arXiv:1805.10726v3 [cs.CV] 9 Jul 2018

A Neurobiological Cross-domain Evaluation Metric for Predictive Coding Networks

Nathaniel Blanchard
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
nblancha@nd.edu

Jeffery Kinnison
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
jkinniso@nd.edu

Brandon RichardWebster
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
brichar1@nd.edu

Pouya Bashivan
McGovern Institute for Brain Research and
Dept. of Brain and Cognitive Sciences MIT
Cambridge, MA 02142
bashivan@mit.edu

Walter J. Scheirer
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
walter.scheirer@nd.edu

Abstract

Achieving a good measure of model generalization remains a challenge within machine learning. One of the highest-performing learning models is the biological brain, which has unparalleled generalization capabilities. In this work, we propose and evaluate a human-model similarity metric for determining model correspondence to the human brain, as inspired by representational similarity analysis. We evaluate this metric on unsupervised predictive coding networks. These models are designed to mimic the phenomenon of residual error propagation in the visual cortex, implying their potential for biological fidelity. The human-model similarity metric is calculated by measuring the similarity between human brain fMRI activations and predictive coding network activations over a shared set of stimuli. In order to study our metric in relation to standard performance evaluations on cross-domain tasks, we train a multitude of predictive coding models across various conditions. Each unsupervised model is trained on next frame prediction in video and evaluated using three metrics: 1) mean squared error of next frame prediction, 2) object matching accuracy, and 3) our human-model similarity metric. Through this evaluation, we show that models with higher human-model similarity are more likely to generalize to cross-domain tasks. We also show that our metric facilitates a substantial decrease in model search time because the similarity metric stabilizes quickly — in as few as 10 epochs. We propose that this metric could be deployed in model search to quickly identify and eliminate weaker models.

1 Introduction

Artificial neural networks were originally designed to mimic the brain at the neuronal level, but their success in this regard has been limited. Predictive coding networks are a class of models [32, 24]

Preprint. Work in progress.

which attempt to replicate the neurobiological theory that biological neural networks are constantly attempting to predict the next input signal. These networks combine the empirical successes of artificial neural networks with modern theories from neuroscience to create unsupervised models with increased biological plausibility. Thus, they have potential for strong generalization for large-scale unsupervised learning — but how much *biological fidelity* (*i.e.*, the correspondence of an algorithm's representations, transformations, and learning rules with those of their counterparts in the brain) do they really possess? Can models be trained in ways that increase their biological fidelity? Similarly, do variations in model structure coincide with variations in biological fidelity? How should biological fidelity be measured — and what would such a metric indicate?

Researchers create models based on biological structures and functions in the hopes that the behavior of their models will resemble that of the biology that inspired them. For each task, the model's measure of success will ideally mimic the success of the original biological system. In spite of recent advances, we need only consider the learning and processing power of the brain to know that machine learning is a far stretch from generalized human performance in many domains. In order to address this, we must first disassociate ourselves from traditional emphasis on the maximization of performance metrics, and begin to analyze and consider new metrics for understanding generalization. Recent work has proposed using representational similarity analysis (RSA) to compare network and brain activations [27]. This theory posits that it is possible to use the brain itself as a benchmark of generalizability. Through this reasoning, measuring similarities in the brain and network's reactions to a shared set of stimuli acts as a sort of Turing Test for model behavior. Ideally, models and brains will become indistinguishable, implying the model has a worldview similar to the brain and thus should achieve similar performance.

Although RSA has been employed in the past to analyze similarities between convolutional neural networks and biological behaviors [18, 40, 41], the potential of incorporating a generalized brain-model similarity metric as a measure of a model's generalizability is largely untested. Additionally, this use of a human brain-model similarity metric in the model search process is, as of yet, an unexplored idea [27]. In this work we seek to fill this gap by extensively examining how human-model similarity metrics vary across hyperparameters, domains, and evaluations. The goal of this work is to present a data-driven understanding of the positive and negative implications of the human-model similarity metric in order to promote it as a tool for studying cross-domain generalization. Here, cross-domain activities include cross-dataset evaluation and cross-task evaluation. We focus our efforts on predictive coding networks in order to maximize the biological plausibility of our model, and to further understand the potential cross-domain application of such networks.

The novelty of this work centers on the following contributions: (1) Proposal and evaluation of a human-model similarity metric to measure model generalizability. (2) Introduction and implementation of a framework to evaluate new machine learning performance metrics. (3) Identification of rapid human-model similarity convergence, in comparison with other metrics. (4) Discovery of human-model similarity as an indicator of a predictive coding network's cross-domain performance.

2 Related work

In this work, we propose a method to evaluate a novel metric. Our method utilizes model search practices to evaluate the metric in a variety of conditions, with the goal of incorporating it into guided model search if it is found to be useful. We implement our method of study using common, unguided methods of model search [9, 1]. In order for our metric to be applicable in a guided model search, its applicability within guided model searches selecting parameters [4, 37, 16, 12, 13, 26, 8, 22] and manipulating neural network architectures [7, 43, 44, 6] needs to be analyzed. Despite the recent work in these domains, the question of which *criterion* should guide such a search, and how to evaluate that criterion, remains largely unaddressed. The most relevant inspiration for our proposed framework comes from Pinto et al. [30], who argued that mass model evaluation was the only foolproof method to evaluate a particular model's viability. In this work, favorable performance on tasks is used to evaluate the biological plausibility of a model; however, we prefer to assess the biological fidelity of candidate models in addition to task performance.

Our proposed metric uses human participant fMRI data to determine human-model similarity. The largest inspiration for this metric is work by Kriegeskorte et al. [20], who described the use of representational dissimilarity matrices (RDM) as an abstract representation of the comparison between

any two models (including, potentially, the brain) using a joint set of stimuli [20, 18, 19, 21, 28]. McClure and Kriegeskorte [27] studied how RDMs could be incorporated into the training process using a "teacher" RDM to guide updates in stochastic gradient descent. They found that this technique improved visual classification performance, but they did not incorporate biological brains into their study. We utilize RDMs in this work to determine the similarity of trained predictive coding models to a human model, but do not attempt to guide the neural network training at this time. Using RDMs, extensive work has been done comparing neural activity of macaques to convolutional neural networks (CNN) [39, 41, 40, 15, 17]. These studies map CNN layers to fine-grained visual areas measured with electrode arrays. This work also often incorporates human subjects' human object-similarity judgments, which have been shown to map well to primate data [28]. Recent findings from Rajalingham et al. [31] show that although artificial neural networks accurately predicted primate patterns of object-level confusion, the networks were not predictive at the image-level. This suggest that new models are needed to capture neural mechanisms with more precision. Rather than focus on CNNs, we opted to study more biologically plausible predictive networks.

We focus our efforts on predictive coding networks [24], even though there are numerous biologically-inspired network models [35, 42, 30, 34, 36] based on the visual systems of other biological beings. In this work we focus on the study of established biologically-inspired predictive coding networks, which are unsupervised and relatively unexplored in many problem domains.

Fong et al. [10] 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 models generalize. The success of this study, alongside the public release of human fMRI data in RDM form by Mur et al. [28] inspired us to use fMRI data in a model evaluation metric.

Our framework for metric evaluation relies on the use of established metrics to mediate the study of our proposed metric. Since predictive coding networks are unsupervised models, we are evaluating them in a cross-domain context, using next frame prediction and object matching tasks. Numerous papers have studied model performance in a cross-task cross-domain context [25, 33, 5, 23]. These works, in general, used alternative data to optimize models for cross-domain performance. Our work centers on training a multitude of unsupervised models with various parameters using one dataset. We then independently evaluate each trained model on its intended task (next frame prediction), and on two other out-of-domain datasets and one out-of-domain task (object matching and human-model similarity).

3 Methods

3.1 PredNet: A biologically-inspired model for vision

PredNet [24] is an unsupervised, biologically inspired, predictive coding network. Predictive coding networks are designed to mimic the neurobiological theories of vision by predicting future events [32]. We select PredNet as a model to experiment with human-model similarity analysis because of its correspondence to neurobiological theory, as opposed to more conventional convolutional neural networks (which are only loosely inspired by biology at the architectural level). We follow the training regime laid out by Lotter et al. [24]. PredNet is trained without supervision, where the model is shown a random set of frame sequences. Upon viewing each frame, the model attempts to predict the next frame. The network is optimized to reduce the next-frame prediction error on the training set.

3.2 Metric evaluation framework

We propose a model evaluation framework to study a metric by varying hyperparameters within a model type, obtaining a Monte Carlo-style statistical sample of the space, and correlating the proposed metric with standard evaluation metrics across models in the sample. This allows us to study the behavior and application of such a metric while avoiding bias from a particular model configuration. In this study, we compare our proposed human-model similarity metric with mean-squared error (MSE) on the next-frame prediction task, as well as object identification accuracy. We use a cross-domain dataset of artificial stimuli for the cross-domain task of object matching. In the experiments, following the protocol established by [24], MSE is computed as the square of the mean pixel-wise difference of the predicted next frame and the actual next frame. Object matching

is evaluated by extracting the activations of the final layer in response to a stimulus image. This is repeated for a gallery of 50 images, and the pairwise cosine similarity is computed, with the lowest value being the predicted match.

3.3 A new neurobiological cross-domain evaluation metric

Here, we define our human-model similarity metric for use in model search. We first discuss the acquisition of human participant fMRI representational dissimilarity matrices (RDMs). We then explain how to create comparable RDMs for each model. Using the human and the model RDMs, we calculate our human-model similarity metric.

Human brain activations to stimuli. We adopted fMRI measurements from several visual cortical regions, including the inferior temporal cortex, of four human subjects in response 92 stimuli. The stimuli were chosen by Kriegeskorte et al. [21] to compare human-primate neural inferior temporal object representations. These measurements were used to construct RDMs, which provide a compact statistic of the representational space in each region. Each participant took part in two recording sessions in which they were presented with 92 stimuli for 300 milliseconds every 3700 milliseconds. Full experimental details can be found in Kriegeskorte et al. [21]. This data was released as part of the RDM toolbox [29]. Following these methods, we average the subject RDMs together into a generalized human-brain RDM, which reduces noise.

PredNet activations to stimuli. Using the exact same set of 92 stimuli, we construct an RDM using model activations as features from PredNet's internal representation units. Predictive coding networks are time-based networks, and thus we present the stimuli for a fixed five frames and record activations from each time step. We discard the first time step as it corresponds to a "blank" prediction.

RDM construction. Given a single feature f and a single stimulus s, the feature value $v = f(s)$, 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 $\vec{v}_j \in V$. Although there are many possible dissimilarity score functions, we use the symmetric function

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

where $\hat{v}_i$ is the unit-vector form of $\vec{v}_i$, and $\bar{v}_i$ is the mean of $\hat{v}_i$. 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 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 feature values are in response to S. The similarity function

$$\Phi(R_1, R_2) := \zeta(\mathrm{vec}(R_1), \mathrm{vec}(R_2)) \tag{4}$$

computes a non-monotonic Spearman's rank correlation coefficient represented by ζ. vec is the ordered process of converting a matrix into a single column vector.

Based on this process, human-model similarity is calculated as the Spearman correlation between the averaged human fMRI RDM and a constructed PredNet model RDM, obtained from the model 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.

4 Experiments

Our experiments evaluate the effect of guiding hyperparameter optimization by human-model similarity within a predictive coding network's intended application, next frame prediction, as well as the cross-domain application of object matching. Human-model similarity is measured by extracting a model RDM from a PredNet's representation layers. The stimuli used were the same as those used to build human RDMs: a collection of 92 objects which range from real human faces to animated objects [20]. Next frame prediction is calculated by measuring the pixel-level error in each prediction for a PredNet model at each frame of a held-out portion of the KITTI dataset [14], following [24]. Finally, object matching accuracy is calculated by probing a PredNet model and then finding the matching object within a gallery of 50 stimuli (chance 0.02). We used randomly generated "Gazoobian" stimuli (originally introduced by Tenenbaum et al. [38]), otherworldly objects that are guaranteed to be unseen in training but could plausibly be found in the real world, and maintain real-world concepts of object hierarchy. If a PredNet model truly corresponds well to the brain, as measured by the human-model evaluation metric, we expect the model will have some capacity for cross-domain, cross-task object matching, regardless of the stimuli presented for object detection.

4.1 Evaluation of training hyperparameters

Initially, we evaluated training hyperparameters in order to test their effect on basic PredNet models. We varied six training 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, and the learning rate. Although these initial experiments are focused on the effects of training hyperparameters, we also coarsely varied the size of the convolutional filters across all layers of PredNet to ensure we did not over-fit to one model architecture. Aside from the filter size for each layer, which ranged from 1×1 to 6×6, we searched a broad hyperparameter search space in order to fully understand the effects of training hyperparameters. The specific space we searched can be found in the supplemental material. Ultimately, for this experiment we trained 95 PredNet models with randomly selected hyperparameterizations using HyperOpt [2, 3], a software package for distributed hyperparameter optimization.

We performed a Monte Carlo-style sampling of the search space, which we used to obtain a statistical sample of model performance. In Table 1 we report the mean and standard deviation of our various metrics for the 95 trained PredNet models summarizing this sample. Next frame prediction varied heavily from model-to-model compared with the other metrics, as reflected in the standard deviation, but was within range of the scores reported by Lotter et al. [24]. The accuracy scores highlight the difficulty of the object matching task, which focuses on fine-grained object matching from a 50 image gallery of stimuli (chance = 0.02). All models performed well above chance, indicating that they were not failing on the task, but were well below the theoretical ceiling of 100% accuracy, which allows room for extensive model tuning. Mean human-human similarity of the four human participants across two sessions was 0.19 with a standard deviation of 0.09, indicating the models were within the range of human-human similarity. We reduced human-model similarity noise by averaging participant RDMs across sessions. In Table 1 we also present the mean scores for all evaluation metrics for the ten models with the highest similarity score. By comparison, the bottom ten models' mean next frame prediction error was 0.314 (SD = 0.138), accuracy was 0.13 (0.15),

Table 1: Summarized statistics of evaluation scores and standard deviations for an initial sample of 95 randomly hyperparameterized PredNet models. These scores indicate the range of scores we expect to obtain from an arbitrary PredNet model. The top ten human-model similarity (HMS) mean score refers to the average score for each metric for the ten models with the highest human-model similarity. The top ten models average shows that models with high human-model similarity also achieve high performance on the other tasks. The object matching task was intentionally designed to be difficult — the model must distinguish fine-grained differences in unseen, fictional Gazoobian objects [38] and task chance is (0.02). Models are trained using KITTI [14] and evaluated on next frame prediction using a heldout set of KITTI data. SD is standard deviation. Pixel MSE is mean squared error of the predicted-to-actual frame at the pixel level.

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 2: Pearson correlations of evaluation metrics for 95 trained PredNet models with random hyperparameters. There is a negative correlation between next frame prediction error and Human-model similarity, because mean squared error (MSE) is a measure of error to be minimized while Human-Model similarity and accuracy are metrics to be maximized. Data is the same as described in Table 1. SD is standard deviation. Pixel MSE is mean squared error at the pixel level.

Variable	**Accuracy**	**Similarity**	**Learning Rate**
Next Frame Prediction Error	-0.785^{**}	$\mathbf{-0.565}^{**}$	0.879^{**}
Object Matching Accuracy	.	$\mathbf{0.609}^{**}$	-0.810^{**}
Human-Model Similarity	.	.	-0.682^{**}

$^{**}p < 0.01$

and human-model similarity was -0.008 (0.027). Thus, human-model similarity shows itself to be a strong metric for predicting generalized model performance across domains and tasks.

In Table 2 we present the Pearson correlations of the evaluation metrics across models. We also include the learning rate, which was the only hyperparameter that significantly correlated with any evaluation metric. Note that while object matching accuracy and human-model similarity are metrics that should be maximized, next frame prediction is evaluated by mean squared error (MSE), a metric to be minimized. Thus, a negative correlation with this metric and the others is a good indicator. These results show that human-model similarity is strongly correlated with both object matching accuracy and next frame prediction accuracy across all scores, strengthening the findings from Table 1 that human-model similarity is a strong indicator of performance.

Given these results, we modulated further experiments by limiting the range of model training hyperparameters. This allowed us to both maximize training speed, which provided a wider variation in models, and to focus on identifying variation within architectural hyperparameters. In the supplemental material, we further justify this decision by examining PredNet's evaluation consistency within an architecture. In the supplemental material, we replicate our findings by cross-dataset training on KITTI evaluating on an alternative dataset, (VLOG) [11], and visa versa. VLOG and KITTI next-frame prediction evaluations correlated almost perfectly whether trained on KITTI (0.992) or VLOG (0.999), indicating PredNet next-frame prediction performance is dataset invariant.

4.1.1 Stability and reproducibility

In the previous section, we found that the number of epochs used for training was not significantly correlated with any evaluation metrics. Thus, we needed to experimentally determine the expected number of epochs needed for training in order to obtain consistent evaluation metrics. The goal of these experiments was threefold: understand model performance in relation to training time, investigate model reproducibility, and minimize training time for future experiments. We trained 62 identically hyperparameterized PredNets but randomly varied the number of training epochs between 10 and 500. On average, models were trained for 220 epochs. Across models, the mean human-model similarity was 0.1688 (SD = 0.0137). The low standard deviation confirms the sta-

bility of the human-model similarity metric over epochs. Mean object matching and next frame prediction scores were 0.0072 (SD = 0.0006) and 0.4482 (SD = 0.0451), confirming the stability of these metrics.

We next investigated how quickly scores stabilized. We identified that each metric had a unique behavior that influenced how to best measure stability: accuracy scores were inconsistent model-to-model, while next frame prediction scores continuously decreased before plateauing. Model-to-model accuracy variability stabilized after training for 100 epochs. We confirmed the stability by comparing the standard deviation before 100 epochs (0.0688) and after 100 epochs (0.0323). We identified the next frame prediction plateau by identifying the first model to dip below the average score across models, which trained for 110 epochs. We confirm this value, 0.0072, as the plateau by comparing it against the lowest MSE across the tested models.

Impressively, we found that human-model similarity was consistent across models, even in models trained for as few as 10 epochs. In order to further assess the stability of human-model similarity we grouped 45 models into 14 sets of identically hyperparameterized models trained for an equal number of epochs. We measured the consistency of human-model similarity using the standard deviation of each set. The mean of human-model similarity across the standard deviations of all sets was 0.010 (SD = 0.006).

The results of these experiments indicate that, across metrics, models are highly reproducible given the same training hyperparameters. An investigation into score stability found that while other scores needed ample training time (~100 epochs) human-model similarity did not require a large amount of training time to be effective. This, along with the strong correlation between human-model similarity and model consistency, indicates that human-model similarity could be used as a litmus test to predict a model's success before it has been trained enough to accurately assess performance. This finding, in conjunction with the correlation between high similarity scores and performance on other tasks, indicates use of human-model similarity as an early predictor could lead to large savings in training time during model screening.

4.2 Evaluation of architectural hyperparameters

We also performed a series of experiments to assess how architectural hyperparameter variations effect PredNet model performance. Based on the experiments conducted in Sec. 4.1 we made two model training assumptions in order to minimize training time and limit the influence of non-architectural hyperparameters. First, we limited the number of epochs used for training to vary between 20 and 60 because of the consistency of human-model similarity with minimal training, as established in Sec. 4.1.1. Second, we minimized the search space of all training hyperparameters except for learning rate, which was found to strongly correlate with model performance. No training hyperparameters, defined in Section 4.1, were fixed at a specific value to avoid over-fitting to one training regime. We used a range epochs in order to balance our need for short training times, in order to obtain a large sample, and to allow next frame prediction and accuracy more time to converge, allowing us to confirm the correlations found in Table 2.

Different architectural hyperparameters were used to assess the stability of model-human similarity across architectures. Architectural hyperparameters included filter sizes for each of the PredNet layer components, the number of PredNet layers, and the number of filters per layer (details in supplemental material). The number of hyperparameters in each network varied between 6-20 depending on the number of layers in the model. Thus, we also exponentially increased the number of models we trained to assess these parameters in order to maximize architectural variability. In all, we trained 1811 models. We found the evaluation metrics for this model sample were within the range of metrics from the training hyperparameters sample (Table 1). The full table for this sample is in the supplemental material.

For our 1811-model sample, we note that next frame prediction was on average lower (mean = 0.063; SD = 0.115), indicating better performance, with a lower standard deviation than what is found in Table 1. An examination of the data shows that 20% of the trained models in the training hyperparameters experiment performed extremely poorly at next frame prediction (MSE > 0.20), while only 12% did so for this set of experiments. This shift coincides with an increase in human-model similarity (mean = 0.120; SD = 0.041). We suspect object matching performance (mean = 0.336; SD = 0.041) does not improve for this sample because of our findings on the metric's stability in

Sec. 4.1.1 — the smaller number of training epochs results in decreased performance, but this decrease is offset by the increase in overall model performance within the sample. By comparison, next frame prediction performance was stable, but did not converge. We confirm our impressions about the object matching metric by correlating it with human-model similarity, paralleling the calculation of Table 2 with our new sample. While human-model similarity average rose to 0.120 (SD = 0.041), human-model similarity correlated with object matching dropped to 0.336 ($p < 0.01$), indicating that human-model similarity is less correlated because these models are trained for far fewer epochs. This further implies the stability of the human-model similarity metric.

Correlations between human-model similarity and next frame prediction drop too. Perhaps most interestingly, accuracy and next frame prediction correlation falls to -0.144 ($p < 0.01$). While weaker correlations for these metrics were expected due to the decreased number of training epochs, as discussed in Sec. 4.1.1, this correlation implies that accuracy and next frame prediction are only barely indicative of each other's performance at this level of training. Correlations with both metrics and human-model similarity were weaker: accuracy and HMS 0.336; next frame and HMS -0.467 ($p < 0.01$ for both). Thus human-model similarity is still predictive.

Experiments from Sec. 4.1.1 indicated that these correlations should return to their previous levels if models are trained for more epochs. To confirm this, we conducted two experiments. First, we examined 525 models with three layers, which require less training, and found correlations across all evaluation metrics dramatically increasing as expected: accuracy and HMS 0.431; next frame and HMS -0.593 ($p < 0.01$ for both). Second, we trained a random subsample of 80 models with layers ranging from 3 to 6 for 150 epochs and found human-model similarity correlates with accuracy with a correlation of 0.365 ($p < 0.05$) and mean squared error (MSE) with a correlation of -0.545 ($p < 0.05$), similar to correlations reported in Table 2. The lower correlations indicate that larger models may take more than 150 epochs for next frame prediction and accuracy scores to stabilize. In summary, the human-model similarity metric is shown to be a stable, generalized metric to predict model performance in a cross-task cross-domain context across architectural and training hyperparameters.

4.3 Human-model similarity on a per-layer basis

We compare measurements for human-model similarity at the layer level instead of the model level across 1302 models, as shown in Table 3. We found that, on average, human-model similarity tends to be highest when comparing all layers of the predictive coding network. We also found that earlier layers are more indicative of human-model similarity performance than later layers. This may be related to the broad technique used to obtain human-participant fMRI data, as discussed in Section 3.3. Previously, Yamins et al. [40] found that different layers of CNNs correspond to different ventral cortical regions (V4 and IT). However, this work focused on fine-grained neurological data. Since the fMRI data persists across all such visual areas, it follows that the full model would correlate best.

Given these results, we investigated the use of all layers (rather than just the last) to match objects, and its effect on object matching performance. Correlating accuracy with the model-human similarity for individual layers resulted in a stronger correlation 0.389 ($p < 0.01$) with the last layer compared with the full model 0.218 ($p < 0.01$). We tested this on a subset of 3-layer models ($N = 409$), which were more fully trained than larger layer models due to the minimal number of epochs used for training. This confirms that the last layer is a better indicator of model performance for accuracy. Conversely, human-model similarities of earlier layers is more indicative of performance on next-frame prediction, which is similar to full model human-model similarity performance. These experiments also imply that predictive coding networks share a hierarchical correspondence to vision.

5 Conclusion and Future Work

In this work, we proposed and evaluated human-model similarity, a metric that quantifies similarity between human fMRI data and machine learning models. By studying this similarity metric in conjunction with other metrics on various model sets we discovered that human-model similarity score is a strong indicator of a predictive coding network's cross-domain cross-task performance. We also discovered that human-model similarity quickly converges, in as few as 10 epochs, whereas other metrics require much more training. Our experiments verified these results across hyperparameters of all kinds. Future work in guided model search should study how incorporating this metric could

Table 3: A comparison of the human-model similarity metric when calculated using a full model (our original evaluation source), an alternative full-view model (averaging the determined similarity scores for each individual layer), and the first and last layers alone. HMS is human-model similarity. SD is standard deviation.

PredNet Reference for RDM Construction	Mean HMS (SD)
All Layers (Concatenated)	**0.130 (0.034)**
All Layers (Averaged)	0.103 (0.033)
First Layer	0.114 (0.037)
Last Layer	0.090 (0.048)

vastly increase model search speed while still obtaining reliable results. While this work has focused on predictive coding networks, in essence how biologically-inspired models might become more biological, human-model similarity should also be studied in other machine learning contexts. Finally, we note that fMRI data is expensive and time consuming to collect, but that all of these experiments were conducted using publicly available fMRI data. This removes a major barrier to deployment of our proposed metric. All code and data for this study will be released following publication.

Acknowledgments

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

References

[1] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[2] James Bergstra, Daniel Yamins, and David D. Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *ICML (1)*, 28:115–123, 2013. URL `http://www.jmlr.org/proceedings/papers/v28/bergstra13.pdf`.

[3] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1): 014008, 2015.

[4] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017.

[6] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Smash: one-shot model architecture search through hypernetworks. *arXiv preprint arXiv:1708.05344*, 2017.

[7] Corinna Cortes, Xavi Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Adanet: Adaptive structural learning of artificial neural networks. *arXiv preprint arXiv:1607.01097*, 2016.

[8] Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *IJCAI*, pages 3460–3468, 2015.

[9] Kai-Bo Duan and S Sathiya Keerthi. Which is the best multiclass svm method? an empirical study. In *International Workshop on Multiple Classifier Systems*, pages 278–285. Springer, 2005.

[10] Ruth Fong, Walter Scheirer, and David Cox. Using Human Brain Activity to Guide Machine Learning. *arXiv:1703.05463 [cs]*, March 2017. URL `http://arxiv.org/abs/1703.05463`. arXiv: 1703.05463.

[11] David F. Fouhey, Wei-cheng Kuo, Alexei A. Efros, and Jitendra Malik. From Lifestyle Vlogs to Everyday Interactions. *arXiv preprint arXiv:1712.02310*, 2017.

[12] Frauke Friedrichs and Christian Igel. Evolutionary tuning of multiple svm parameters. *Neurocomputing*, 64:107–117, 2005.

[13] Beatriz A Garro and Roberto A Vázquez. Designing artificial neural networks using particle swarm optimization algorithms. *Computational intelligence and neuroscience*, 2015:61, 2015.

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[15] Ha Hong, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613, 2016.

[16] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An evaluation of sequential model-based optimization for expensive blackbox functions. In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*, pages 1209–1216. ACM, 2013.

[17] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, 6:32672, September 2016. ISSN 2045-2322. doi: 10.1038/srep32672. URL `http://www.nature.com/srep/2016/160907/srep32672/full/srep32672.html`.

[18] Nikolaus Kriegeskorte. Relating Population-Code Representations between Man, Monkey, and Computational Models. *Frontiers in Neuroscience*, 3(3):363–373, 2009. ISSN 1662-453X. doi: 10.3389/neuro.01.035.2009.

[19] Nikolaus Kriegeskorte. Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage*, 56(2):411–421, 2011. URL `http://www.sciencedirect.com/science/article/pii/S1053811911000978`.

[20] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2, November 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605405/`.

[21] Nikolaus Kriegeskorte, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6):1126–1141, December 2008. ISSN 0896-6273. doi: 10.1016/j.neuron.2008.10.043. URL `https://www.cell.com/neuron/abstract/S0896-6273(08)00943-4`.

[22] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Efficient hyperparameter optimization and infinitely many armed bandits. In *ICML 2016 workshop on AutoML (AutoML 2016)*, 2016.

[23] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Domain adaptation with randomized multilinear adversarial networks. *arXiv preprint arXiv:1705.10667*, 2017.

[24] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *Proceedings of the International Conference on Learning Representations*, 2017.

[25] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F. Fei-Fei. Label Efficient Learning of Transferable Representations acrosss Domains and Tasks. In *Advances in Neural Information Processing Systems*, pages 164–176, 2017.

[26] Dougal Maclaurin, David K Duvenaud, and Ryan P Adams. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, pages 2113–2122, 2015.

[27] Patrick McClure and Nikolaus Kriegeskorte. Representational Distance Learning for Deep Neural Networks. *Frontiers in computational neuroscience*, 10:131, 2016.

[28] Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A. Bandettini, and Nikolaus Kriegeskorte. Human Object-Similarity Judgments Reflect and Transcend the Primate-IT Object Representation. *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00128. URL `http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00128/full`.

[29] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Comput Biol*, 10(4):e1003553, 2014. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003553`.

[30] Nicolas Pinto, David Doukhan, James J. DiCarlo, and David D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol*, 5(11):e1000579, 2009. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000579`.

[31] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James 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. *bioRxiv*, page 240614, February 2018. doi: 10.1101/240614. URL `https://www.biorxiv.org/content/early/2018/02/12/240614`.

[32] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

[33] Zhongzheng Ren and Yong Jae Lee. Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery. *arXiv:1711.09082 [cs]*, November 2017. URL `http://arxiv.org/abs/1711.09082`. arXiv: 1711.09082.

[34] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999.

[35] Mark Roper, Chrisantha Fernando, and Lars 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.

[36] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–426, 2007.

[37] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[38] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[39] Daniel Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. In *Nature neuroscience*, volume 19, pages 356 – 365, 2016. URL `http://www.nature.com/neuro/journal/v19/n3/abs/nn.4244.html`.

[40] Daniel Yamins, Ha Hong, Charles Cadieu, and James J. DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in neural information processing systems*, pages 3093–3101, 2013.

[41] Daniel Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James 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. URL `http://www.pnas.org/content/111/23/8619.short`.

[42] Davide Zoccolan, Nadja Oertelt, James J. DiCarlo, and David D. Cox. A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences*, 106(21):8748–8753, 2009. URL `http://www.pnas.org/content/106/21/8748.short`.

[43] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[44] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*, 2017.