Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?

Martin Schrimpf[*,1,2], **Jonas Kubilius**[*,3,4], **Ha Hong**[5], **Najib J. Majaj**[6], **Rishi Rajalingham**[1], **Elias B. Issa**[7], **Kohitij Kar**[1,3], **Pouya Bashivan**[1,3], **Jonathan Prescott-Roy**[1], **Franziska Geiger**[3], **Kailyn Schmidt**[1], **Daniel L. K. Yamins**[8,9], **James J. DiCarlo**[1,2,3]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139
[2]Center for Brains, Minds and Machines, MIT, Cambridge, MA 02139
[3]McGovern Institute for Brain Research, MIT, Cambridge, MA 02139
[4]Brain and Cognition, KU Leuven, Leuven, Belgium
[5]Bay Labs Inc., San Francisco, CA 94102
[6]Center for Neural Science, New York University, New York, NY 10003
[7]Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027
[8]Department of Psychology, Stanford University, Stanford, CA 94305
[9]Department of Computer Science, Stanford University, Stanford, CA 94305

The internal representations of early deep artificial neural networks (ANNs) were found to be remarkably similar to the internal neural representations measured experimentally in the primate brain. Here we ask, as deep ANNs have continued to evolve, are they becoming more or less brain-like? ANNs that are most functionally similar to the brain will contain mechanisms that are most like those used by the brain. We therefore developed *Brain-Score* – a composite of multiple neural and behavioral benchmarks that score any ANN on how similar it is to the brain's mechanisms for core object recognition – and we deployed it to evaluate a wide range of state-of-the-art deep ANNs. Using this scoring system, we here report that: (1) DenseNet-169, CORnet-S and ResNet-101 are the most brain-like ANNs. (2) There remains considerable variability in neural and behavioral responses that is not predicted by any ANN, suggesting that no ANN model has yet captured all the relevant mechanisms. (3) Extending prior work, we found that gains in ANN ImageNet performance led to gains on Brain-Score. However, correlation weakened at $\geq 70\%$ top-1 ImageNet performance, suggesting that additional guidance from neuroscience is needed to make further advances in capturing brain mechanisms. (4) We uncovered smaller (i.e. less complex) ANNs that are more brain-like than many of the best-performing ImageNet models, which suggests the opportunity to simplify ANNs to better understand the ventral stream. The scoring system used here is far from complete. However, we propose that evaluating and tracking model-benchmark correspondences through a Brain-Score that is regularly updated with new brain data is an exciting opportunity: experimental benchmarks can be used to guide machine network evolution, and machine networks are mechanistic hypotheses of the brain's network and thus drive next experiments. To facilitate both of these, we release Brain-Score.org: a platform that hosts the neural and behavioral benchmarks, where ANNs for visual processing can be submitted to receive a Brain-Score and their rank relative to other models, and where new experimental data can be naturally incorporated.

computational neuroscience | object recognition | deep neural networks

Correspondence: ***mschrimpf@mit.edu (M.S.), qbilius@mit.edu (J.K), dicarlo@mit.edu (J.J.D.)***

[*] **Equal contribution**

Introduction

Deep convolutional artificial neural networks (ANNs) (LeCun et al., 2015) were derived in part from findings in visual neuroscience (see Yamins and DiCarlo (2016) for review) and are now the leading models in machine vision and other areas of AI. Soon after Krizhevsky et al. (2012)'s initial results, it was found that by evolving deep ANNs to achieve gains in performance (through either architecture search or weight training), those ANNs developed internal feature representations that are remarkably similar to neural representations recorded in mid and high levels of the non-human primate ventral visual processing stream (Yamins et al., 2013, 2014; Khaligh-Razavi and Kriegeskorte, 2014) (see Yamins and DiCarlo (2016) for review). More recent work has extended this same "performance-driven" ANN approach to the human visual system (Güçlü and van Gerven, 2015; Cichy et al., 2016; Kubilius et al., 2016), to lower levels of visual processing (Cadena et al., 2017), to auditory processing (Kell et al., 2018), and to the rodent tactile system (Zhuang et al., 2017).

The models from this early work outlined above outperformed all other neuroscience models at the time and yielded reasonable scores on predicting response patterns from both single unit activity and fMRI. It was also suggested that ANNs with improved task performance were likely to be even better matches to the primate visual stream (Yamins et al. (2014); Yamins and DiCarlo (2016)). We thus ask, as model performance has increased from Alexnet's 57.67% top-1 on ImageNet to up to 85.4% (Mahajan et al., 2018)[1] today, are these even better models of the primate visual stream?

To answer this question, we here propose *Brain-Score* to evaluate any ANN on how brain-like it is – focusing on the parts of the brain that have been implicated in visual object recognition. Brain-Score is a composite benchmark consisting of neural and behavioral benchmarks, where each benchmark refers to the application of a metric to a particular dataset. Neural metrics assess the similarity of internally observable signals: image-evoked feature activations in ANNs and image-evoked recorded neural activations in different primate brain regions. Behavioral metrics assess the similarity of the "outputs" of ANNs and primates, such as predictions on match-to-sample tasks. For this study, we assembled a base set of neural and be-

[1]This model was unfortunately unavailable at the time of writing and we thus excluded it from the following analyses. The best ImageNet model included here is PNASNet with 82.9% top-1 performance.

havioral benchmarks: neural recordings from cortical areas V4 and IT in macaque monkeys and behavioral data from humans. We then evaluated dozens of state-of-the-art deep ANNs on these three brain benchmarks and the resulting Brain-Score. We note that while neuroscience models have in the past been influential as building blocks of today's deep neural networks, we do not claim that Brain-Score will automatically yield better models for machine learning. This is a benchmark for the brain sciences, encouraging a quantified evaluation of models on neural and behavioral data. Further, while benchmarking is common in machine learning, most of neuroscience still lacks standardized datasets and tools to perform such benchmarks routinely, and there is little awareness of the benefits of a rigorous evaluation of proposed models of brain mechanisms. Brain-Score is our attempt to bridge this gap and provide the means to track progress in brain-like model development on a scale that is larger than individual laboratories.
Our main contributions are the following:

- We replicate prior work (Yamins et al., 2014) showing that ANNs that have higher ImageNet performance tend to be more functionally similar to the ventral visual stream, and we extend that work by demonstrating that many state-of-the-art deep ANNs simultaneously score well on all three of the brain benchmarks (V4, IT, and behavior).

- We report that correlation between ImageNet performance and neural data prediction is weak for recent models (i.e. those with ImageNet top-1 performance $\geq 70\%$), but variability between models appears non-trivial. In other words, some ANNs appear to predict neural responses better than others but not because they perform better on ImageNet.

- We show that an ANN's ImageNet performance correlates robustly with behavioral metrics, meaning that the image-by-image patterns of behavior of high-performing ANNs mostly resemble and predict those of primates. Similar to predicting neural response data, there is significant variability among ANNs, and the ANNs with the highest ImageNet performance (NASNet and PNASNet) predict behavioral data considerably worse than older models such as ResNet-101.

- We identify DenseNet-169, CORnet-S (a new shallow recurrent network) and ResNet-101 as the current top three models of the mechanisms underlying primate object recognition (under our current set of benchmarks).

- To enable fast evaluations of neural networks on brain data, we release a platform, Brain-Score.org, that hosts the neural and behavioral data and accompanying metrics, where ANNs for visual processing can easily be submitted to receive a Brain-Score and their rank relative to other models. The Brain-Score and the ranking of models will be updated regularly as new experimental benchmarks are added and new ANNs become available. We also plan to include anatomical benchmarks in a future update of Brain-Score. This platform can easily be extended with new data, such as human fMRI recordings.

Stepping back, we suggest that evaluating and tracking the anatomical, neural, and behavioral correspondences through a Brain-Score that is regularly updated with new brain data is an exciting opportunity: brain measurements can be used to guide ANN evolution to emulate brain functions not yet captured by ANNs and to make ANNs that are more human-like in their patterns of success and failures. For the brain and cognitive sciences, the ANN that best emulates the brain simultaneously becomes the current best understanding of how the brain actually works, and the driver of next experiments.

Brain Benchmarks

In the following section we outline the benchmarks that models are measured against. A benchmark consists of a metric applied to a specific set of experimental data, which here can be either neural recordings or behavioral measurements.

Neural. The purpose of neural metrics is to establish how well internal representations of a source system (e.g., a neural network model) match the internal representations in a target system (e.g., a primate). Unlike typical machine learning benchmarks, these metrics provide a principled way to prefer some models over others even if their outputs are identical. We outline here one common metric, Neural Predictivity, which is a form of a linear regression.

Neural Predictivity: Image-Level Neural Consistency Neural Predictivity is used to evaluate how well responses $\mathbf{X}$ to given images in a source system (e.g., a deep ANN) predict the responses in a target system (e.g., a single neuron's response in visual area IT). As inputs, this metric requires two assemblies of the form stimuli $\times$ neuroid where neuroids can either be neural recordings or model activations. First, source neuroids are mapped to each target neuroid using a linear transformation:

$$y = \mathbf{X}w + \epsilon,$$

where w denotes linear regression weights and ϵ is the noise in the neural recordings. This mapping procedure is performed on multiple train-test splits across stimuli. In each run, the weights are fit to map from source neuroids to a target neuroid using training images, and then using these weights predicted responses y' are obtained for the held-out images. We used the neuroids from V4 and IT separately to compute these fits. To obtain a neural predictivity score for each neuroid, we compare predicted responses y' with the measured neuroid responses y by computing the Pearson correlation coefficient r:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(y'_i - \overline{y}')}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 (y'_i - \overline{y}')^2}} \quad \textbf{(1)}$$

A median over all individual neuroid neural predictivity values (e.g., all measured target sites in a target brain region) is

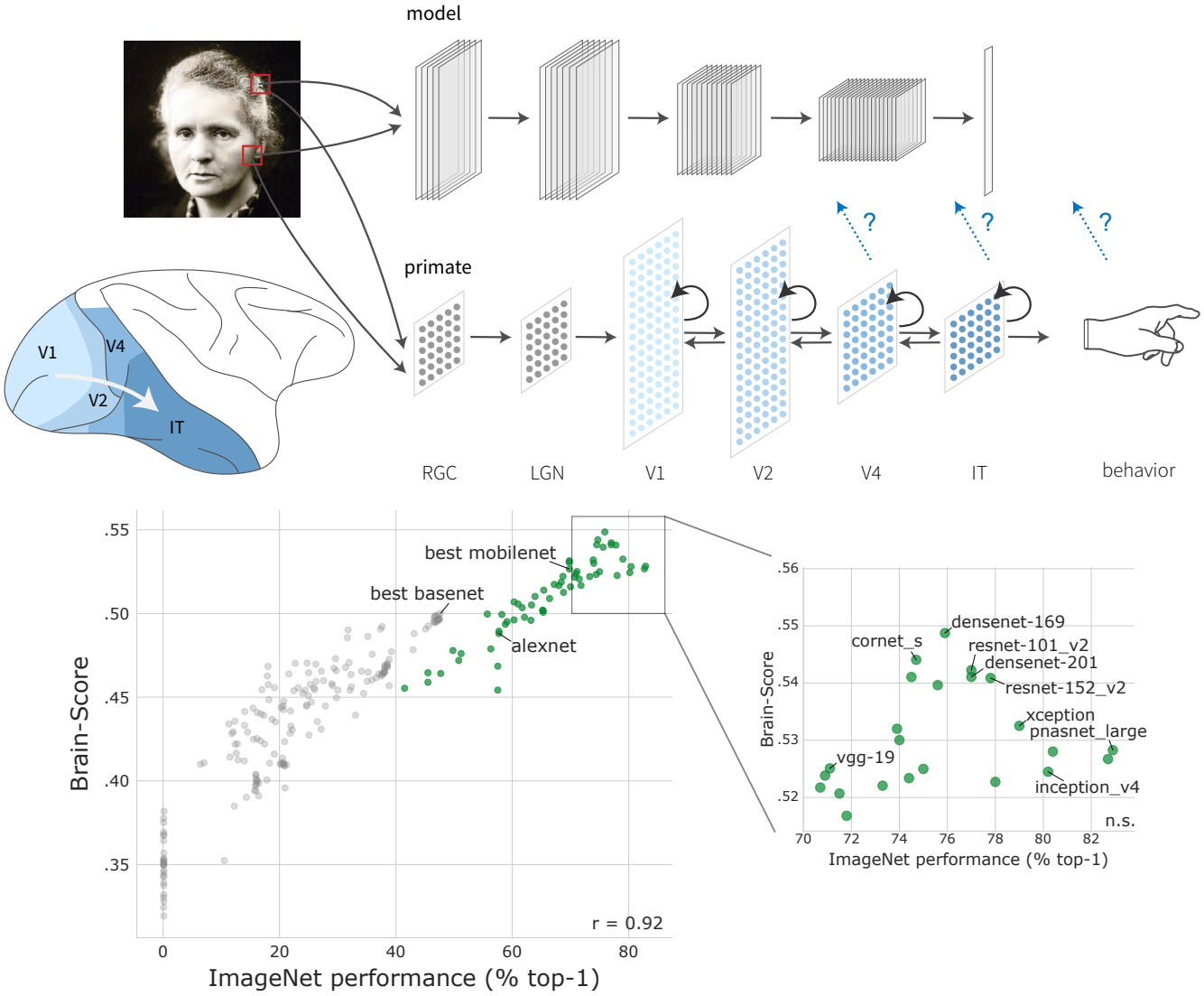

Figure 1: **Overview of the Brain-Score.** We compare neural networks using two classes of metrics: neural metrics compare the internal activations to regions of the macaque ventral stream, and behavioral metrics compare the similarity in outputs. Brain-Score is correlated with ImageNet performance for small, randomly combined models (gray dots) but becomes weak for current state-of-the-art models (green dots) at $\geq 70\%$ top-1 performance.

computed to obtain a predictivity score for that train-test split (median is used since responses are typically distributed non-normally). The final neural predictivity score for the target brain region is computed as the mean across all train-test splits.

We further estimate the internal consistency between neural responses by splitting neural responses in half across repeated presentations of the same image and computing Spearman-Brown-corrected Pearson correlation coefficient (Eq. 1) between the two splits across images for each neuroid.

In practice, we found that standard linear regression is comparably slow given a large dimensionality of the source system and not sufficiently robust. Thus, following Yamins et al. (2014), we use a partial least squares (PLS) regression with 25 components. We further optimized this procedure by first projecting source features into a lower-dimensional space using principal components analysis. The projection matrix is obtained for the features of a selection of ImageNet images, so that the projection is constant across train-test splits. This projection matrix is then used to transform source features. Results reported here were obtained by retaining 1000 principal components from the feature responses per layer to 1000 ImageNet validation images that captured the most variance of a source model.

Neural Recordings The neural dataset currently used in both neural benchmarks included in this version of Brain-Score is comprised of neural responses to 2,560 naturalistic stimuli in 88 V4 neurons and 168 IT neurons (cf. Fig. 1), collected by Majaj et al. (2015). The image set consists of 2,560 grayscale images in eight object categories (animals, boats, cars, chairs, faces, fruits, planes, tables). Each category contains eight unique objects (for instance, the "face" category has eight unique faces). The image set was generated by pasting a 3D object model on a naturalist background. In each image, the position, pose, and size of an object was randomly selected in order to create a challenging object recognition

task both for primates and machines. A circular mask was applied to each image (see Majaj et al. (2015) for details on image generation).

Two macaque monkeys were implanted three arrays each, with one array placed in area V4 and the other two placed on the posterior-anterior axis of IT cortex. The monkeys passively observed a series of images (100 ms image duration with 100 ms of gap between each image) that each subtended approximately 8 deg visual angle. To obtain a stable estimate of the neural responses to each image, each each image was re-tested about 50 times (re-tests were randomly interleaved with other images). In the benchmarks used here, we used an average neural firing rate (normalized to a blank gray image response) in the window between 70 ms and 170 ms after image onset where the majority of object category-relevant information is contained (Majaj et al., 2015).

Behavioral. The purpose of behavioral benchmarks it to compute the similarity between source (e.g., an ANN model) and target (e.g., human or monkey) behavioral responses in any given task. For core object recognition tasks, primates (both human and monkey) exhibit behavioral patterns that differ from ground truth labels. Thus, our primary benchmark here is a behavioral response pattern metric, not an overall accuracy metric, and higher scores are obtained by ANNs that produce and predict the primate patterns of successes and failures. One consequence of this is that ANNs that achieve 100% accuracy will not achieve a perfect behavioral similarity score.

Even within the visual behavioral domain of core object recognition, there are many possible behavioral metrics. We here use the metric of the image-by-image patterns of difficulty, broken down by the object choice alternatives (termed *I2n*), because recent work (Rajalingham et al., 2018) suggests that it has the most power to distinguish among alternative ANNs (assuming that sufficient amounts of behavioral data are available).

I2n: Normalized Image-Level Behavioral Consistency Source data (model features) for a total of i images are transformed first into a $i_b \times c$ matrix of c object categories and i_b images with behavioral data available using the following procedure. First, images where behavioral responses are not available (namely, $i - i_b$ images) are used to build a c-way logistic regression from source data to a c-value probability vector for each image, where each probability is the probability that a given object is in the image. This regression is then used to estimate probabilities for the held-out i_b images. For each image, all normalized target-distractor pair probabilities are computed from the c-way probability vector. For instance, if an image contains a dog and the distractor is a bear, the target-distractor score is $\frac{p(\text{dog})}{p(\text{dog})+p(\text{bear})}$.

In order to compare source and target data, we first transform these raw accuracies in the $i_b \times c$ response matrix to a d' measure for each cell in the $i_b \times c$ matrix:

$$d' = Z(\text{Hit Rate}) - Z(\text{False Alarms Rate}),$$

where Z is the estimated z-score of responses, Hit Rate is the accuracy of a given target-distractor pair while the False Alarms Rate corresponds to how often the observers incorrectly reported seeing that target object in images where another object was presented. For instance, if a given image contains a dog and distractor is a bear, the Hit Rate for the dog-bear pair for that image comes straight from the $i_b \times c$ matrix, while in order to obtain the False Alarms Rate, all cells from that matrix that did not have dogs in the image but had a dog as a distractor are averaged, and 1 minus that value is used as a False Alarm Rate. All d' above 5 were clipped. This transformation helps to remove bias in responses and also to diminish ceiling effects (since many primate accuracies were close to 1), but empirically observed benefits of d' in this dataset are small; see Rajalingham et al. (2018) for a thorough explanation.

The resulting response matrix is further refined by subtracting the mean Hit Rate across trials of the same target-distractor pair (e.g., for dog-bear trials, their mean is subtracted from each trial). Such normalization exposes variance unique to each image and removes global trends that may be easier for models to capture. For instance, dog-bear trials on average could have been harder than dog-zebra trials. Without this normalization, a model might score very well by only capturing this tendency. After normalization, all responses are centered around zero, and thus capturing only global trends but not each image's idiosyncrasies would be insufficient for a model to rank well.

After normalization, a Pearson correlation coefficient r_{st} between source and target data is computed using Eq. 1. We further estimate noise ceiling, that is, how well an ideal model could perform given the noise in the measured behavioral responses, by dividing target data in half across trials, computing the normalized d' $i_b \times c$ matrices for each half, and computing the Pearson correlation coefficient r_{tt} between the two halves. If source data is produced by a stochastic process, the same procedure can be carried out on the source data, resulting in the source's reliability r_{ss}.

The final behavioral predictivity score of each ANN is then computed by:

$$r = \frac{r_{st}}{\sqrt{r_{ss} r_{tt}}}$$

All models that we tested so far produced deterministic responses, thus $r_{ss} = 1$ in our scoring.

Primate behavioral data The behavioral data used in the current round of benchmarks was obtained by Rajalingham et al. (2015) and Rajalingham et al. (2018). Here we focus on only the human behavioral data, but the human and non-human primate behavioral patterns are very similar to each other (Rajalingham et al., 2015, 2018).

The image set used in this data collection was generated in a similar way as the images for V4 and IT using 24 object categories. In total, the dataset contains 2,400 images (100 per object). For this benchmark, we used 240 (10 per object) of these images for which the most trials were obtained. 1,472 human observers responded to briefly presented images on Amazon Mechanical Turk. At each trial, an image was presented for 100 ms, followed by two response choices, one

corresponding to the target object present in the image and the other being one of the remaining 23 objects (i.e., a distractor object). Participants responded by choosing which object was presented in the image. Thus, over three hundred thousand responses for each target-distractor pair were obtained from multiple participants, resulting in a 240 (images) $\times$ 24 (objects) response matrix when averaged across participants.

Brain-Score. To evaluate how well a model is doing overall, we computed the global Brain-Score as a composite of neural V4 predictivity score, neural IT predictivity score, and behavioral I2n predictivity score (each of these scores was computed as described above). The Brain-Score presented here is the mean of the three scores. This approach does not normalize by different scales of the scores so it may be penalizing scores with low variance but it also does not make any assumptions about significant differences in the scores, which would be present in ranking.

Candidate Models

For this round of evaluations, we sought to benchmark most commonly used neural network families: AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2015), Inception (Szegedy et al., 2015a,b, 2016), InceptionResNet (Szegedy et al., 2016), SqueezeNet (Iandola et al., 2016), DenseNet (Huang et al., 2017), MobileNet (Howard et al., 2017), and (P)NASNet (Zoph and Le, 2016; Liu et al., 2017). Most of pre-trained models were available in TensorFlow (Abadi et al., 2016), either via their Keras (Chollet et al., 2015) or Slim interface. For AlexNet, SqueezeNet, ResNet-18 and ResNet-34, we used their PyTorch implementation (Paszke et al., 2017).

To further map out the space of possible architectures and a baseline of neural, behavioral, and performance scores, we included an in-house-developed family of models with up to moderate ImageNet performance, termed *BaseNets*: lightweight AlexNet-like architectures with six convolutional layers and a single fully-connected layer, captured at various stages of training. Various hyperparameters were varied between BaseNets, such as the number of filter maps, nonlinearities, pooling, learning rate scheduling, and so on, and formed a basis for the CORnet family of models (Kubilius et al., 2018b).

We also tested CORnet-S, a new model that was developed with the goal of rivaling the best models on *Brain-Score* while being significantly shallower than competitors by leveraging bottleneck architecture and recurrence (Kubilius et al., 2018b). CORnet-S is composed of four recurrent areas with two to three convolutions each and a fully-connected layer at the end. For Neural Predictivities, we used activations at multiple internal layers of the networks. Layers were pre-selected by hand to include layers at multiple depths in each model and respecting the natural structuring (e.g., the outputs of a ResNet block were used, not the internal activations within the block). To keep the regression manageable, features were further downsampled with PCA to 1,000 dimensions. After testing every layer on both V4 and IT, we report the model's score as the score of the best layer per region. Going forward, we are imagining more flexible methods for mapping model layers to brain regions, such as combining the activations of multiple layers. For CORnet-S, which already commits to a mapping to brain regions, we use the pre-defined mapping of the model. Behavioral scores were obtained using the final pre-readout layer of a network (i.e., the layer just before the last weight layer after which features are transformed into 1,000-dimensional outputs specific to the ImageNet task). In this case, features were not downsampled because typically dimensionality of the readout layer was sufficiently low to compute the scores quickly.

Results

We examined a wide range of deep neural network trained on ImageNet and compared their internal representations with neural recordings in non-human visual cortical areas V4 and IT and with human behavioral measurements.

Ranking of state-of-the-art networks. Table 1 summarizes the scores for each model on the range of brain benchmarks. The Brain-Score against ImageNet performance is shown in Fig. 1. The strongest model under our current set of benchmarks is DenseNet-169 with a Brain-Score of .549, closely followed by CORnet-S with a Brain-Score of .544 and ResNet-101 with a Brain-Score of .542. The current top-performing models on ImageNet from the machine learning community all stem from the DenseNet and ResNet families of models. DenseNet-169 and ResNet-101 are also among the highest-scoring models on the IT neural predictivity and the behavioral predictivity respectively with scores of .604 on IT (DenseNet-169, layer *conv5_block16_concat*) and .378 on behavior ResNet-101, layer *avg_pool*). VGG families win V4 with a score of .672 for VGG-19 (layer *block3_pool*).

The best models from the BaseNet baseline family of models lag behind the winning models with a Brain-Score of .500 and a behavioral score of .256 but still perform reasonably well on V4 (.654) and IT (.592). Several observations for other model families are also worth noting: while ANNs from the Inception architectural family improved on ImageNet performance over subsequent versions, its Brain-Score decreased. Another natural cluster emerged with AlexNet and SqueezeNet at the bottom of the ranking: despite reasonable scores on V4 and IT neural predictivity, their behavioral scores are sub-par.

Interestingly, models that score high on brain data are also not the ones ranking the highest on ImageNet performance, suggesting a potential disconnect between ImageNet performance and fidelity to brain mechanisms. For instance, despite its superior performance of 82.90% top-1 accuracy on ImageNet, PNASNet only ranks 13$^{\text{th}}$ on the overall Brain-Score. Models with an ImageNet top-1 performance below 70% show a strong correlation with Brain-Score of .92 ($p < 10^{-14}$) but above 70% ImageNet performance, there was no significant correlation ($p >> .05$, cf. Fig. 1). To investigate this potential disconnect further, we next analyzed the specific scores on neural and behavioral data.

Brain-Score	model	neural predictivity V4	neural predictivity IT	behavioral predictivity	top-1 accuracy ImageNet
.549	densenet-169	.663	**.606**	.378	75.90
.544	cornet_s	.650	.600	.382	74.70
.542	resnet-101_v2	.653	.585	**.389**	77.00
.541	densenet-201	.655	.601	.368	77.00
.541	densenet-121	.657	.597	.369	74.50
.541	resnet-152_v2	.658	.589	.377	77.80
.540	resnet-50_v2	.653	.589	.377	75.60
.533	xception	.671	.565	.361	79.00
.532	inception_v2	.646	.593	.357	73.90
.532	inception_v1	.649	.583	.362	69.80
.531	resnet-18	.645	.583	.364	69.76
.530	nasnet_mobile	.650	.598	.342	74.00
.528	pnasnet_large	.644	.590	.351	**82.90**
.528	inception_resnet_v2	.639	.593	.352	80.40
.527	nasnet_large	.650	.591	.339	82.70
.527	best mobilenet	.613	.590	.377	69.80
.525	vgg-19	**.672**	.566	.338	71.10
.524	inception_v4	.628	.575	.371	80.20
.523	inception_v3	.646	.587	.335	78.00
.522	resnet-34	.629	.559	.378	73.30
.521	vgg-16	.669	.572	.321	71.50
.500	best basenet	.652	.592	.256	47.64
.488	alexnet	.631	.589	.245	57.70
.469	squeezenet1_1	.652	.553	.201	57.50
.454	squeezenet1_0	.641	.542	.180	57.50

Table 1: Brain-Scores and individual performances for state-of-the-art models

Scores on individual neural and behavioral benchmarks. Previous studies observed that models with higher classification performance tend to better predict neural data (Yamins et al., 2014). Here we extend that work by demonstrating that this performance-driven approach holds in a broad sense when evaluated on multiple deep neural networks in a wide range of ImageNet performance regimes, but fails to produce a network exactly matching the brain when reaching human performance levels (see Fig. 1). On individual scores, the correlation of ImageNet performance and Brain-Score varies substantially (Fig. 2). For instance, V4 single site responses are predicted best not only by VGG-19 (ImageNet top-1 performance 71.10%) but also Xception (79.00% top-1). Similarly, IT single site responses are predicted best by DenseNet-169 (.606; 75.90% top-1) but even BaseNets (.592; 47.64% top-1) and MobileNets (.590; 69.80% top-1) are very close to the same IT neural predictivity score. In contrast, the correlation between ImageNet performance and behavioral predictivity remains robust with AlexNet (57.50% top-1) or BaseNets performing substantially worse than the best models. However, the top-performing models on the behavioral score are not the state-of-the-art models on ImageNet: ResNet-101 ranks the highest on behavioral score (.389) but has 77.37% ImageNet top-1 performance, compared to PNASNet that achieves higher ImageNet performance (82.90% top-1) but a substantially lower behavioral score (.351). In fact, the correlation between ImageNet top-1 performance and behavioral score appears to be weakening, with models performing well on ImageNet exhibiting little correlation to behavioral scores, suggesting that better consistency with behavioral data might not be achieved by continuing the efforts to push ImageNet performance higher. Overall, despite the lack of clear trend at high ImageNet performance regimes, the performance-to-neural correlation is .68 ($p < 10^{-28}$) in V4, .80 ($p < 10^{-47}$) in IT, and performance correlates with behavior at .93 ($p < 10^{-91}$).

While all our current predictivity scores only summarize the average, it is clear that individual image-wise predictions are misaligned. In particular, out of the total $5,520$ images, over half of the images ($3,388$, 61.38%) are already relatively well aligned between PNASNet (the best ImageNet model) and humans, as measured by no more than a $1d'$ difference between human and model predictions. A substantial number of images ($1,918$, 34.75%) are easier for humans than models ($\Delta d' > 1$), meaning that further performance gains will simultaneously improve the behavioral score. However, on some images (214, 3.88%), the model outperforms humans ($\Delta d' < -1$) which might be desirable in a typical machine learning challenge but in fact hurts the model's behavioral score as it tends to make the model more misaligned from humans.

We further analyzed the correlation of the neural scores to the behavioral score to determine the need for all individual benchmarks. We found that there was a moderate but not perfect

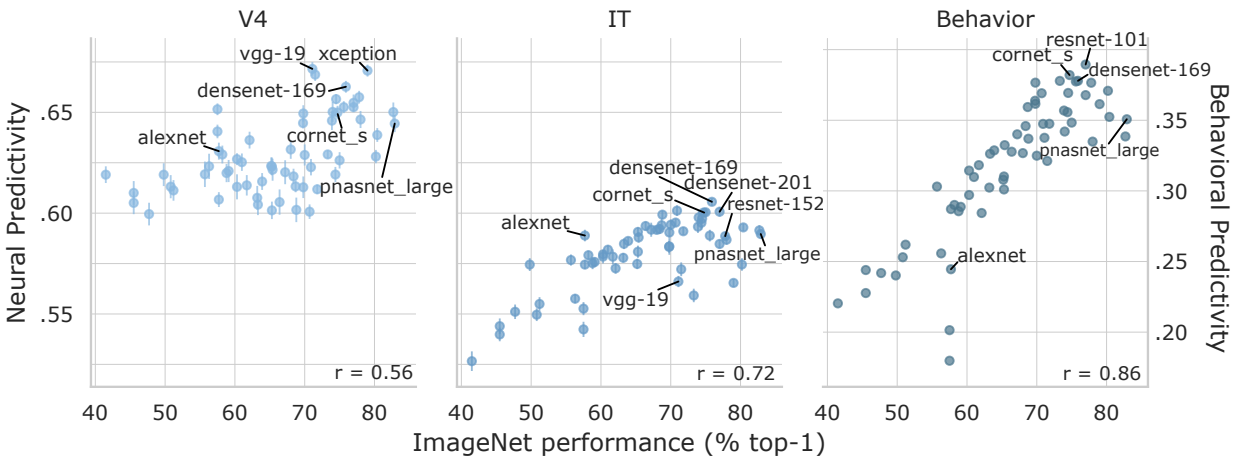

Figure 2: **Predictivities of all models on neural and behavioral benchmarks.** We evaluated regions V4 and IT using the *neural predictivity* as well as behavioral recordings using *I2n*. Current best models are: VGG-19 on V4, DenseNet-169 on IT and ResNet-101 on behavior. Notably, DenseNet-169, CORnet-S and ResNet-101 are strong models on all three benchmarks. Noise ceilings are .892 for V4, .817 for IT and .497 for behavior. Error bars indicate s.e.m.

correlation (.66 for behavior-to-V4 and .86 for behavior-to-IT) which provides justification for the full composite set of benchmarks outlined here. Put another way, this result confirms that even these first neural scores provide additional constraints on the mechanisms of the ventral stream beyond our high-resolution behavioral score.

Current models still fall short of reaching benchmark ceilings: The best ANN model V4 predictivity score is .663, which is below the internal consistency ceiling of these V4 data (.892). The best ANN model IT predictivity score is .604, which is below the internal consistency ceiling of these IT data (.817). And the best ANN model behavioral predictivity score is .378, which is below the the internal consistency ceiling of these behavioral data (.497).

Discussion

We here present an initial framework for quantitatively comparing any artificial neural network to the brain's neural network for visual processing. With even the relatively small number of brain benchmarks that we have included so far, the framework already reveals interesting patterns: It extends prior work showing that performance correlates with brain similarity, and our analysis of state-of-the-art networks yielded DenseNet-169, CORnet-S and ResNet-101 as the current best models of the primate visual stream. On the other hand, we also find a potential disconnect between ImageNet performance and Brain-Score: many of the best ImageNet models fall behind other models on Brain-Score, with the winning DenseNet-169 not being the best ImageNet model, and even small networks ("BaseNets") with poor ImageNet performance achieving reasonable scores.

We do not believe that our initial set of chosen metrics is perfect, and we expect the metrics to evolve in several ways:

By including more data of the same type used here. More neural sites collected with even the same set of images will provide more independent data samples, ensuring that models do not implicitly overfit a single set of benchmarks. Moreover, more data from more individuals will allow us to better estimate between-participant variability (i.e., the noise ceiling), establishing the upper bound of where models could possibly be (see below).

By acquiring the same types of data using new images. Presently, our datasets use naturalistic images, generated by pasting objects on a random backgrounds. While these datasets are already extremely challenging, we will more stringently be able to test model ability to generalize beyond its training set by expanding our datasets to more classes of images (e.g., photographs, distorted images (Geirhos et al., 2018), artistic renderings (Kubilius et al., 2018a), images optimized for neural responses (Bashivan et al., 2018)).

By acquiring the same types of data from other brain regions. The current benchmarks include V4, IT and behavioral readouts, but visual stimuli are first processed by the retina, LGN, V1 and V2 in the ventral stream. Including spiking neural data from these regions further constrains models in their early processing. Moreover, top-down modulation and control warrants recordings outside the ventral stream in regions such as PFC.

By adding qualitatively new types of data. Our current set of neural responses consists of recordings from implanted electrode arrays, but in humans, fMRI recordings are much more common. Local Field Potential (LFP), ECoG, and EEG/MEG could also be valuable sources of data. Moreover, good models of the primate brain should not only predict neural and behavioral responses but should also match brain structure (anatomy) in terms of number of layers, their order, connectivity patterns, ratios of numbers of neurons in different areas, and so on. Finally, to scale this framework to a more holistic view of the brain, adding benchmarks for other tasks and domains outside of core object recognition is essential.

By providing better experimental estimates of the ceilings of each component score. Note that it is still difficult to establish whether the ANN models are truly plateauing in their brain similarity – as implied in the results presented above – or if we are observing the limitations of our experimental datasets. For instance, neural ceilings only reflect the internal consistency of individuals neurons and, in that sense, are only an upper bound on the ceiling. That is, those neural responses are collected from individual monkeys, and it may be unreasonable to expect a single model to correctly predict every monkey's neuron responses. A more reasonable ceiling might therefore need to reflect the consistency of an *average* monkey, leaving individual variabilities aside. However, in typical neuroscience experiments, recordings from only two monkeys are obtained, making it currently impossible to directly estimate these potentially lower ceilings.

Behavioral ceilings, on the other hand, might not be prone to such ceiling effects as they are already estimated using multiple humans responses (i.e. the "pooled" human data, see Rajalingham et al. (2015, 2018)). However, reaching consistency with the pooled human behavioral may not be the only way that one might want to use ANN models to inform brain science, as the across-subject variation is also an important aspect of the data that models should aim to inform on.

By developing new ways to compute the similarity between models and data. Besides computing neural predictivity, there are multiple possible ways and particular parameter choices. Others have used for instance different versions of linear regression (Agrawal et al., 2014), RDMs (Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016) or GLM (Cadena et al., 2017). We see neural predictivity as the current strongest form of comparing neural responses because it maps between the two systems and makes specific predictions on a spike-rate level. One could also use entirely new types of comparison, such as precise temporal dynamics of neural responses that are ignored here, even though they are likely to play an important role in brain function (Wang et al., 2018), or causal manipulations that may constrain models more strongly (Rajalingham and DiCarlo, 2018).

By developing brain scores that are tuned separately for the non-human primate and the human. Our current set of benchmarks consist of recordings in macaques and behavioral measurements in humans and models are thus implicitly assumed to fit both of these primates. We do not believe that one ANN model should ultimately fit both species, so we imagine future versions of Brain-Score will treat them separately.

We caution that while Brain-Score reveals *that* one model is better than another, it does not yet reveal *why* that is the case. Due to current experimental constraints, we are not yet able to use Brain-Score to actually train a model. Both of these are key goals of our ongoing work.

To aid future efforts of aligning neural networks and the brain, we are building tools that allow researchers to quickly get a sense how their model scores against the available brain data on multiple dimensions, as well as compare against other models. Researchers can use our online platform Brain-Score.org to obtain all available brain data, submit new data and score their models on standardized benchmarks. The online platform provides an interface for submitting candidate models which are then automatically run on the current version of all benchmarks (code open-sourced at github.com/brain-score) and notify the submitting user about scores.

By providing this initial set of benchmarks we hope to ignite a discussion and further community-wide efforts around even better metrics, brain data and models. In this respect, our field is far closer to the beginning than the end, but it is important to get started and this is our version of such a start. We hope that Brain-Score will become a way of keeping track of computational models of the brain in terms of "how close we are" and quickly identifying the strongest model for a specific benchmark.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 705498 (J.K.), US National Eye Institute (R01-EY014970, J.J.D.), Office of Naval Research (MURI-114407, J.J.D), the Simons Foundation (SCGB [325500, 542965], J.J.D). This work was also supported in part by the Semiconductor Research Corporation (SRC) and DARPA.

Author Contributions. M.S., J.P.R., F.G., and J.J.D. designed the platform. H.H., N.M., and K.S. collected neural data. E.B.I., K.K., R.R., and K.S. collected behavioral data. J.K., P.B., H.H., N.M., E.B.I., R.R., P.B., D.L.K.Y, and J.J.D developed metrics and benchmarks. M.S., J.K., H.H., J.P.R, and D.L.K.Y implemented metrics and benchmarks. M.S. and J.K. compared models. M.S, J.K., and J.J.D. wrote the paper.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint*, mar 2016. URL http://arxiv.org/abs/1603.04467.

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep ann image synthesis. In *Cognitive Computational Neuroscience*, 2018. URL https://ccneuro.org/2018/Papers/ViewPapers.asp?PaperNum=1222.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *bioRxiv*, page 201764, 2017.

François Chollet et al. Keras. https://keras.io, 2015.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schuett, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv preprint*, 2015. URL http://arxiv.org/abs/1512.03385.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural

Networks for Mobile Vision Applications. *arXiv preprint*, apr 2017. URL http://arxiv.org/abs/1704.04861.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, aug 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.243. URL http://arxiv.org/abs/1608.06993.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv preprint*, 2016. ISSN 0302-9743. doi: 10.1007/978-3-319-24553-9. URL http://arxiv.org/abs/1602.07360.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 2018.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems (NIPS)*, 2012. ISSN 10495258. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007. URL https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

Jonas Kubilius, Kohitij Kar, Kailyn M Schmidt, and James J DiCarlo. Can deep neural networks rival human ability to generalize in core object recognition? In *Cognitive Computational Neuroscience*, 2018a. URL https://ccneuro.org/2018/Papers/ViewPapers.asp?PaperNum=1234.

Jonas Kubilius, Martin Schrimpf, and James DiCarlo. CORnet: Modeling Core Object Recognition. *arXiv preprint*, sep 2018b.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive Neural Architecture Search. *arXiv preprint*, 2017. URL https://arxiv.org/pdf/1712.00559.pdfhttp://arxiv.org/abs/1712.00559.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018.

Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418, sep 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5181-14.2015. URL http://www.ncbi.nlm.nih.gov/pubmed/26424887http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4588611http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5181-14.2015.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Rishi Rajalingham and James J DiCarlo. Reversible inactivation of different millimeter-scale regions of primate it results in different patterns of core object recognition deficits. *bioRxiv*, page 390245, 2018.

Rishi Rajalingham, Kailyn Schmidt, and James J DiCarlo. Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, 35(35):12127–12136, 2015.

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. *Journal of Neuroscience*, pages 0388–18, 2018.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint*, 2014. ISSN 09505849. doi: 10.1016/j.infsof.2008.09.005. URL http://arxiv.org/abs/1409.1556.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, sep 2015a. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298594. URL http://arxiv.org/abs/1409.4842.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint*, 2015b. ISSN 08866236. doi: 10.1109/CVPR.2016.308. URL https://arxiv.org/pdf/1512.00567.pdfhttp://arxiv.org/abs/1512.00567.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv preprint*, 2016. ISSN 01678655. doi: 10.1016/j.patrec.2014.01.008.

Jing Wang, Devika Narain, Eghbal A Hosseini, and Mehrdad Jazayeri. Flexible timing by temporal scaling of cortical responses. *Nature neuroscience*, 21(1):102, 2018.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, feb 2016. ISSN 1097-6256. doi: 10.1038/nn.4244. URL http://www.nature.com/doifinder/10.1038/nn.4244.

Daniel L K Yamins, Ha Hong, and Charles Cadieu. Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. In *Neural Information Processing Systems (NIPS)*, 2013. URL http://papers.nips.cc/paper/4991-hierarchical-modular-optimization-ofhttp://machinelearning.wustl.edu/mlpapers/papers/NIPS2013{_}4991{%}5Cnpapers3://publication/uuid/E90976F4-5E4C-482D-B785-561E5A45B9D2.

Daniel L. K. 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, jun 2014. ISSN 0027-8424. doi: 10.1073/pnas.1403112111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1403112111http://www.ncbi.nlm.nih.gov/pubmed/24812127http://www.pubmedcentral.nih.gov/articlerender.fcgi.

Chengxu Zhuang, Jonas Kubilius, Mitra JZ Hartmann, and Daniel L Yamins. Toward goal-driven neural network models for the rodent whisker-trigeminal system. In *Advances in Neural Information Processing Systems*, pages 2552–2562, 2017.

Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. *arXiv preprint*, nov 2016. URL http://arxiv.org/abs/1611.01578.