A Neurobiological Evaluation Metric for Neural Network Model Search

Nathaniel Blanchard
Computer Science and Engineering
University of Notre Dame
nblancha@nd.edu

Jeffery Kinnison
Computer Science and Engineering
University of Notre Dame
jkinniso@nd.edu

Brandon RichardWebster
Computer Science and Engineering
University of Notre Dame
brichar1@nd.edu

Pouya Bashivan
McGovern Institute for Brain Research and
Dept. of Brain and Cognitive Sciences, MIT
bashivan@mit.edu

Walter J. Scheirer
Dept. of Computer Science and Engineering
University of Notre Dame
walter.scheirer@nd.edu

Abstract

Neuroscience theory posits that the brain's visual system coarsely identifies broad object categories via neural activation patterns, with similar objects producing similar neural responses. Artificial neural networks also have internal activation behavior in response to stimuli. We hypothesize that networks exhibiting brain-like activation behavior will demonstrate brain-like characteristics, e.g., stronger generalization capabilities. In this paper we introduce a human-model similarity (HMS) metric, which quantifies the similarity of human fMRI and network activation behavior. To calculate HMS, representational dissimilarity matrices (RDMs) are created as abstractions of activation behavior, measured by the correlations of activations to stimulus pairs. HMS is then the correlation between the fMRI RDM and the neural network RDM across all stimulus pairs. We test the metric on unsupervised predictive coding networks, which specifically model visual perception, and assess the metric for statistical significance over a large range of hyperparameters. Our experiments show that networks with increased human-model similarity are correlated with better performance on two computer vision tasks: next frame prediction and object matching accuracy. Further, HMS identifies networks with high performance on both tasks. An unexpected secondary finding is that the metric can be employed during training as an early-stopping mechanism.

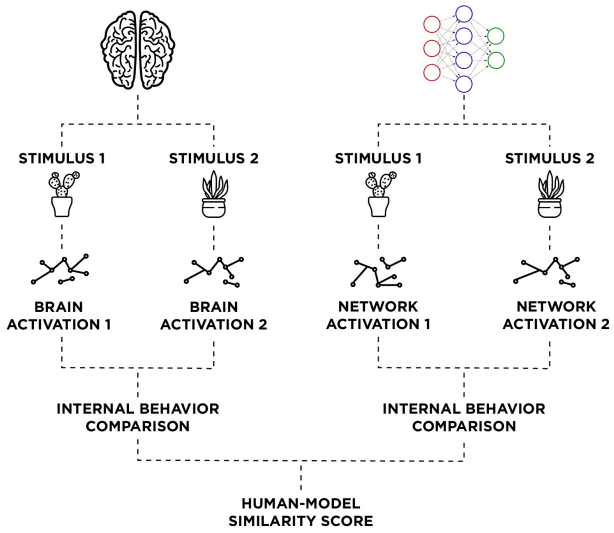

Figure 1. A primary goal of biologically-inspired deep learning work is achieving generalization capabilities that more closely resemble those of biological brains. Along these lines, we propose that model search frameworks for neural network training can be guided by a human-model similarity metric. The metric correlates internal activation behavior of the human brain and neural networks over shared stimuli. In this work, we examine the specific case of fMRI recordings [23] and predictive coding networks [29]. Internal behavior is measured by the dissimilarity in activations between two stimuli. Human-model similarity is the comparison of internal behavior of a brain and a model on a stimulus set, where higher similarity implies better model generalization.

1. Introduction

Researchers originally designed artificial neural networks based on neurobiological structure and function in the hope that such networks would approximate the performance of the biology that inspired them [44]. With the advent of modern deep learning techniques, neural networks are finally beginning to realize this original objective across some pattern recognition problems [25]. However, we need only consider the learning and processing power of the brain to know that neural network performance is a far stretch from generalized human capabilities [5, 16, 40, 41]. This shortcoming has inspired researchers to design new networks which better approximate neurobiological structure, utilizing the architectural elements of machine learning to build networks that embody modern theories of brain organization [29, 30, 42, 43, 47, 53]. In this paper we look beyond structural similarity and consider behavioral similarity between biological brains and trained networks (*i.e.*, models), as measured by similarity of activation behavior across a set of stimuli. We hypothesize that networks with increased behavioral similarity will exhibit better generalization capabilities across different visual recognition tasks.

One neurobiologically-inspired network is the unsupervised predictive coding network [29, 39]. Predictive coding networks combine the empirical successes of neural networks with insights from computational neuroscience to train unsupervised models with increased biological fidelity (*i.e.*, the correspondence of an algorithm's representations, transformations, and learning rules with those of their counterparts in the brain). The networks are designed [29] and demonstrated [30] to embody the theory that in-the-wild biological vision systems are continuously predicting the next input signal [39]. Additionally, these networks are trained using unsupervised video data, something also done by biological beings [25], allowing large-scale unsupervised learning. Finally, the networks have been shown to perform well on at least two different tasks: next frame prediction and object matching [29].

Predictive coding networks are architecturally designed to emulate neural processing. However, the ability of biological beings to generalize and adapt extends from both structure and internal behavior. Internally, the visual system processes similar objects with similar patterns of cell activations [7, 14, 33]. This activation behavior is an observable manifestation of the brain's ability to generalize beyond its experience, such as automatically allowing the classification of unseen instances of object classes (*e.g.*, correctly identifying a car despite never having seen this specific car before). We hypothesize that predictive coding networks which mimic the brain's visual behavior will exhibit increased biological fidelity and thus possess strong generalization ability compared to coding networks that do not exhibit this behavior. To test this hypothesis we investigate a new human-model similarity metric (HMS) that evaluates the networks for internal behavioral similarity to fMRI recordings of the human brain (Fig. 1).

Systems such as predictive coding networks and biological brains both exhibit internal behavior through their neural activations. Therefore, assessing networks for internal behavior indicative of biological fidelity requires measuring the similarity of activations. To do this, we make use of the recently established technique of representational similarity analysis (RSA) [23, 32]. RSA utilizes a set of stimuli to quantify behavioral similarity from activations. For any brain or network, the activation power can be measured in response to a stimulus. The internal behavior can then be defined as the dissimilarity in activations over a set of stimuli. In the case of visual recognition, we expect like-stimuli to have like-activations. We utilize a set of stimuli selected to exhibit a range of both similar and dissimilar objects [24].

One problem with assessing the similarity of activation behavior of biological beings and neural networks is the absence of a one-to-one mapping between the neurons of the brain and the neurons of the networks. With RSA, complex systems are abstracted into representational dissimilarity matrices (RDMs), composed of the internal behavior of the system, which is the activation dissimilarity over the set of stimuli. The full process to abstract a system into an RDM is illustrated in Fig. 2. Two input systems can be directly mapped when both are abstracted into RDMs with the same stimuli. Our proposed HMS metric measures human-model similarity as the correlation of a human fMRI RDM and a neural network RDM.

We evaluate the HMS metric in a Monte Carlo scenario across a broad range of hyperparameterized networks, data domains, and alternative network metrics. This approach allows us to explore the range of internal behavioral similarity that we can expect to find in predictive coding networks. Additionally, this method allows us to consider how a metric for human-model similarity could be used in the model search process for neural network training. While RSA has been employed to analyze similarities between convolutional neural networks (CNNs) and biological behaviors [21, 32, 51, 52], a generalized human-model similarity metric as an evaluation of a network's neurbiological fidelity and its use in neural network model search has remained largely untested until now. Our goal is to present a data-driven study of the HMS metric in order to promote it as a tool for studying generalization in computer vision.

In summary, we make the following contributions: (1) The introduction and evaluation of a new human-model similarity metric, dubbed HMS, to measure network generalizability.[1] (2) The implementation of a metric evaluation framework to assess new machine learning performance metrics. (3) The discovery of HMS as an indica-

[1] https://github.com/CVRL/human-model-similarity

tor of a predictive coding network's performance via experiments on the KITTI [15], VLOG [13], and "Gazoobian Object" [48] datasets. (4) The identification of HMS as an early stopping mechanism for training.

2. Related Work

How to best evaluate machine learning algorithms is an ongoing discussion. Traditional evaluations focus on external performance on a dataset, but there are no guarantees against overfitting or unpredictable network performance on real-world data [49]. One alternative evaluation regime is visual psychophysics, which monitors neural network performance while increasingly perturbing stimuli [26, 40, 41]. This evaluation centers on the observation that a network which inconsistently recognizes perturbed stimuli cannot be trustworthy. However, these evaluations are still focused on creating variability within a dataset, offering no guarantee that the network is not simply overfit to it. Moving beyond datasets, our proposed evaluation metric HMS quantifies consistency in a network's internal behavior by directly comparing against the internal behavior of one of the most generalizable vision systems in the world: the biological brain [7, 14, 33].

HMS uses human participant fMRI data as ground-truth internal behavior that leads to good generalization. The comparison between networks and human fMRI data is inspired by Kriegeskorte *et al*. [23], who described how network or neural activations could be abstracted into an RDM. An RDM is an abstract representation that can be directly compared against another RDM, as long as both are created from a joint set of stimuli. Fig. 2 shows how internal behavior is calculated and abstracted into RDMs, and how RDMs can be compared. Sec. 3.3 describes the formal RDM creation process. Kriegeskorte has a long history of utilizing RDMs to study neural behavior [21, 22, 23, 24, 34, 35].

With respect to the intersection between neuroscience and machine learning, the neuroimaging technique of fMRI has been used as ground-truth for designing features [8], interpreting neural network features [19, 28], and studying network performance [46]. Fong *et al*. [12] recently found that raw fMRI data could be used to weight support vector machines to improve performance, indicating that coarse-level brain data can potentially help machine learning networks generalize. The success of that study, alongside the public release of human fMRI data in RDM form by Nili *et al*. [35] further motivated us to use fMRI data as ground-truth in our network evaluation. The specific contributions fMRI data can make in expanding our understanding of neural networks are still to be explored, but to our knowledge this is the first instance of fMRI data being deployed for neural network model search, where the task is to screen different hyperparameter and architecture configurations for models that perform well on a given task. There is significant recent interest in optimization methods, search strategies, and infrastructure for neural network model search [3, 10, 18, 27, 36]. In this context, our work represents a new capability for such searches.

Extensive research has been performed comparing the neural activity of macaques to CNNs [17, 20, 50, 51, 52]. These studies map CNN layers to anatomical visual areas measured with electrode arrays. Recently, research has shown that these internal representations are not predictive of primate behavior at the image level [38, 45], suggesting CNNs are not mimicking internal behavior well enough. Given these recent findings, we opted to study more biologically plausible predictive coding networks [29, 39]. These networks are unsupervised and are relatively unexplored for many problem domains, but yield state-of-the-art performance for problems such as next frame prediction. We selected the PredNet architecture because of research establishing its emergent properties that are consistent with biological vision [30], meaning it is not grounded only in theory. Nonetheless, there are many biologically-inspired neural network architectures [37, 42, 43, 47, 53], and interest in them continues to grow [2]. All such networks warrant an investigation into internal behavior as well.

3. Methods

In this section we introduce the core methodologies surrounding the HMS metric. First, we introduce the biologically-inspired predictive coding network used for the experiments. We then explain the evaluation framework that was used to study HMS, and discuss the computer vision tasks (object matching and next frame prediction) that network performance was evaluated on. Finally, we detail the metric itself (Fig. 2), explaining: (1) the abstraction of both the fMRI recordings and neural networks into individual RDMs by measuring activations in response to stimuli, and (2) the correlation of the fMRI and the neural network RDMs, which results in the HMS score.

3.1. PredNet: A Biologically-Inspired Network

PredNet [29] is a recently introduced, unsupervised, biologically inspired, predictive coding network. Its architecture consists of multiple layers (which can vary based on configuration) each incorporating representation neurons (convolutional LSTM units), which output layer-specific predictions at each time step when processing a sequence of data. This output is then compared against a target to calculate an error term, which is propagated laterally and vertically throughout the network. We follow the PredNet training regime laid out by Lotter *et al*. [29]. PredNet is trained without supervision: the network is shown a randomly sampled set of sequential frame sequences and upon viewing each frame, the network attempts to predict the next

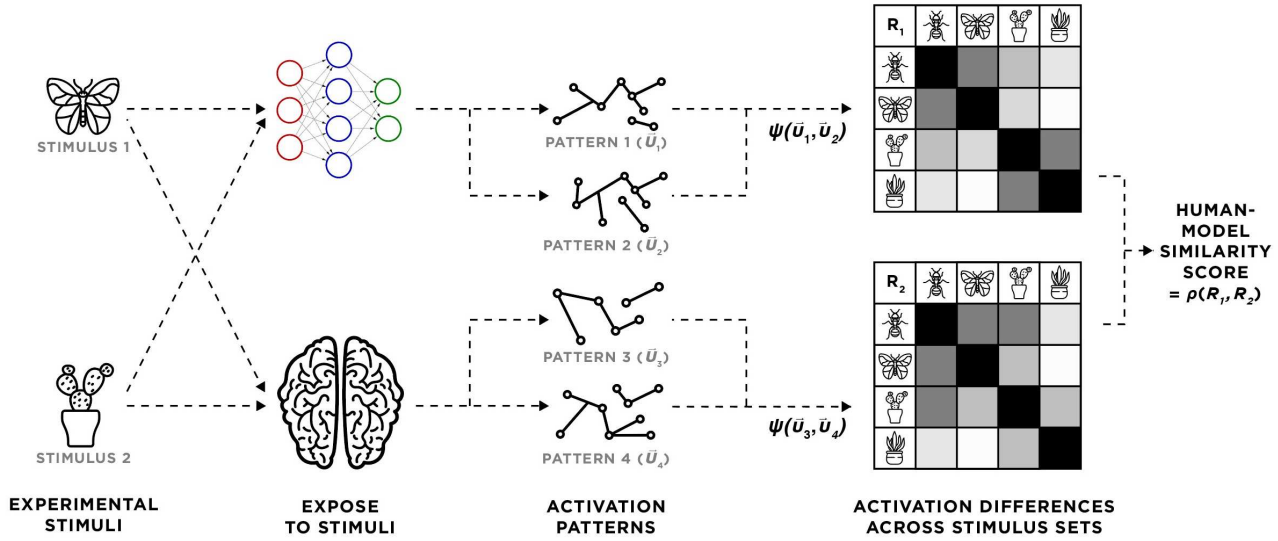

Figure 2. The proposed human-model similarity metric HMS is calculated by comparing the neural activation behavior from two systems: predictive coding networks and fMRI recordings of the human brain. Neural activations are obtained by exposing the systems to stimuli. We abstractly summarize a source based on its internal behavior, generating a similarity score via ψ from activation patterns for each stimulus pair. We then store this internal behavior to the stimuli in an RDM (R_1 and R_2 above). Finally, the HMS metric ρ is equal to the Spearmen's rank correlation coefficient of the internal behavior of the two sources as measured by the stimulus pairs.

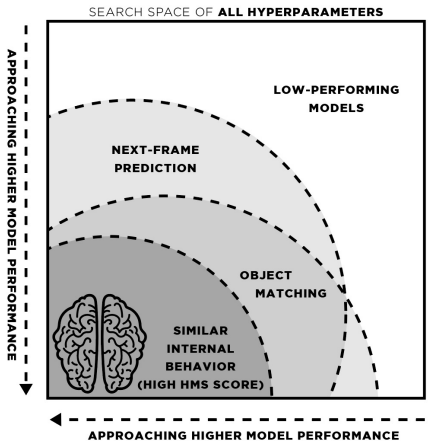

Figure 3. We assess our proposed HMS evaluation metric on randomly hyperparameterized predictive coding networks to study a Monte Carlo-style statistical sample of the space. We evaluate each network with three metrics: HMS, an object matching accuracy metric, and a next frame prediction error metric. We then compare metric performance across the full set of trained networks. We find that networks with higher HMS have high performance on other computer vision metrics, and performance is linked both across and within networks.

frame. The network is optimized to reduce the next frame prediction error on the training set.

3.2. Metric Evaluation Framework

Because we are focused on improving generalizability, we assess the value of HMS as a predictor of other, more standard, performance measures. This involves varying hyperparameters within a network type, obtaining a Monte Carlo-style statistical sample of the search space, and correlating HMS with standard computer vision evaluation metrics across networks in the sample (Fig. 3). We analyze the networks by studying the mean, standard deviation, and Spearman's rho (correlation coefficient) of several performance metrics across the set of sampled networks. We ensure significance by reporting Spearman's rho p values, which correspond to the likelihood that correlations occur by chance. We also adhere to Cohen's standard recommendation for interpreting effect sizes [6], and do not consider small correlations (less than 0.2) when comparing two different metrics — even if they reached statistical significance. Further, we perform Bonferroni correction, which conservatively adjusts significance to counteract the multiple testing problem where multiple inferences increase the likelihood of erroneous inferences [9]. In all of our results, Bonferroni adjusted p values are reported.

In this study, we correlate HMS with mean-squared error (MSE) on the next frame prediction task (the default mode of PredNet), as well as object matching accuracy. In the experiments, following the protocol established by Lotter *et al*. [29], MSE is computed as the square of the mean pixel-wise difference of the predicted next frame and the actual next frame. Object matching accuracy is evaluated

by first extracting the neural activations of the final layer in response to a probe image. Neural activations from the final layer are then extracted across a gallery of 50 images, one of which is the same object with altered lighting, color, viewing angle, or a combination thereof. Cosine similarity is computed between the probe and the gallery activations, and the gallery image with the highest activation similarity to the probe is the predicted match.

3.3. The HMS Metric for Model Search

Several steps are involved in computing the proposed HMS metric (with the major ones highlighted in Fig. 2). The RDM creation process described below follows the procedure of the RSA toolbox [35].

Stimuli selection for RDM construction. The stimuli were chosen by Kriegeskorte *et al*. [24] to compare human-primate neural inferior temporal (IT) object representations. The stimuli were selected to provide a hierarchical range of dissimilar and similar objects, such as animate and inanimate objects, not human and human objects, and face and body objects. The full set of stimuli is described in Sec. 1.1 of the Supp. Mat.

Human fMRI collection. The human fMRI data were released as part of the Representational Dissimilarity Toolbox [35]. All data are in RDM format, meaning we did not directly process the fMRI data, but instead received the set already in usable form. As such, anyone can utilize this data without specific fMRI domain knowledge, which makes the HMS metric broadly applicable to machine learning tasks. Although data from four participants were collected over two sessions, following the methods of Mur *et al*. [34], we averaged the subject RDMs together into a mean human brain RDM, which reduced noise. RDMs were constructed from activations in the bilateral IT region of the brain.

The full details of the human fMRI data collection can be found in [24]. Nonetheless, for completeness we briefly describe the procedure Kriegeskorte *et al*. [24] used to collect human fMRI data. Eight RDMs were constructed from fMRI recordings of four subjects over two sessions in response to 92 stimuli. Recordings were from measurements of $1.95 \times 1.95 \times 2mm^3$ within an occipitotemporal measurement slab ($5cm$ thick). Subjects were presented with a random sequence of the 92 stimuli. Each stimulus was displayed for 300 milliseconds, every 3700 milliseconds, with four seconds between stimuli. Not all voxels were used to construct the RDM. Voxels of interest were selected based on voxel responses to stimuli from an independent dataset. No spatial smoothing or voxel averaging was performed.

PredNet activations to stimuli. Using the exact same set of 92 stimuli, we construct an RDM using network activations as features from PredNet's internal representation neurons. Specifically, activations are recorded from the convolutional LSTM units. Predictive coding networks are time-based networks, and thus we present the stimuli for a fixed five frames and record activations at each time step. We discard the first time step as it corresponds to a "blank" prediction. Activation patterns from PredNet for this style of stimuli presentation mimic biological neural responses for perception [30].

RDM construction. Given a single feature f and a single stimulus s, $v = f(s)$, where v is the value of feature f in response to s. Likewise, the vector

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}^T = \begin{bmatrix} f_1(s) \\ f_2(s) \\ \vdots \\ f_n(s) \end{bmatrix}^T \tag{1}$$

can represent the feature values of a collection of n features, $f_1,\ f_2,\ ...,\ f_n$, in response to s. If one expands the representation of s to a set of m stimuli $S = s_1,\ s_2, ...,\ s_m$, the natural extension of $\vec{v}$ is the set of feature value collections $V = \vec{v}_1,\ \vec{v}_2, ...,\ \vec{v}_m$, in which $s_i \in S$ is paired with $\vec{v}_i \in V$ for each $i = 1, 2, ..., m$. The last step prior to constructing an RDM is to define the dissimilarity score between any two $\vec{v}_i \in V$ and $\vec{v}_j \in V$. We use the symmetric function

$$\psi(\vec{v}_i, \vec{v}_j) := 1 - \frac{(\vec{v}_i - \bar{v}_i) \cdot (\vec{v}_j - \bar{v}_j)}{\|\vec{v}_i - \bar{v}_i\|_2 \|\vec{v}_j - \bar{v}_j\|_2} \tag{2}$$

where $\bar{v}$ is the mean of the features in $\vec{v}$. An RDM R may then be constructed from S, V, and ψ as:

$$R = \begin{bmatrix} \psi(\vec{v}_1, \vec{v}_2) & \psi(\vec{v}_1, \vec{v}_3) & \dots & \psi(\vec{v}_1, \vec{v}_m) \\ & \psi(\vec{v}_2, \vec{v}_3) & \dots & \psi(\vec{v}_2, \vec{v}_m) \\ & & \ddots & \vdots \\ & & & \psi(\vec{v}_{m-1}, \vec{v}_m) \\ & & & \end{bmatrix} \tag{3}$$

Human-model similarity (HMS). Given any two RDMs R_1 and R_2 from the same set of stimuli S, one can compute their similarity to determine how similar the activation behavior is in response to S. The similarity function

$$HMS = \rho(\hat{R}_1, \hat{R}_2) \tag{4}$$

computes a Spearman's rank correlation coefficient represented by ρ, where $\hat{R}$ is the flattened RDM.

Thus HMS is calculated as the correlation between the averaged human fMRI RDM and a constructed PredNet network RDM, obtained from the network activations to the stimuli. The resulting score is defined over the real interval $[-1, 1]$, with 1 indicating perfect correlation, -1 indicating perfect negative correlation, and 0 indicating the two RDMs are completely uncorrelated.

Evaluation Task	Metric	Mean (SD)	Top Ten HMS Mean (SD)
Next Frame Prediction Error	Pixel MSE	0.092 (0.148)	0.009 (0.003)
Object Matching	Accuracy	0.367 (0.134)	0.459 (0.049)
Human-Model Similarity	RDM Correlation	0.106 (0.055)	0.178 (0.011)

Table 1. A statistical overview of evaluation scores for a sample of 95 randomly hyperparameterized PredNet networks. These scores indicate the range of scores we expect to obtain from an arbitrary PredNet network. The top ten HMS mean score refers to the average score for each metric for the ten networks with the highest human-model similarity. The top ten average shows that networks with high HMS also achieve high performance on the other tasks. The object matching task was intentionally designed to be difficult — the network must distinguish fine-grained differences in unseen, fictional "Gazoobian" objects [48] where task chance is (0.02). Networks are trained using KITTI [15] and evaluated on next frame prediction using a held-out set of KITTI data. Pixel MSE is mean squared error of the predicted-to-actual frame at the pixel level. SD is standard deviation.

4. Experiments

Our experiments assess how the biological fidelity of predictive coding networks affects two computer vision tasks: next frame prediction and object matching. We identify biological fidelity as more similar internal activation behavior to human fMRI, as measured through RDMs.

Four datasets are utilized. We evaluate HMS, as described in Sec. 3.3, on a dataset of 92 stimuli with a range of similar and dissimilar objects, from real human faces to animated objects [23]. Computer vision capabilities are evaluated on two tasks: next frame prediction and object matching accuracy, as described in Sec. 3.2. Next frame prediction is assessed by measuring pixel-level MSE on the KITTI dataset [15], a video dataset composed of image sequences from a car mounted camera. We also experimented with another video dataset, VLOG [13]. For object matching, we used a randomly generated "Gazoobian Objects" dataset (following the procedure described by Tenenbaum *et al*. [48]), composed of otherworldly objects guaranteed to be unseen in training. Gazoobian stimuli mirror the stimuli presentation of the HMS stimuli. Even though these objects are well out of domain compared to the natural images used for training, humans are able to generalize to them with ease [48], making them an excellent basis from which to study model generalization at inference time. Objects were varied in rotation, lighting, color, or a combination thereof. Example images from all datasets can be found in Sec. 1 of the Supp. Mat.

4.1. Does HMS Discover Models that Generalize?

Initially, we evaluated a random Monte Carlo-style sample of hyperparameters in order to test how HMS, next frame prediction, and object matching varied across PredNet networks. In typical model search fashion, we varied six hyperparameters including the number of training epochs, the number of video sequences used for validation after training for an epoch, the number of video sequences used to train within an epoch, the batch size, the learning rate, and the size of the convolutional filters across all layers. The exact space searched is listed in Sec. 2.1 of the Supp. Mat. We trained 95 4-layer PredNets with randomly selected hyperparameters using HyperOpt [4], a software package for distributed hyperparameter optimization.

In Table 1 we report the metrics' mean and standard deviation for the 95 trained PredNets. Next frame prediction was within range of Lotter *et al*. [29]. The accuracy scores highlight the difficulty of the object matching task, which focuses on specific object matching from a 50 image gallery of stimuli (chance $= 0.02$). The evaluation scores indicate our parameters were well suited for sampling: performance was above chance but below ceiling. Impressively, the mean HMS was within a standard deviation of the average human-human similarity score of 0.19 (SD $= 0.09$). We also confirmed that these results are stable in a cross-dataset context using the VLOG dataset [13] (these experiments are discussed in Sec. 3 of the Supp. Mat.).

We next examined how high HMS similarity corresponds with other metrics by looking at the 10 networks with the highest HMS scores (reported in Table 1). These networks achieve much higher performance over the set of all networks on the two computer vision tasks. We also examined the 10 networks with the lowest HMS, and note that they have much worse than average performance: mean next frame prediction error was 0.314 (SD $= 0.138$), mean object matching accuracy was 0.13 (0.15), and mean HMS was -0.008 (0.027). This shows HMS is an effective metric for predicting performance. Networks with high HMS perform well, and those with low HMS perform poorly.

To be useful, HMS needs to be an effective predictor across all models, not just high and low performing models. We verified that, across all models, higher HMS is associated with higher performance on the other metrics by computing Spearman's rho across the sampled networks (Table 2). Further, the p values of these correlations are the probability our findings occur by chance, with $p < 0.001$ indicating a less than 0.001 probability (0.1%) that our correlations occur by chance (see Sec 3.2. for details of these safeguards). The strength of the correlations between the metrics are moderate to strong, with $p < 0.001$. This con-

Variable	Accuracy	HMS	Learning Rate
Next Frame Prediction Error	-0.791**	**-0.646****	0.635**
Object Matching Accuracy	.	**0.575****	-0.517**
Human-Model Similarity	.	.	-0.452**

**$p < 0.001$

Table 2. Spearman's rho of metrics for 95 trained PredNets with random hyperparameters. The correlations confirm that HMS is predictive of network performance on other metrics. The negative correlation between Next Frame Prediction Error and the two other metrics occurs because next frame prediction is measured by error, which should minimized, while HMS and Accuracy are metrics to be maximized. Precautions taken in determining statistical significance are described in Sec. 3.2. Learning rate was correlated with each metric, but was not determined to be a significant contributing factor to HMS as a predictor of network performance after partial correlation analysis.

firms that HMS is predictive of network performance on computer vision tasks. Additionally, we calculated correlation scores for all hyperparameters to verify that no individual parameter was responsible for these results. We include the learning rate (LR) hyperparameter in Table 2 because it is moderately correlated with the other metrics.

The correlation with LR indicated a possible risk that LR is strongly influencing the results. We investigated its influence with a partial correlation analysis, which measures the relationship between metrics while controlling for the influence of LR. The correlations between metrics from Table 2 were not statistically significant ($p < 0.001$); however, the sample size was too small for the breadth of LRs tested. We addressed this by repeating the partial correlation on a much larger set of networks ($N = 1811$). For this sample, the partial correlations between the metrics were statistically significant ($p < 0.001$), with similar correlation strength for the sample. This confirms HMS is significantly correlated with the other metrics regardless of the influence of LR on training. More discussion of this experiment can be found in Sec. 2.2 of the Supp. Mat.

All of the findings discussed above provide evidence that HMS is an effective search metric. HMS was indicative of performance for both computer vision tasks across all models (via correlation) and extremes (top and bottom models). Networks which exhibited more brain-like internal behavior generalized better to other evaluation tasks.

4.2. Metric Stability During Model Search

How stable are our evaluation metrics during network training? Do evaluations of network performance vary across identically hyperparameterized models? If HMS fluctuates wildly during training, it may be an unreliable indicator of performance. Through further experimentation, we found that this is not the case, and show that HMS is a predictor earlier in training than the other metrics.

Within-network stability. We first investigated how the metrics varied during training on a sample of 74 4-layer PredNets trained for 150 epochs, evaluating performance every 5 epochs. We focused our analysis on 10 networks where MSE was below 0.01 by the 150th epoch, implying convergence. We found each metric had its own predictable behavior, illustrated by a representative network in Fig. 4, consistent across hyperparameters. Once HMS was stable (SD ≤ 0.01) for 25 epochs it remained so. Object matching accuracy tended to start higher, before dropping, and finally rising again as the training unfolded. Finally, next frame prediction error either continuously decreased, leading to a good network, or increased, leading to a degenerate network. The correlations from Table 2 imply that any of the metrics could be used as a predictor, but the training behavior offers insight into how these metrics would need to be utilized. HMS stabilizes first, after an average of 32 epochs. Accuracy stabilizes next, after an average of 66.5 epochs (SD = 36), although some scores did not plateau but continued to increase. In cases where next frame prediction error (MSE) decreased with training, it typically decreased throughout all 150 epochs, making a poor indicator

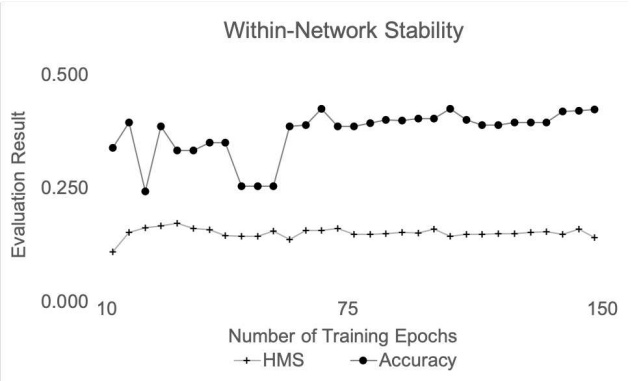

Figure 4. Within-network stability analysis for a representative PredNet model. We find that each metric has its own stereotypical behavior during training. Object matching accuracy is inconsistent early in training, but eventually stabilizes and continues to increase. Next frame prediction error (MSE) either falls consistently (the case shown above) or rises unpredictably, but is heavily dependent on training time. HMS is inconsistent very early in training, but stabilizes more quickly than accuracy, which is unstable for longer, or MSE, which requires a long training time before stabilizing. These findings mean that HMS can be used to identify poor-performing networks for early stopping in network search.

of performance. Note that in the 95 model sample, MSE was the only metric correlated with the number of epochs ($-0.332, p < 0.001$). Further details, results, and experiments on across-network stability can be found in Sec. 5 of the Supp. Mat.

4.3. A Mechanism for Early Stopping

An outcome of the findings from Sec. 4.2 is that our proposed HMS metric can be employed during network training as a way to discard (*i.e.*, stop training) models that will ultimately perform poorly. To demonstrate this, we conducted a *post hoc* analysis of the 95 PredNets from Sec. 4.1. On the left-hand side of Fig. 5 we present time saved by early stopping with HMS and accuracy using the convergence criteria of Sec. 4.2. Overall, early stopping with HMS could have reduced training time by 67% at no cost to final performance. We also tested a threshold strategy which considered a network to be stopped during training if its HMS score was below a threshold of 0.161 (the mean HMS from Table 1 plus one standard deviation). Only 13 of the 95 models (13.7%) were above this threshold. The right-hand side of Fig. 5 depicts the accuracy scores of the models with respect to the side of the HMS threshold they are on. Our analysis shows that even with a high threshold to stop training, and the loss of some models with high performance, most retained models are high performing and were more likely to have high performance on both tasks. Additionally, in this case the highest performance model on both computer vision tasks is retained, but a number of other retained higher performing models have trivial differences in performance, and would be just as useful had the top model been discarded. Complete details for these experiments and additional results can be found in Sec. 6 of the Supp. Mat.

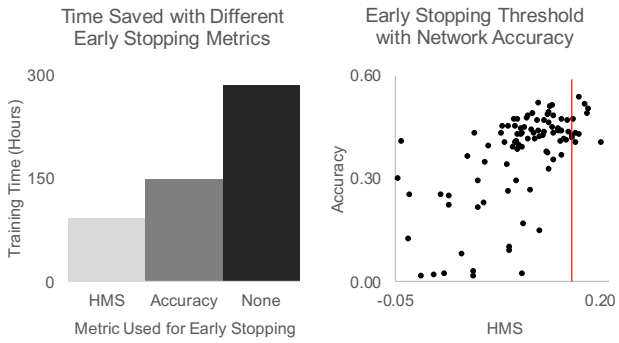

Figure 5. The left-hand plot shows how HMS-driven early stopping for the sample of 95 PredNets cut down training time by 67%, using the criteria for convergence of SD $\leq$ 0.01 for 25 epochs. Using the same convergence criteria for accuracy was not as effective. The right-hand scatter plot shows the accuracy of models above and below an early stopping threshold (0.161). 82 models left of the line would be discarded at no cost to final performance. These experiments utilize the findings for metric stability established in Sec. 4.2 to quantify the potential outcome of utilizing early stopping on the model sample.

5. Discussion

There are several benefits to utilizing HMS over traditional human-model comparisons. (1) HMS is useful for model searches because activation patterns for learned representations emerge early in training, whereas other evaluations require fully training a network. (2) There is evidence that HMS is indicative of a model's ability to generalize to unseen data and tasks, since PredNet models with higher HMS are more likely to perform well on both object matching accuracy and next frame prediction. (3) Compared with other evaluations of perceptive consistency, such as visual Psychophysics [40, 41], HMS is much less computationally expensive. Consider the computational cost of HMS evaluation compared with the accuracy evaluation, which utilizes psychophysical stimuli (varying lighting and texture). HMS only requires a network to process 92 stimuli. The PredNet accuracy metric requires the network to process 51 stimuli (1 probe, 50 gallery) per trial, for 500 trials (25,500 stimuli). (4) We used fMRI data as a benchmark because it overcomes the difficulty in labeling the correct similarity between different objects. For example, as humans we instinctively know a pair of faces should have highly similar activation behavior, but what about a hand and face? Neural data provides an implicit answer to this question. One concern that can be raised is the perceived difficulty of obtaining fMRI data. Fortunately, there is a growing open science movement within neuroscience. The fMRI data used in this study is publicly available and can be utilized by anyone [35], and it is far from the only data available. Vast public fMRI repositories exist for vision, text, and audio tasks, and researchers do not need to be experts in order to utilize them. A few examples are the Donders repository [11], OpenNeuro [1], and Oasis [31] brains.

We believe that networks with more biological fidelity in function will be essential to overcome the shortcomings of today's networks in replicating biological vision. The future of artificial intelligence research will need to bridge the gap between network structure and internal behavior, which requires reassessing how we evaluate networks. In the past, unexpected network behavior has blinded-sided researchers, *e.g.*, susceptibility to adversarial images. And it is important to remember that current networks are not consistent with human behavior [40, 41]. Evaluations measuring internal behavior should prove useful for avoiding unforeseen issues, and may help us achieve the next generalization breakthrough.

6. Acknowledgements

Funding was provided under IARPA contract D16PC00002 and NSF DGE 1313583.

References

[1] https://openneuro.org/. 8

[2] D. G. Barrett, A. S. Morcos, and J. H. Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *arXiv preprint arXiv:1810.13373*, 2018. 3

[3] P. Bashivan, M. Tensen, and J. J. DiCarlo. Teacher guided architecture search. *arXiv:1808.01405 [cs]*, Aug. 2018. arXiv: 1808.01405. 3

[4] J. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *ICML (1)*, 28:115–123, 2013. 6

[5] M. F. Bonner and R. A. Epstein. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology*, 14(4):e1006111, Apr. 2018. 2

[6] J. Cohen. *Statistical power analysis for the behavioral sciences. 2nd edition.* Lawrence Erlbaum Associates, 1988. 4

[7] M. N. Coutanche and G. E. Koch. Creatures great and small: Real-world size of animals predicts visual cortex representations beyond taxonomic category. *NeuroImage*, 183:627–634, Dec. 2018. 2, 3

[8] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, Jul 1985. 3

[9] O. J. Dunn. Estimation of the means of dependent variables. *Ann. Math. Statist.*, 29(4):1095–1111, 12 1958. 4

[10] T. Elsken, J. H. Metzen, and F. Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018. 3

[11] FAIRsharing Team. Donders Repository, 2018. type: dataset. 8

[12] R. Fong, W. J. Scheirer, and D. Cox. Using human brain activity to guide machine learning. *Scientific Reports*, 8:5397, March 2018. 3

[13] D. F. Fouhey, W.-c. Kuo, A. A. Efros, and J. Malik. From lifestyle vlogs to everyday interactions. In *IEEE/CVF CVPR*, 2018. 3, 6

[14] F. E. Garcea, J. Almeida, M. H. Sims, A. Nunno, S. P. Meyers, Y. M. Li, K. Walter, W. H. Pilcher, and B. Z. Mahon. Domain-specific diaschisis: Lesions to parietal action areas modulate neural responses to tools in the ventral stream. *Cerebral Cortex*, 2018. 2, 3

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3, 6

[16] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schuett, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018. 2

[17] H. Hong, D. L. Yamins, N. J. Majaj, and J. J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613, 2016. 3

[18] C.-H. Hsu, S.-H. Chang, D.-C. Juan, J.-Y. Pan, Y.-T. Chen, W. Wei, and S.-C. Chang. Monas: Multi-objective neural architecture search using reinforcement learning. *arXiv preprint arXiv:1806.10332*, 2018. 3

[19] I. Kalfas, K. Vinken, and R. Vogels. Representations of regular and irregular shapes by deep convolutional neural networks, monkey inferotemporal neurons and human judgments. *PLOS Computational Biology*, 14(10):e1006557, 2018. 3

[20] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, 6:32672, Sept. 2016. 3

[21] N. Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3(3):363–373, 2009. 2, 3

[22] N. Kriegeskorte. Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage*, 56(2):411–421, 2011. 3

[23] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, Nov. 2008. 1, 2, 3, 6

[24] N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, Dec. 2008. 2, 3, 5

[25] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015. 2

[26] J. Z. Leibo, C. d. M. d'Autume, D. Zoran, D. Amos, C. Beattie, K. Anderson, A. G. Castañeda, M. Sanchez, S. Green, A. Gruslys, et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018. 3

[27] C. Liu, B. Zoph, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. Progressive neural architecture search. In *ECCV*, September 2018. 3

[28] B. Long and T. Konkle. The role of textural statistics vs. outer contours in deep CNN and neural responses to objects. In *Conference on Computational Cognitive Neuroscience*, page 4, Philadelphia, Pennsylvania, 2018. 3

[29] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *ICLR*, 2017. 1, 2, 3, 4, 6

[30] W. Lotter, G. Kreiman, and D. Cox. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. *arXiv preprint arXiv:1805.10734*, 2018. 2, 3, 5

[31] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, Dec. 2010. 8

[32] P. McClure and N. Kriegeskorte. Representational distance learning for deep neural networks. *Frontiers in Computational Neuroscience*, 10:131, 2016. 2

[33] H. E. Moss, J. M. Rodd, E. A. Stamatakis, P. Bright, and L. K. Tyler. Anteromedial temporal cortex supports fine-grained differentiation among objects. *Cerebral Cortex*, 15(5):616–627, 2004. 2, 3

[34] M. Mur, M. Meys, J. Bodurka, R. Goebel, P. A. Bandettini, and N. Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in Psychology*, 4, 2013. 3, 5

[35] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014. 3, 5, 8

[36] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean. Efficient neural architecture search via parameter sharing. In *ICML*, 2018. 3

[37] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 5(11):e1000579, 2009. 3

[38] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. 3

[39] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79, 1999. 2, 3

[40] B. RichardWebster, S. E. Anthony, and W. J. Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE T-PAMI*, 2018. To Appear. 2, 3, 8

[41] B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *ECCV*, 2018. 2, 3, 8

[42] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019, 1999. 2, 3

[43] M. Roper, C. Fernando, and L. Chittka. Insect bio-inspired neural network provides new evidence on how simple feature detectors can enable complex visual generalization and stimulus location invariance in the miniature brain of honeybees. *PLoS Computational Biology*, 13(2):e1005333, 2017. 2, 3

[44] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 2

[45] A. Rosenfeld, M. D. Solbach, and J. K. Tsotsos. Totally looks like – how humans compare, compared to machines. In *Mutual Benefits of Cognitive and Computer Vision Workshop*, 2018. 3

[46] J. Sacramento, R. P. Costa, Y. Bengio, and W. Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In *NIPS*, 2018. 3

[47] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE T-PAMI*, 29(3):411–426, 2007. 2, 3

[48] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. 3, 6

[49] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE CVPR*, 2011. 3

[50] D. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. In *Nature Neuroscience*, volume 19, pages 356 – 365, 2016. 3

[51] D. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *NIPS*, 2013. 2, 3

[52] D. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. 2, 3

[53] I. Yildirim, W. Freiwald, and J. Tenenbaum. Efficient inverse graphics in biological face processing. *bioRxiv*, 2018. 2, 3