences presented in this article, has a $\Phi = 0, \frac{1}{6}, ..., 1$.

The goal of MLDS is to estimate the parameters of the stochastic model of perceptual differences, Ψ. A perceptual



(a) Ψ with no skewness (b) Ψ with positive skewness

Figure 1: Two plots showing the differences in an unskewed (a) and positively skewed (b) perceptual scale, Ψ. Underneath are the object sequences that correspond to the estimated perceptual scale. Φ is the nominal value/ratio to blend the intermediate objects in the sequence. See Methods Sect. for complete details.

scale represents the estimated suprathresholds of perceived distances through the repeated trials in the two-alternative forced choice (2AFC) where an observer is instructed to select the pair (out of two) which is most similar in structure and texture. The stimuli change over time, so the selection of their similarity is dependent upon temporal change. This choice is repeated numerous times to generate the complete perceptual scale of the sequence.

Stimuli generation: To generate the stimuli for a perceptual scale we used a software framework ShapeNet which contains 55 ImageNet classes. There are anywhere from tens to over thousands of instances for each of the 55 classes. There are two ways to define your view-port — the perspective from which the virtual camera is pointed at an object — in ShapeNet: panorama and canonical. Panorama allows the user to define the number of steps to take in 360° of rotation around a selected axis (x, y, z) and direction ($-$ or $+$), and canonical is a selection of six view-ports one for each combination of axis and direction. To simplify our selection and reduce the sheer number of possible generated images (from infinite possibilities), we chose to use the built-in canonical view-port selection.

After we proceeded to generate every object instance with their respective six view-ports, we needed to determine which instance pairs (a and b) had adequate overlap before forming a composite of the pair. The higher the overlap in the instance pair, the more significant the object features of the pair will be blended. If the overlap is low, blending of background and object will occur instead and cause the object to appear to fade instead of blending with its pair. To determine which instance pairs we can use for creating composites, we created a mask for each object and computed the Jaccard Index,

$$\mathrm{J}(a,b) = \frac{|a \cap b|}{|a \cup b|} \tag{1}$$

on the corresponding pair to determine the intersection between a and b. The Jaccard Index ranges from $[0,1]$, or no overlap to perfect overlap, respectively. For every instance pair in ShapeNet, we computed the Jaccard index, then selected the 10 pairs — for each instance pair — which had the highest Jaccard index. When the Jaccard index was high enough that the top 10 pairs had no statistically significant difference, 10 pairs were randomly selected from that set. Combined with the six canonical perspectives, we started with an initial set of 89.1k unique pairs. We then generated a sequence for each pair of objects.

Before creating a composite image of (a,b), we needed to ensure each image contained primarily low-level vision features such as edges, gradients, and local morphology. First, every image is converted to grayscale using the average of the RGB image channels, $\frac{R+B+G}{3}$. Then a Gaussian blur with $\sigma = 3$ is applied to the image to deteriorate higher features. At this point, each image is a grayscale image with each object looking closer to an unrecognizable blob than a known object in the 55 ImageNet classes.

At this point, each pair was blended using Alpha blending, $A(a',b',\alpha) = a'(1-\alpha) + b'(\alpha)$ where a' and b' are the modified source images, and α is in the range $[0,1]$ representing the percentage to include from image b'. To generate a sequence, the nominal scale $\Phi = 0, \frac{1}{6}, ..., 1$ is used as input to generate 7 composite images of pair (a,b).

Relating to our terminology, there are two types of sequences: class-level sequences and instance-level sequences. An instance-level sequence is a sequence using pair of images, (a,b), where $a \in A$ and $b \in B$, and A and B are two sets of instances for two classes. An instance-level sequence has 7 composite images but no perceptual scale — for our experiments — but a class-sequence has no composite images but has a perceptual scale. A class-level sequence is a set of all instance-level sequences aggregated together using MLDS to create the perceptual scale. In total there were 1450 class-level perceptual scales aggregated. Henceforth, we will use the notation $\Psi(A,B)$ to specify the perceptual scale for class (A,B).

Human Observer Experiments: A human baseline must be established to analyze how machine learning models compared to their human counterparts. While (Maloney & Yang, 2003) used a traditional approach to psychophysical experimentation with in-the-lab monitoring of participants, we use a modern crowd-sourced approach, called Amazon Mechanical Turk (AMT). In all, we had 2,390 participants submitting $\sim$ 2.7m perceptual responses. Rather than one individual responding to different instances of the same object sequence to obtain the perceptual scale — as would have been done in a lab setting under the original MLDS paradigm — we feed all of the participant responses to the stimuli into the MLDS optimization routine. Instead of having the set of suprathresholds that might be of one person, we have the mean set of suprathresholds for the population. Due to the vast number of stimuli and needed responses, this was the only way to get the estimations we needed. In this respect, Psychophysical-Score offers a superior approach to Brain-Score because of

the difficulty in obtaining large numbers of cortical measurements from subjects.

Skewness: Perceptual scales appear different for each pair; however, we needed a way to quantitatively compare the 1450 different perceptual scales. We developed a method to measure the skewness of the perceptual space based on the skewness of probability distributions. Recall that the skewness of a probability distribution is a measurement of where the probability mass is located to its tail. For example, in a normal distribution, the probability mass is precisely in the center — the mean, median, and mode are equal — with symmetric tails. In a right-tailed positively skewed distribution, the probability mass is closer to zero and the mode $<$ median $<$ mean. A left-tailed negatively skewed distribution has the opposite with mean $>$ median $>$ mode.

Since a perceptual scale has a few similar properties of a probability distribution, we can use the idea of skewness to generate a similar skewness for perceptual scales. However, one fundamental difference is that a perceptual scale is discrete, which poses complications to using previously known skewness calculations. One option to overcome this challenge would be to use a best-fit approximation using a non-symmetric distribution such as the Beta distribution, but this comes at the cost of introducing approximation errors. Alternatively, we can use a discrete integral to approximate the area under the curve and then rescale to match the standard format of skewness $(-\infty, \infty)$:

$$\mathrm{S}(\Psi) := -2 * \left(\frac{(\sum_{i=0}^{6} \psi_i) - 1}{5} - \frac{1}{2} \right) \tag{2}$$

If the $\sum \psi \simeq \frac{1}{2}$, then $\mathrm{S}(\Psi) \simeq 0$, and is considered to have a neutral interspace. An object sequence with a neutral interspace can be seen in Fig. 1a. If $\sum \psi > \frac{1}{2}$, then $\mathrm{S}(\Psi) > 0$, which is positive skewness and a positively skewed interspace (see Fig. 1b). Finally, if $\sum \psi < \frac{1}{2}$, then $P(\psi) < 0$ has negatively skewed interspace. Given the skewness value for each perceptual scale, we define Psychophysical-Score as the ranked-correlation, monotonic relationship, between a set of perceptual scales measured from humans and those measured from a machine learning model.

Experiments

What did we learn about human vision? The original sequences used in Meloney et al. (Maloney & Yang, 2003) demonstrated the utility of the MLDS procedure on an ordered sequence of color patches, so the first thing that we needed to determine was if humans could even perform the task with sequences that were significantly more complex (see Fig. 3). Meloney et al. describe two axiomatic validations an observer must meet for the procedure to work. The first test is that an observer must be able to order the sequence given the complete sequence, and second, meet what is called the Six-Point Property. The first test is a given. The second test is a subset of the first: observers must be able to order any random three images in the sequence (see (Maloney & Yang, 2003) for complete details). Out of the 1,485 possible sequences using all 55 objects in ShapeNet, only two sequences did not fully qualify (see Figs. 3(1a) & 3(1b)).

While these two tests indicate observers can properly perform the task for individual sequences, it does not indicate whether the measured suprethresholds for the perceptual space are an adequate representation of the population — i.e., are the perceptual spaces more than just the product of chance? To determine this, we perform a Pearson's Chi-Squared Test on the null hypothesis, h_0, that the perceptual spaces are the product of chance. Each space is represented as a quantitative value, what we call the skewness of the perceptual space (see Methods). The null hypothesis is unequivocally rejected with a $p \cong 0$ and $\alpha < 0.001$. Thus, h_1, is true, and the human-obtained perceptual scales are not randomly sampled signals.

Thus the results so far indicate 1) human observers can perform the task, and 2) the scales that result from the observations mean something beyond the noise.

The skewness for each perceptual scale can be seen from a high-level perspective in the "Human" plot in Fig. 2 and Table 1. One potential complication in analyzing variance is that AMT presents an uncontrolled environment, which naturally introduces extraneous factors in observer responses. However, even in a controlled environment, there will be variance among observers. To analyze the results of the AMT participants, and consistent with other findings (Germine, Duchaine, & Nakayama, 2011; Crump, McDonnell, & Gureckis, 2013), we assume the variance in perception has been accurately captured.

Figure 2: Each skewness, $\mathrm{S}(\Psi)$, of a perceptual scale, Ψ, plotted for three observers: Humans, Gabor, and CORnet-S. The red lines indicate two standard deviations for the observer. Notice the two orders of magnitude difference between the human observers and the CORnet-S observer. Gabor filters have roughly half the amount of variance and have near-random chance level variance. See Table 1 for a complete list of variances by model, and the Results section for further analysis.

A possible reason for the variance is the varying quality of the object sequences — not all sequences blend well. For an estimate of the quality of a blend, we can use the human noise factor (HNF) that is estimated along with a perceptual scale using MLDS (see (Maloney & Yang, 2003) for complete details). As with perceptual scales, in (Maloney & Yang, 2003), the HNF is for one observer. But for this experiment, it represents the population's noise factor. While we use HNF for blend quality analysis for human perception, the

HNF has no bearing on the described procedure to calculate Psychophysical-Score.

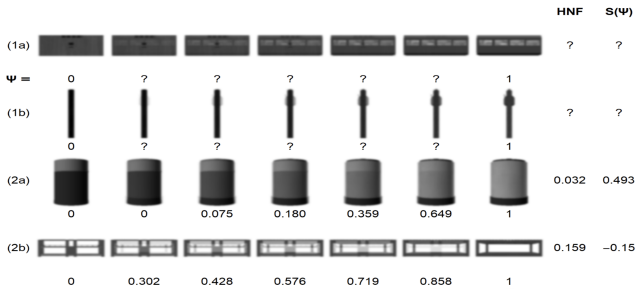


Figure 3: Six object sequences with their corresponding perceptual scale derived from human observers, Ψ. The two columns on the right give the value of the human noise factor (HNF) and $S(\Psi)$, the skewness of Ψ. See Results section for the description of how each sequence 1-4 differs and corresponding analyses.

Human-Machine Learning Comparison: How do human observers compare to machine learning models? We perform the 2AFC procedure on seven traditional CNNs (AlexNet, GoogleNet, ResNet, SqueezeNetv1.0, SqueezeNetv1.1, VGG-16, and VGG-19) (Krizhevsky, 2014; Szegedy et al., 2015; He et al., 2016; Iandola et al., 2016; Simonyan & Zisserman, 2015) and five biologically inspired models (Gabor, HMAX, HT-L3, PredNet, and CORnet) (Fogel & Sagi, 1989; Riesenhuber & Poggio, 1999; Cox & Pinto, 2011; Lotter et al., 2017; Kubilius et al., 2019), each representing significant advancements at the time of their publication. The procedure for humans differs from machine observers in two subtle ways: 1) we use many human observers whereas each machine observer is its own, and 2) machine observers use ℓ_2 distance for their difference measurement — the mechanism by which human observers assess distance is unknown. While procedural differences are minimal, we did make one substantial change to each machine observer that differs from their published form. Before performing the procedure, each machine observer was modified to better match our intentions of measuring low-level vision. In general, the later layers of each model were removed, which are thought to correspond with the later part of the human visual cortex (Rajalingham et al., 2018b).

Each model performs our proposed procedure so that we can compare scales from human observers to machine observers. Looking at the results in Table 1, we can see a substantial difference between human and machine performance. Human observers have a larger variance compared to random chance, while all machine observers have variance less than random chance. This is consistent with what one should expect given human perceptual scales are aggregate. However, Gabor filters standout because they are close to random chance, an order of magnitude higher than any other machine observer (Table 1). Something that differs with Gabor filters that is not present in the other models is that Gabor filters are not trained with data. In this respect, Gabor filters differ the most from human observers as well because humans have learn something visual throughout their lives. Although all of the learned machine models had a variance which was an order of magnitude smaller, it is possible that had each of these models been trained hundreds of times separately to better replicate the crowd-sourcing procedure, the variance would be higher. However, given this and the way the perceptual scales are acquired, we don't believe these fluctuations would be substantial.

Table 1: Ordered List of the variances, σ^2, in the skewness, $S(\Psi)$

-	σ^2
Human	0.06154
Random	0.03047
Gabor	0.02704
HT-L3	0.00622
HMAX	0.00607
VGG-19	0.00314
AlexNet	0.00302
SqueezeNet v1.0	0.00301
VGG-16	0.00264
CORnet-S	0.00250
GoogleNet	0.00230
ResNet-18	0.00196
SqueezeNet	0.00133
PredNet	0.00082

In each of the trained machine observers, even the biological ones, the classification layer treats each class as equally important. We don't know yet how human observers comparatively represent objects, but from the results, it is clear that different objects have different representational values. For machine observers, since each object has equal representational value, the skewness of the perceptual space is only present due to the loss in the training regime — the training regime forces skewness towards zero but only succeeds if there is no error. This is a huge difference between machine and human observers and something the computer vision community needs to consider.

Comparison to Brain-Score: Next we analyze whether or not our procedure produces results that align with the Brain-Score (Schrimpf et al., 2020) metric, which is a measurement of a model's biological fidelity from the perspective of neural activity. If a computer vision model has a high Brain-Score, it highly correlates with the neural activity of monkeys, but just because a machine observer has a high correlation, does it mean that the machine observer is "seeing" like a human observer? To answer this, we look at the monotonic relationship between the perceptual scales of a machine observer to the perceptual scales of human observers using Spearman's Rho. The correlation of all results can be seen in Fig. 4. Due to the symmetric property of perceptual scales, we represent them as either absolute correlation or its Psychophysical-Score.

The Psychophysical-Score for each machine observer is then compared to its published Brain-Score on `www.brain-score.org`. When the correlation results of the

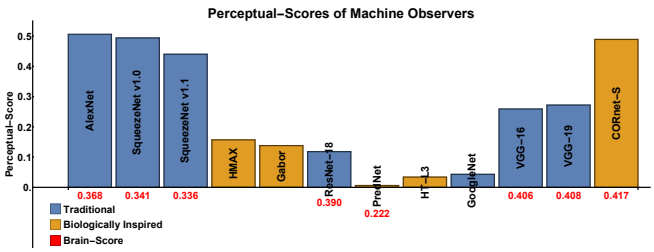


Figure 4: A visualization of the Psychophysical-Scores for each machine observer and each corresponding Brain-Score. Some models do not have a Brain-Score because they are not published on `www.brain-score.org`. Note the "U-shape" for both Psychophysical-Scores and Brain-Scores, where high Brain-Scores and Psychophysical-Scores are correlated. One outlier exists, ResNet-18. See Results for further analysis of individual Psychophysical-Scores and Brain-Scores for these models.

machine observers are compared as in Fig. 4, it becomes immediately apparent that Psychophysical-Score and Brain-Score are converging on similar observations, i.e., models which have a higher Psychophysical-Score also have a higher Brain-Score, and vice versa. However, ResNet remains an outlier. A potential reason for this deviation might be that when the later layers were removed, the skip connections which are only present in ResNet, were also removed. Skip connections provide early information to later layers by bypassing intermediate layers. However, the removal of these skip connections may have created a bifurcation in corresponding neural activity (includes the skip connections) and perceptual behavior.

PredNet is a particularly interesting case because it is biologically inspired in design but performs near the bottom on Brain-Score and has $\sim$0 as a Psychophysical-Score. PredNet was designed to operationalize a complex predictive coding model of visual processing. This biological mechanism may differ from the neural activity aligned with object recognition, causing a lower Brain-Score. However, this doesn't explain why its Psychophysical-Score is so low. We raise two possibilities for this: 1) by removing all but one predictive block, it may have eliminated quality error propagation (the output before ℓ_2 distance is computed), or 2) behavioral results may differ depending on which of the next predicted frames from the model are used — we used the first predicted frame.

Finally, and most importantly, the highest scoring model published on `www.brain-score.org`, CORnet, also scores high on Psychophysical-Score (although not the highest, there is no statistical difference when compared to AlexNet). CORnet-S is designed for and trained with Brain-Score, which is derived from the neural activity of monkeys, so one would expect CORnet-S to perform high on Brain-Score. But even more, since CORnet-S also has a high Psychophysical-Score, it indicates CORnet-S may have even more similarity to biology than initially considered. First, the neural activity recorded is representing authentic perceptual behavior to some degree. Second, using Brain-Score's representation of neural activity to train a model has significant potential utility. And third, Psychophysical-Score can be used to predict whether a model will have the potential to score high for neural activity, without the need for a difficult and expensive acquisition process — though the relationship is correlative, not causal.

Discussion

The capability of human vision has long inspired the pursuit of sophisticated artificial vision systems — generations of computer vision scientists have worked towards this end. The early advancements in artificial vision were often in tandem with an improved understanding of human visual systems. Unfortunately, the deep learning age has allowed computer vision scientists to lose sight of this original goal of mimicking human visual systems. Thankfully, scientists such as (Schrimpf et al., 2020; Rajalingham et al., 2018a; Gerhard et al., 2013; Heath et al., 1996; Eberhardt et al., 2016; Kubilius et al., 2019; Lotter et al., 2017; Dapello et al., 2020; RichardWebster, Anthony, & Scheirer, 2018) have returned to the grassroots of artificial vision. Arguably, the Brain-Score metric published by Schrimpf et al. (Schrimpf et al., 2020) is currently the best metric scientists have at estimating the relationship between modern vision algorithms and the neural activity of the brain. Even on its own, Brain-Score has shown to have huge potential (Kubilius et al., 2019). But now that Psychophysical-Score can complement Brain-Score by estimating the relationship between the perception of machine and human observers, the opportunity for the development of new biologically-inspired algorithms is apparent.

Psychophysical-Score and Brain-Score have the unique ability to inform scientists at every scale of artificial vision algorithm development. Take for example something as simple as the skip connections in ResNet-18. Previously it has been hypothesized that they model the pyramidal neurons in the cerebral cortex. But if Brain-Score and Psychophysical-Score differs so greatly, we should ask questions such as: does this really model a component of the brain? Did we misunderstand that component of the brain? Or is there a missing piece we don't know about? While previously questions like these have been asked rhetorically, they can now be asked with the hope of at least obtaining a partial answer. Instead of artificial models learning *what* to perceive from input images, using metrics like Psychophysical-Score and Brain-Score, a model can also learn *how* to perceive directly from human perception and neural activity. If a model can learn how to perceive, maybe it can learn from humans how to do exceedingly difficult tasks such as one-shot learning, generalization, or behavioral prediction.

References

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... Yu, F. (2015). *ShapeNet: An Information-Rich 3D Model Repository* (Tech. Rep.

No. arXiv:1512.03012 [cs.GR]). Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Cox, D. D., & Pinto, N. (2011). Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *Ieee fg*.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, 03). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), 1-18. Retrieved from https://doi.org/10.1371/journal.pone.0057410 doi: 10.1371/journal.pone.0057410

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2020/06/17/2020.06.16.154542 doi: 10.1101/2020.06.16.154542

Dickson, B. (2021). *Is neuroscience the key to protecting ai from adversarial attacks? — venturebeat.* https://venturebeat.com/2021/01/08/is-neuroscience-the-key-to-protecting-ai-from-adversarial-attacks/. ((Accessed on 01/25/2021))

Eberhardt, S., Cader, J., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In *Nips*.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Fogel, I., & Sagi, D. (1989). Gabor filters as texture discriminator. *Biological cybernetics*, *61*(2), 103–113.

Gerhard, H. E., Wichmann, F. A., & Bethge, M. (2013). How sensitive is the human visual system to the local statistics of natural images? *PLoS Computational Biology*, *9*(1), e1002873.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201–210.

Guthrie, E. R. (1946). Psychological facts and psychological theory. *Psychological Bulletin*, *43*(1), 1.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778). doi: 10.1109/CVPR.2016.90

Heath, M., Sarkar, S., Sanocki, T., & Bowyer, K. (1996). Comparison of edge detectors: a methodology and initial study. In *Computer vision and pattern recognition, 1996. proceedings cvpr'96, 1996 ieee computer society conference on* (pp. 143–148).

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

IARPA, I. A. R. P. A. (n.d.). *Microns.* https://www.iarpa.gov/index.php/research-programs/microns ((Accessed on 01/25/2021))

Kasthuri, N. (2020). *Hhrg-116-ap10-wstate-kasthurin-20200205.pdf.* https://docs.house.gov/meetings/AP/AP10/20200205/110447/H ((Accessed on 01/25/2021))

Krizhevsky, A. (2014). *One weird trick for parallelizing convolutional neural networks.*

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in neural information processing systems* (Vol. 32, pp. 12805–12816). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/7813d1590d

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Lotter, W., Kreiman, G., & Cox, D. (2017). Deep predictive coding networks for video prediction and unsupervised learning. In *Proceedings of the International Conference on Learning Representations*.

Maloney, L. T., & Yang, J. N. (2003, 10). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 5-5. Retrieved from https://doi.org/10.1167/3.8.5 doi: 10.1167/3.8.5

Marcus, G. (2018). Deep learning: A critical appraisal. *CoRR*, *abs/1801.00631*. Retrieved from http://arxiv.org/abs/1801.00631

of Health: NIH, N. I. (2021). *Brain initiative.* https://braininitiative.nih.gov/. ((Accessed on 01/25/2021))

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), e10.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Pont-Tuset, J. (2020). *Google ai blog: Open images v6 — now featuring localized narratives.* https://ai.googleblog.com/2020/02/open-images-v6-now-feat ((Accessed on 01/25/2021))

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018a). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *bioRxiv*, 240614.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018b, 02/2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep ar-

tificial neural networks [preprint]. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/10.1101/240614v4.full.pdf doi: https://doi.org/10.1101/240614

RichardWebster, B., Anthony, S., & Scheirer, W. (2018). Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1. doi: 10.1109/TPAMI.2018.2849989

RichardWebster, B., Kwon, S. Y., Clarizio, C., Anthony, S. E., & Scheirer, W. J. (2018a). Visual psychophysics for making face recognition algorithms more explainable. *European Conference on Computer Vision 2018*, *15*. Retrieved from https://par.nsf.gov/biblio/10144516 doi: 10.1007/978-3-030-01267-0

RichardWebster, B., Kwon, S. Y., Clarizio, C., Anthony, S. E., & Scheirer, W. J. (2018b, September). Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the european conference on computer vision (eccv)*.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025.

Roads, B. D., & Love, B. C. (2021). Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3547–3557).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition.*

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 1-9). doi: 10.1109/CVPR.2015.7298594