BRAIN SCORE: ANN≈BRAIN?

Arnav Mago 2021101074
Pranjali Bishnoi 2021101038

OVERVIEW

- Introduction
- What is Brain Score?
- Neural Benchmarks
- Behavioral Benchmarks
- Candidate Models
- Results
- Ways to improve
- Future work

INTRODUCTION

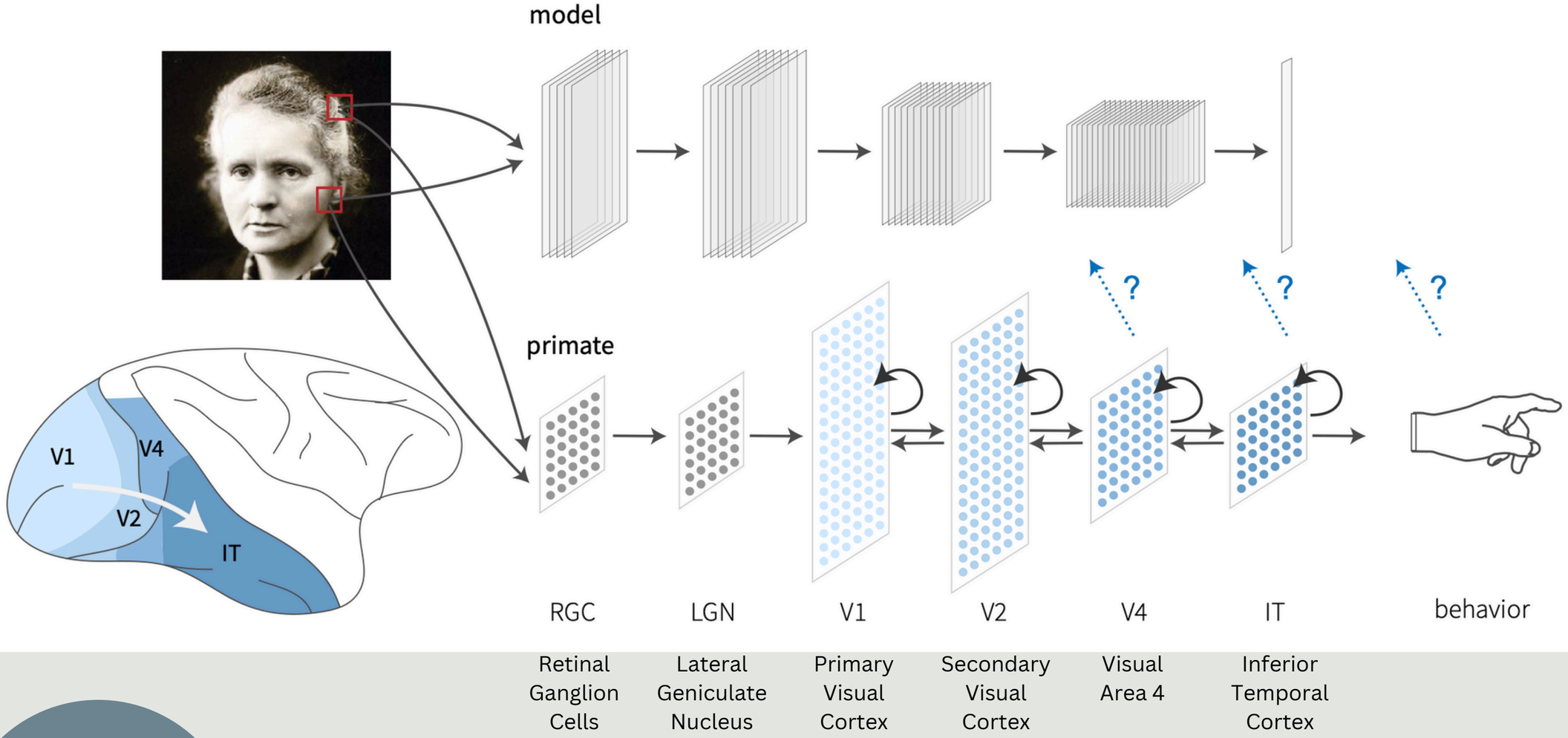
The internal representations of early deep ANNs were found to be remarkably similar to the internal neural representations measured experimentally in the primate brain. The paper evaluates whether deep ANNs have become more or less brain-like as they have continued to evolve.

WHAT IS BRAIN SCORE?

Brain Score is a framework for quantitatively comparing artificial neural networks (ANNs) to the brain's neural mechanisms. It is 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.

The papers theorizes that ANNs that are most functionally similar to the brain will contain mechanisms that are most like those used by the brain.

Neural and Behavior because monkey data was collected via neural


model
primate
V1
V4
V2
IT
?
?
?
RGC
LGN
V1
V2
V4
IT
behavior
Retinal Ganglion Cells
Lateral Geniculate Nucleus
Primary Visual Cortex
Secondary Visual Cortex
Visual Area 4
Inferior Temporal Cortex

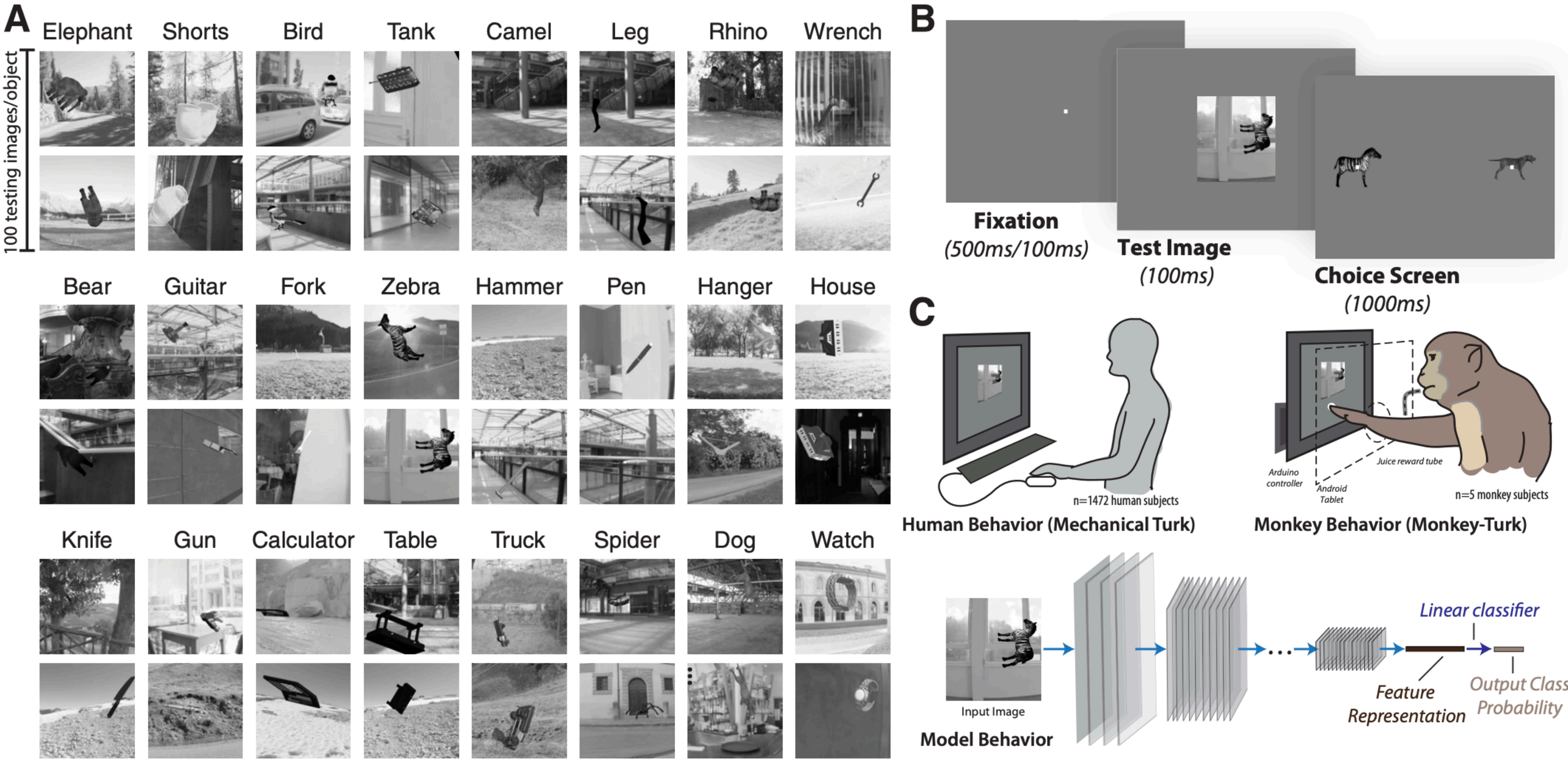
BRAIN BENCHMARKS

Neural

The purpose of the neural metrics is to establish how well the internal representaions of the ANNs match those of the brain

Behavioral

The purpose of behavioral metrics it to compute the similarity between ANN responses and brain (human/primate) behavioral responses in any given task


A
Elephant
Shorts
Bird
Tank
Camel
Leg
Rhino
Wrench
100 testing images/object
Bear
Guitar
Fork
Zebra
Hammer
Pen
Hanger
House
Knife
Gun
Calculator
Table
Truck
Spider
Dog
Watch
B
Fixation
(500ms/100ms)
Test Image
(100ms)
Choice Screen
(1000ms)
C
n=1472 human subjects
Human Behavior (Mechanical Turk)
Arduino controller
Android Tablet
Juice reward tube
n=5 monkey subjects
Monkey Behavior (Monkey-Turk)
Input Image
Model Behavior
Linear classifier
Feature Representation
Output Class Probability

NEURAL BENCHMARKS

Neural Predictivity: Used to predict how well the responses **X** to given images in an ANN match the internal representation in a target system (ex- neurons in the IT region) using a linear transformation.

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

y *= (stimuli x neuroid) in target system,* ***X*** *= (stimuli x neuroid) in ANN,* ***w*** *= regression weights,* ***ε*** *= noise in neural recordings*

They used the neuroids from **V4** and **IT** separately to compute the fits from the source neuroid to the target neuroid.
To obtain a neural predictivity score for each neuroid, they compare predicted responses y'with the measured neuroid responses y by computing the pearson correlation coeff **r**.

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

The median over all neuroid predicted values is the prediction for that run.
The average over all runs is the prediction for the brain region.

NEURAL BENCHMARKS

Neural Recordings: The data is comprised of neural responses to 2,560 naturalistic stimuli (*grayscale images, eight object categories, eight unique objects per category*) in **88 V4 neurons** and **168 IT neurons**.

Two macaque monkeys were implanted three arrays each, with one array placed in area V4 and two placed on the posterior-anterior axis of IT cortex.

The monkeys 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, each image was randomly re-tested about 50 times.

BEHAVIORAL BENCHMARKS

I2n *(Normalized Image-Level Behavioral Consistency):* Measures image-by-image patterns of difficulty, broken down by the object choice alternatives.

Model features for i images are first transformed into a $i_b \times c$ matrix of c object categories and i_b images with behavioral data available.

Images where behavioral responses are not available (namely, $i - i_b$ images) are used to build a c-way logistic regression from model features to a c-dim probability vector for each image, where each element is the probability that a given object is in the image.

For each image, normalized target-distractor pair probs are computed from the probability vector.

If image contains a dog and the distractor is a bear,
target-distractor pair probability: $\frac{p(\text{dog})}{p(\text{dog})+p(\text{bear})}$

BEHAVIORAL BENCHMARKS

To compare the source and target data, convert each cell in the $i_b xc$ matrix to d'.

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

Z *= estimated score of z-responses,* ***Hit Rate*** *= accuracy of a given target-distractor pair,* ***False Alarms Rate*** *= how often the observers incorrectly reported seeing that target object in images where another object was presented*

The resulting response matrix is normalized by subtracting the mean Hit Rate across trials of the same target-distractor pair.

After normalization, a Pearson correlation coefficient $\mathbf{r}_{st}$ between source and target data is computed. We estimate the noise ceiling 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 $\mathbf{r}_{tt}$ 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 $\mathbf{r}_{ss}$.

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

BEHAVIORAL BENCHMARKS

Human Behavioral Data:
The images used were 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 the benchmark, the paper uses 240 (10 per object) of these images for which the most trials were obtained.

1,472 human observers responded to briefly presented images. 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).

Brain Score: The Brain-Score is computed as the mean of the neural V4 predictivity score, neural IT predictivity score, and behavioral I2n predictivity score.

CANDIDATE MODELS

Models Benchmarked:

- AlexNet, VGG, ResNet, Inception, SqueezeNet, DenseNet, MobileNet, and (P)NASNet.

BaseNets Family:

- An in-house-developed family of models with up to moderate ImageNet performance, termed BaseNets
- AlexNet-like architectures with six convolutional layers and a single fully-connected layer.
- Varied hyperparameters laid the foundation for the CORnet family

CORnet-S Model:

- 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. CORnet-S is composed of four recurrent areas with two to three convolutions each and a fully-connected layer at the end

CANDIDATE MODELS

Neural Predictivities:

- Used activations from multiple internal layers, downsampled with PCA to 1,000 dimensions.
- Best-performing layer score reported per region (V4 & IT).

Behavioral Scores:

- Derived from the pre-readout layer without dimensionality reduction.

Future Direction:

- Plan to combine activations from multiple layers for better brain-region mapping.

RESULT

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

RESULT

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.

Top Brain-Score Models: DenseNet-169 (.549) leads, followed closely by CORnet-S (.544) and ResNet-101 (.542) are the the current best models of the primate visual stream.

Neural Predictivity Winners:

- V4 region: VGG-19 (.672, 71.1%) and Xception (.671, 79%)
- IT region: Best predicted by DenseNet-169 (.606, 75.9%), but BaseNets (0.592,47.6%) and MobileNets (0,590,69.8%) perform almost as well.

Behavioral Predictivity Winner:

- ResNet-101 (.389) despite lower ImageNet accuracy (77.37%).
- PNASNet (82.9% top-1) performs worse behaviorally (.351), showing higher ImageNet accuracy doesn’t guarantee better behavioral consistency.

RESULT

ImageNet vs. Brain Performance Disconnect:

- Strong correlation (.92) between ImageNet performance and Brain-Score for models below 70% top-1 accuracy
- No significant correlation for models above 70% ImageNet accuracy
- The winning DenseNet-169 is not the best ImageNet model, and even small networks ("BaseNets") with poor ImageNet performance achieved reasonable scores. Top ImageNet models don't necessarily rank highest on brain benchmarks (e.g. PNASNet ranks 13th on Brain-Score despite 82.90% ImageNet accuracy).
- Models with higher ImageNet performance generally better predict neural data

RESULT

Model-Human Alignment:

- Of 5,520 images:
  - 61.4% are well-aligned between PNASNet and humans.
  - 34.75% are easier for humans.
  - 3.88% are easier for models, lowering the behavioral score.

ANN Models Fall Short of Benchmarks:

- V4: Best ANN score: .663 (ceiling: .892).
- IT: Best ANN score: .604 (ceiling: .817).
- Behavioral: Best ANN score: .378 (ceiling: .497).
- Current models fail to reach the benchmark ceilings.

RESULT

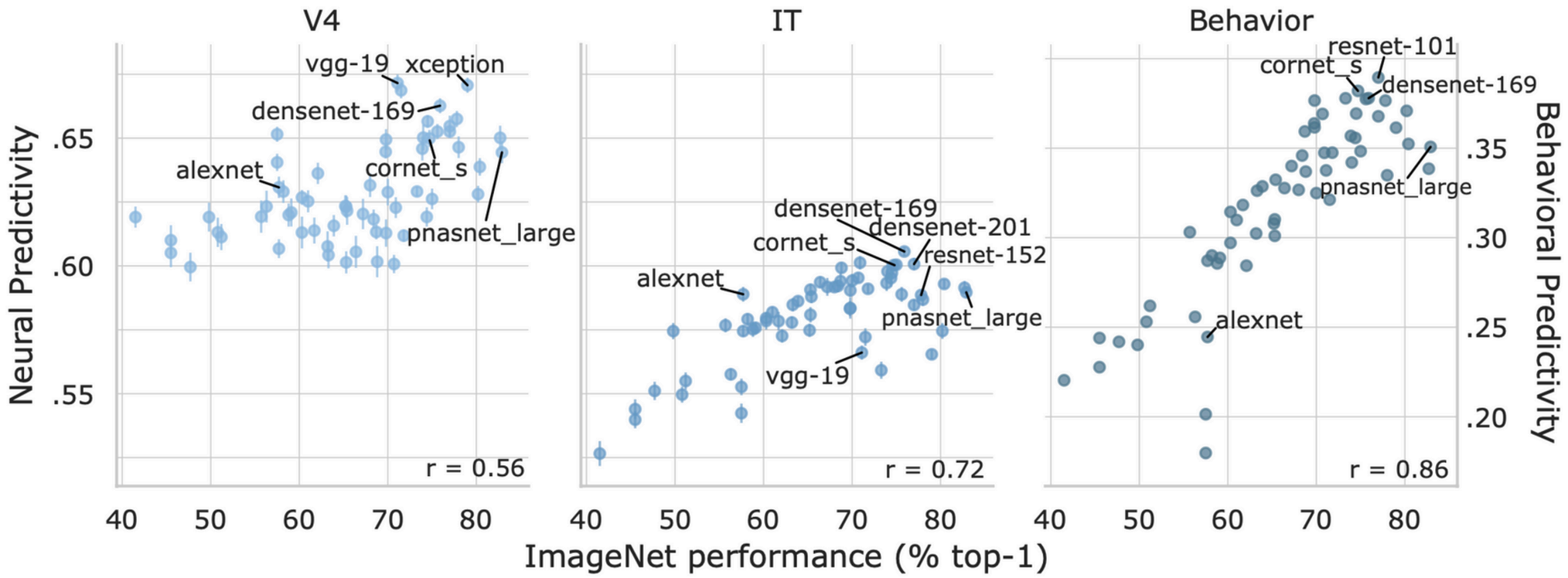
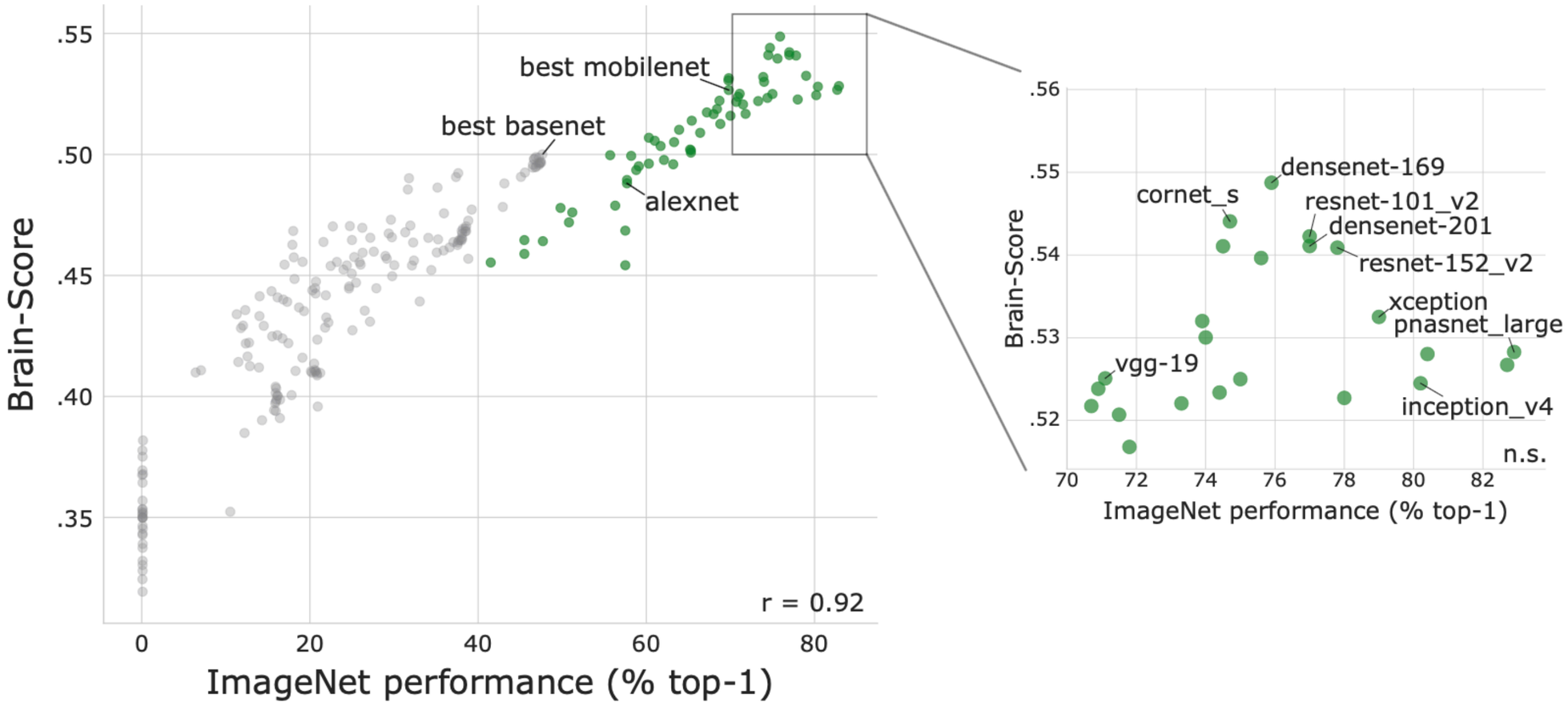


Figure 2: **Predictivities of all models on neural and behavioral benchmarks.** We evaluated regions V4 and IT using the *neural predictivity* as well as behavioral recordings using *l2n*. Current best models are: VGG-19 on V4, DenseNet-169 on IT and ResNet-101 on behavior. Notably, DenseNet-169, CORnet-S 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.

RESULT


.55
.50
.45
.40
.35
Brain-Score
best mobilenet
best basenet
alexnet
r = 0.92
0
20
40
60
80
ImageNet performance (% top-1)
.56
.55
.54
.53
.52
Brain-Score
densenet-169
cornet_s
resnet-101_v2
densenet-201
resnet-152_v2
xception
pnasnet_large
vgg-19
inception_v4
n.s.
70
72
74
76
78
80
82
ImageNet performance (% top-1)

WAYS TO IMPROVE BRAIN-SCORE METRICS

- **Adding more neural recording sites** (even using the same image set) would provide more independent data samples, which helps prevent models from overfitting to the specific benchmarks currently in use.

- **Collecting data from more individual subjects** (more monkeys or humans) would allow for better estimation of between-participant variability. This improved understanding of variability between subjects would help establish a more accurate "noise ceiling" - the theoretical upper limit of how well any model could possibly perform.

WAYS TO IMPROVE BRAIN-SCORE METRICS

- **More diverse images** (photographs, distorted images, artistic renderings), and **data from additional brain regions** (retina, LGN, V1, V2 (earlier regions in visual processing), PFC (higher-level regions outside the ventral stream)).
- Including fMRI, LFP, ECoG, EEG/MEG data alongside current electrode recordings, **fMRI recordings are much more common in humans**, 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.
- **Extending benchmarks to other cognitive tasks** and domains for a more holistic brain model evaluation.

WAYS TO IMPROVE BRAIN-SCORE METRICS

- Need to develop **new ways to compute model-brain similarity** beyond current neural predictivity measures, potentially including temporal dynamics or causal manipulations.
- Current benchmarks mix macaque neural and human behavioral data; **future versions should evaluate models separately for each species.**

FUTURE WORK

- Brain-Score shows which models are better but not why; cannot yet use scores to train models directly.
- Community Platform - Brain-Score.org provides tools for researchers to evaluate models against brain data and compare results, with open-source code and standardized benchmarks.

Thank You