Simulating Brain Regions in CNNs

Incorporating neurophysiological constraints into models

1. Brain Score (2020)

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]

Summary

- Are Deep Networks becoming more or less brain-like with increased task performance?
- Benchmarks:
  - Neural metrics - assess similarity of image-evoked feat activations in ANNs and corresponding neural activations in different primate brain regions
  - Behavioural metrics - assess similarity of the outputs of ANNs and primates (predictions and match-to-sample tasks)

Neural Predictivity

- How well the internal representations of a source system (neural network) match the internal representations of target system (primate brain)
- Source neuroids are mapped to each target neuroid using a linear transformation

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

- Neural Predictivity scores were obtained by comparing predicted responses against held-out, measured neuroid responses and computing Pearson coefficient

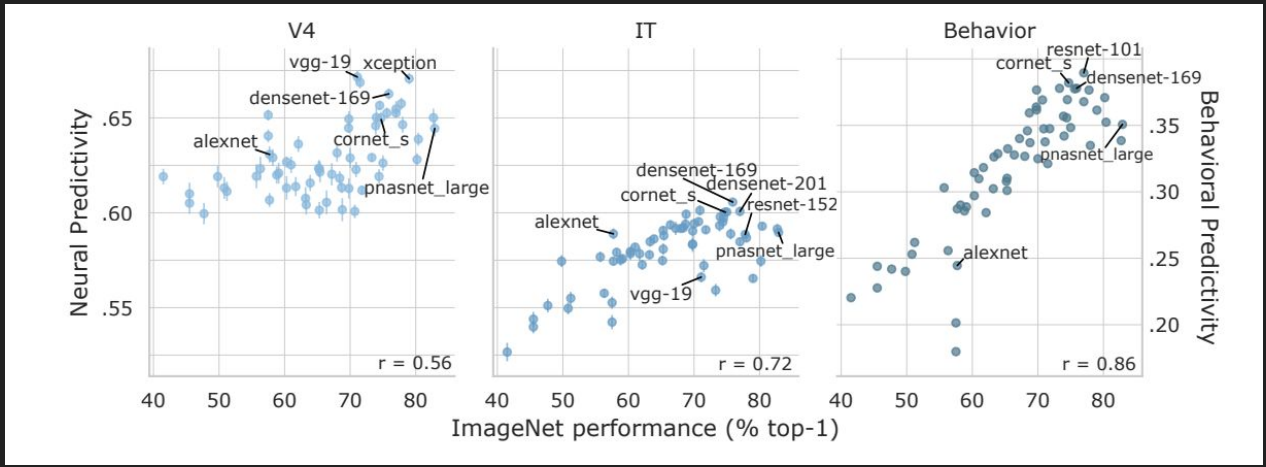
$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(y'_i - \overline{y}')}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2(y'_i - \overline{y}')^2}}$$

Behavioural Benchmark

- How similarly do target and source systems differ in erroring?
- Image-by-image patterns of difficulty broken down by the object choice alternatives (I2n) is the behavioural metric chosen

Score computation

- Final Brain Score: mean of Neural V4 predictivity, Neural IT predictivity, behavioural I2n predictivity
- Mapping model layers involved computing V4 and IT neural predictivity scores for each ‘block’ and the best block being compared with primate counterpart



2. CORnet-S (2019)

Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs

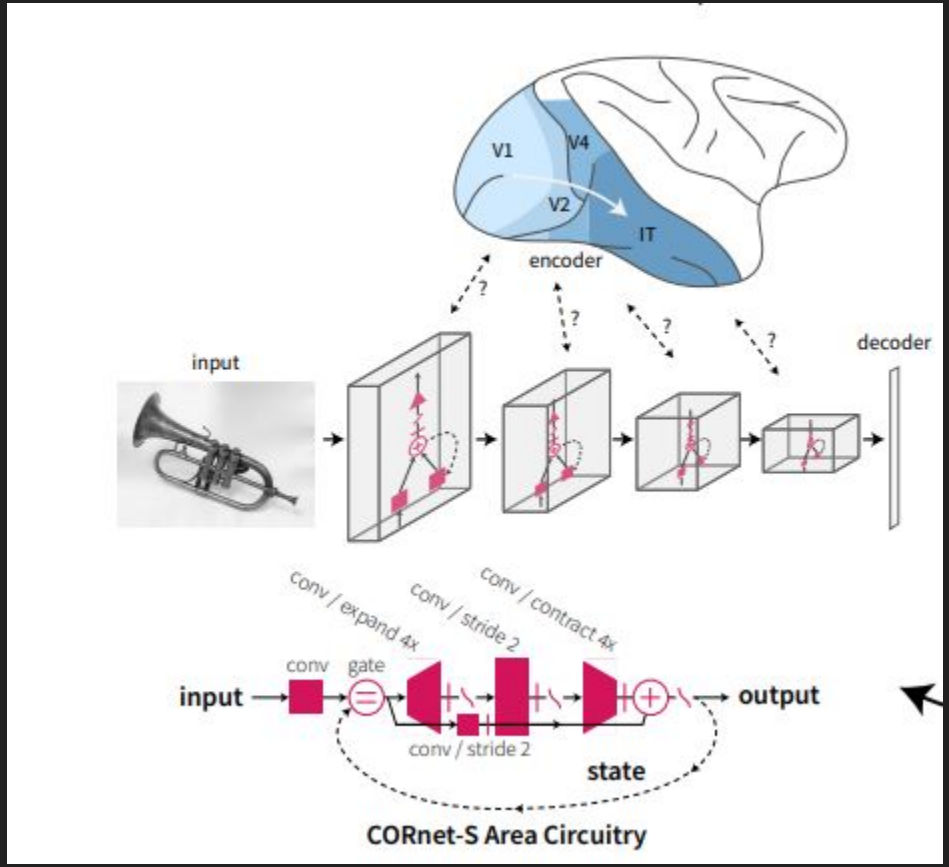
Jonas Kubilius[*,1,2], Martin Schrimpf[*,1,3,4], Kohitij Kar[1,3,4], Rishi Rajalingham[1], Ha Hong[5], Najib J. Majaj[6], Elias B. Issa[7], Pouya Bashivan[1,3], Jonathan Prescott-Roy[1], Kailyn Schmidt[1], Aran Nayebi[8], Daniel Bear[9], Daniel L. K. Yamins[9,10], and James J. DiCarlo[1,3,4]

Motivation

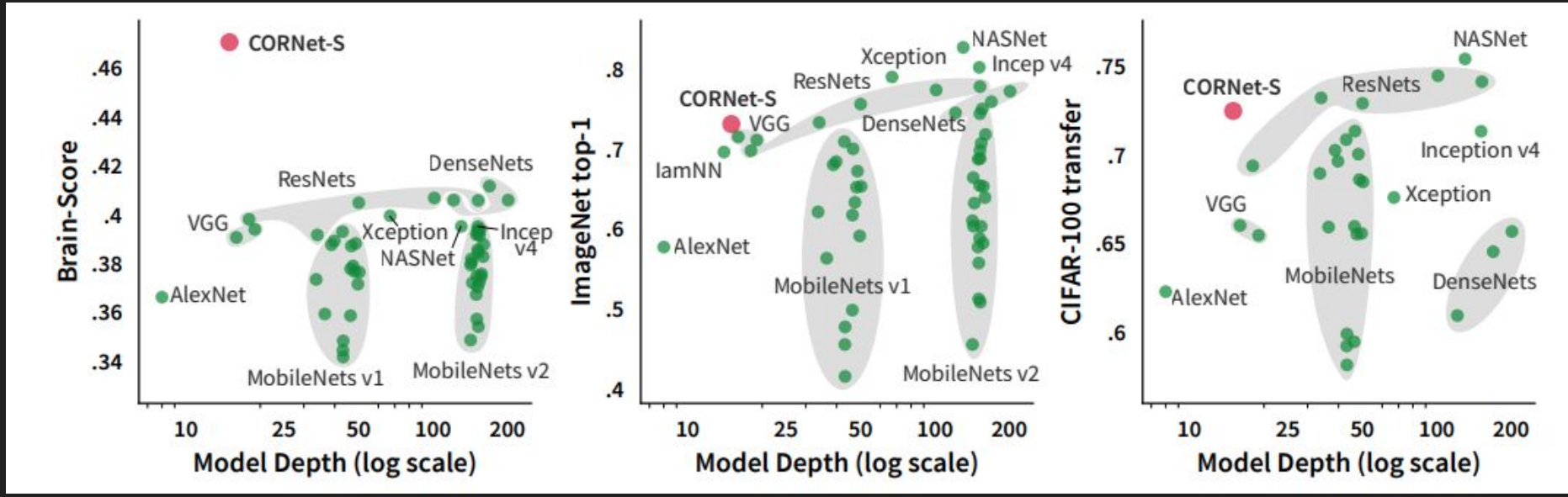
- ANNs were able to *partly* account for how neurons in intermediate layers in primate visual system will respond to an image
- Strong models of the brain opens up non-invasive interfaces to investigate the primate visual system
- Can these models capture brain process even more stringently? Not as a byproduct?
- Proposal: Aligning ANNs to neuroanatomy might lead to more compact, interpretable and functionally brain-like ANNs.

CORnet-S

- A shallow, recurrent anatomical structure of the ventral visual stream
- Compact but also retains a strong 71.3% on ImageNet top-1
- Design Criteria:
  - Predictivity
  - Compactness
  - Recurrence
- Four computational areas analogous to V1, V2, V4 and IT and a linear category decoder mapping last areas' neurons to behaviour choice



https://github.com/dicarlolab/cornet


Brain-Score
.46
.44
.42
.4
.38
.36
.34
CORNet-S
DenseNets
ResNets
VGG
Xception
NASNet
Incep v4
AlexNet
MobileNets v1
MobileNets v2
10
25
50
100
200
Model Depth (log scale)
ImageNet top-1
.8
.7
.6
.5
.4
NASNet
Xception
Incep v4
ResNets
CORNet-S
VGG
DenseNets
IamNN
AlexNet
MobileNets v1
MobileNets v2
10
25
50
100
200
Model Depth (log scale)
CIFAR-100 transfer
.75
.7
.65
.6
NASNet
CORNet-S
ResNets
Inception v4
VGG
Xception
MobileNets
DenseNets
AlexNet
10
25
50
100
200
Model Depth (log scale)

3. Simulating Visual System at the front of CNNs (2020)

Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations

Joel Dapello[*,1,2,3], **Tiago Marques**[*,1,2,4]
Martin Schrimpf[1,2,4], **Franziska Geiger**[2,5,6,7], **David D. Cox**[8,3], **James J. DiCarlo**[1,2,4]

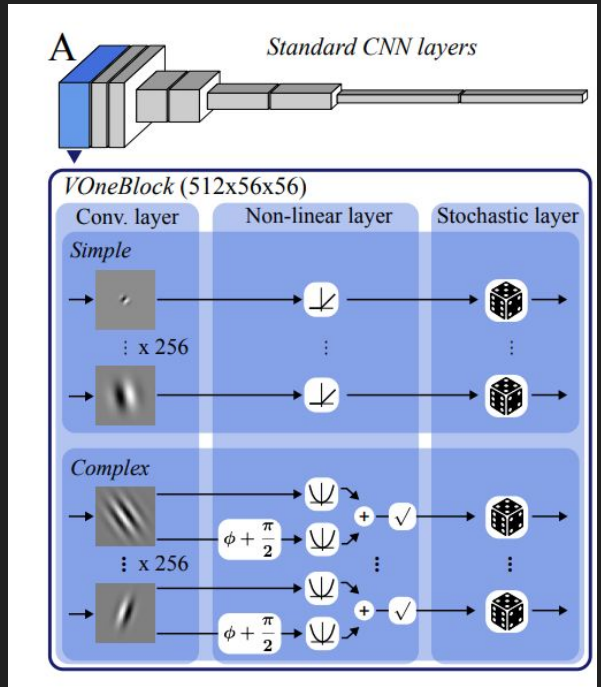
Motivation

- How can we develop CNN's that robustly generalise like human vision?
- Is there a neurobiological prior that can improve CNN robustness to adversarial attacks and image corruptions?
- Correlation - the more *biological* a CNN’s "V1" is, the more adversarially robust it is
- VOneNet - biologically constrained neural network that simulates V1 as the front-end

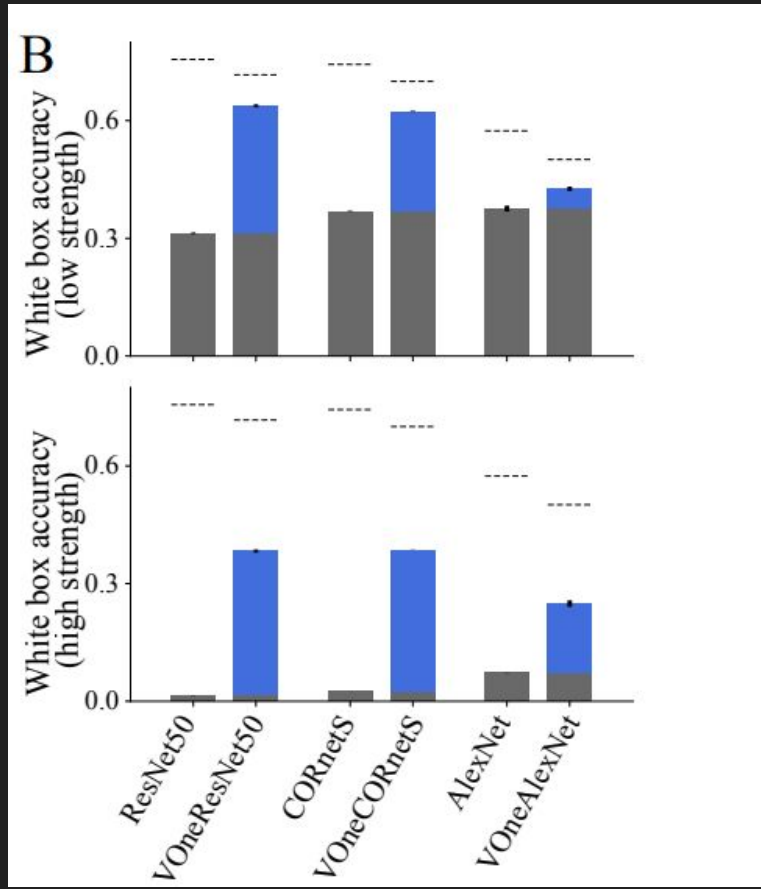
VOneBlock - The Biologically Constrained Front End

- a fixed-weight, mathematically parameterized CNN model that approximates primate neural processing of images up to and including area V1
- VOneBlock components are mapped to specific neuronal populations in V1

- Conv layer - Gabor Filter Bank tuned to approximate empirical primate V1 data
- Non-linear layer (simple or complex cell non-linearities)
- Stochastic layer (V1-Poisson stochasticity generator)



Model	Overall Mean [%]	Perturbation Mean [%]	Perturbation White box [%]	Perturbation Corruption [%]	Clean [%]
Base ResNet50 [4]	$43.6_{\pm 0.0}$	$27.6_{\pm 0.1}$	$16.4_{\pm 0.1}$	$38.8_{\pm 0.3}$	$\mathbf{75.6}_{\pm 0.1}$
AT_{L_∞} [76]	49.0	42.3	**52.3**	32.3	62.4
$ANT^{3\times 3}$+SIN [59]	48.0	34.9	17.3	**52.6**	74.1
VOneResNet50	$\mathbf{54.3}_{\pm 0.1}$	$\mathbf{45.6}_{\pm 0.2}$	$51.1_{\pm 0.4}$	$40.0_{\pm 0.3}$	$71.7_{\pm 0.1}$



Implications

- Forcing biological constraints onto the network leads to similar primate behaviour - robustness to attacks
- The VOneBlock drives downstream layers to learn representations more robust to attacks
- VOneNets improve robustness by being able to generalise to all forms of adversarial attacks and image corruptions without additional training overhead
- Less training to achieve more human-like behaviour

4. Aligning IT and Model Representations (2023)

Published as a conference paper at ICLR 2023

ALIGNING MODEL AND MACAQUE INFERIOR TEMPORAL CORTEX REPRESENTATIONS IMPROVES MODEL-TO-HUMAN BEHAVIORAL ALIGNMENT AND ADVERSARIAL ROBUSTNESS

Joel Dapello[*,1,2,3], **Kohitij Kar**[*,1,2,4,6],
Martin Schrimpf[1,2,4], **Robert Geary**[1,2,3], **Michael Ferguson**[1,2,4] **David D. Cox**[5], **James J. DiCarlo**[1,2,4]

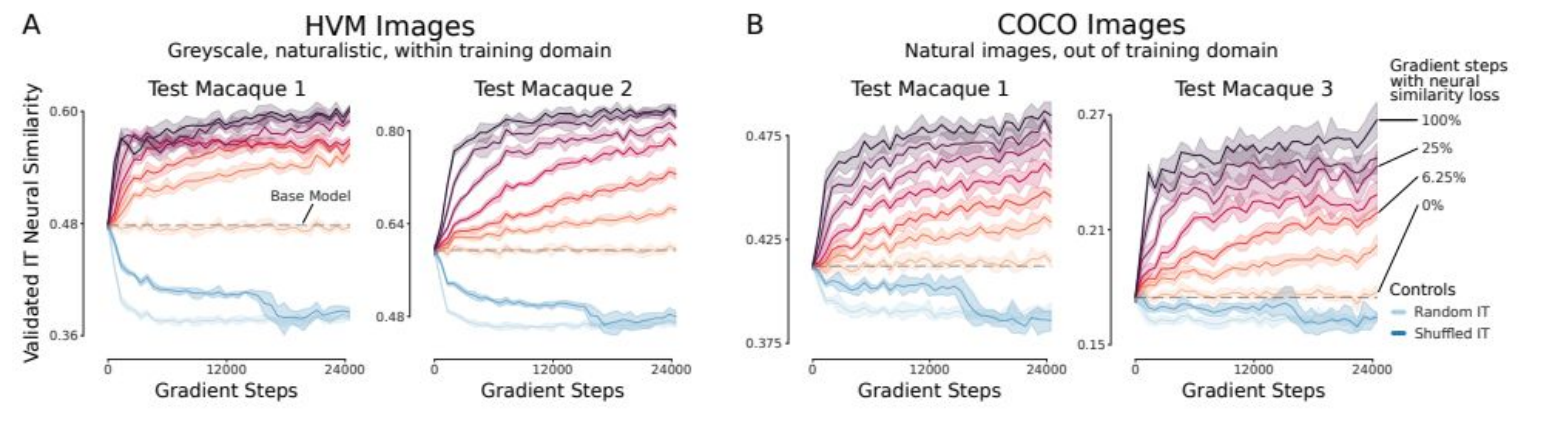
Motivation

- Despite similarities and resemblance to the primate visual system, output behaviour of CNNs do not fully match that of a primate in same tasks
- Adversarial attacks
- Instead of directly simulating a brain region, can forcing CNN to have similar representations be a viable approach?
- Focus on IT cortex neural representations
- V1 Paper - Towards robust vision by multi-task learning on monkey visual cortex (diff lab)

Aligning

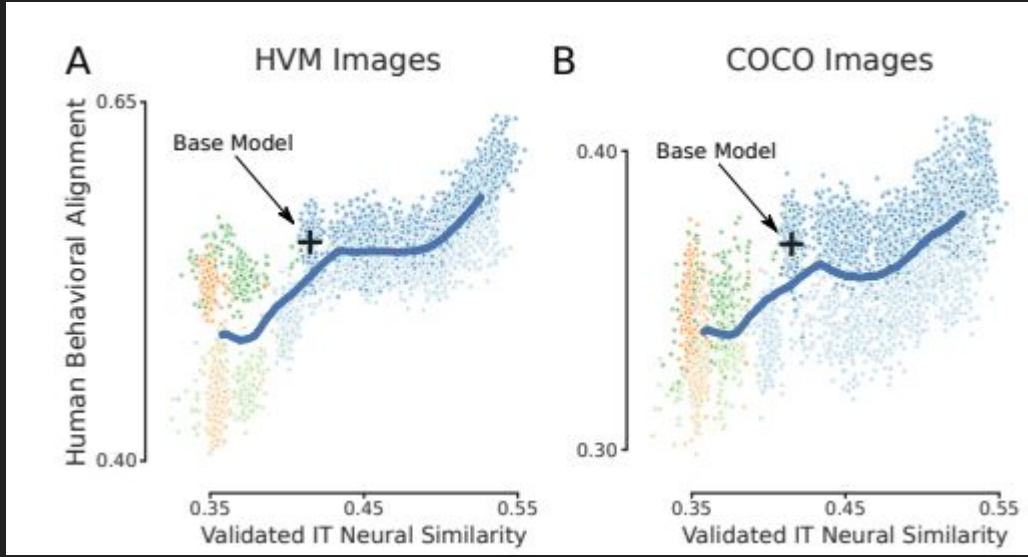
- Developed a method to align late layer “IT” representations of CNN pretrained on ImageNet and HVM images to biological IT representations while model continues to be optimised for classification
- Multi-loss formulation
  - Standard categorical cross entropy for model prediction of ImageNet and HVM labels
  - Centered Kernel Alignment (CKA) based loss penalising ‘IT’ Layer
- CKA loss made sure CNN’s “IT” layer representations aligned with primate IT representations
- CKA measures linear subspace alignment invariant to isotropic scaling and orthonormal transformations
- Behavioural Alignment through BrainScore’s i2n metric

IT-likeness Results



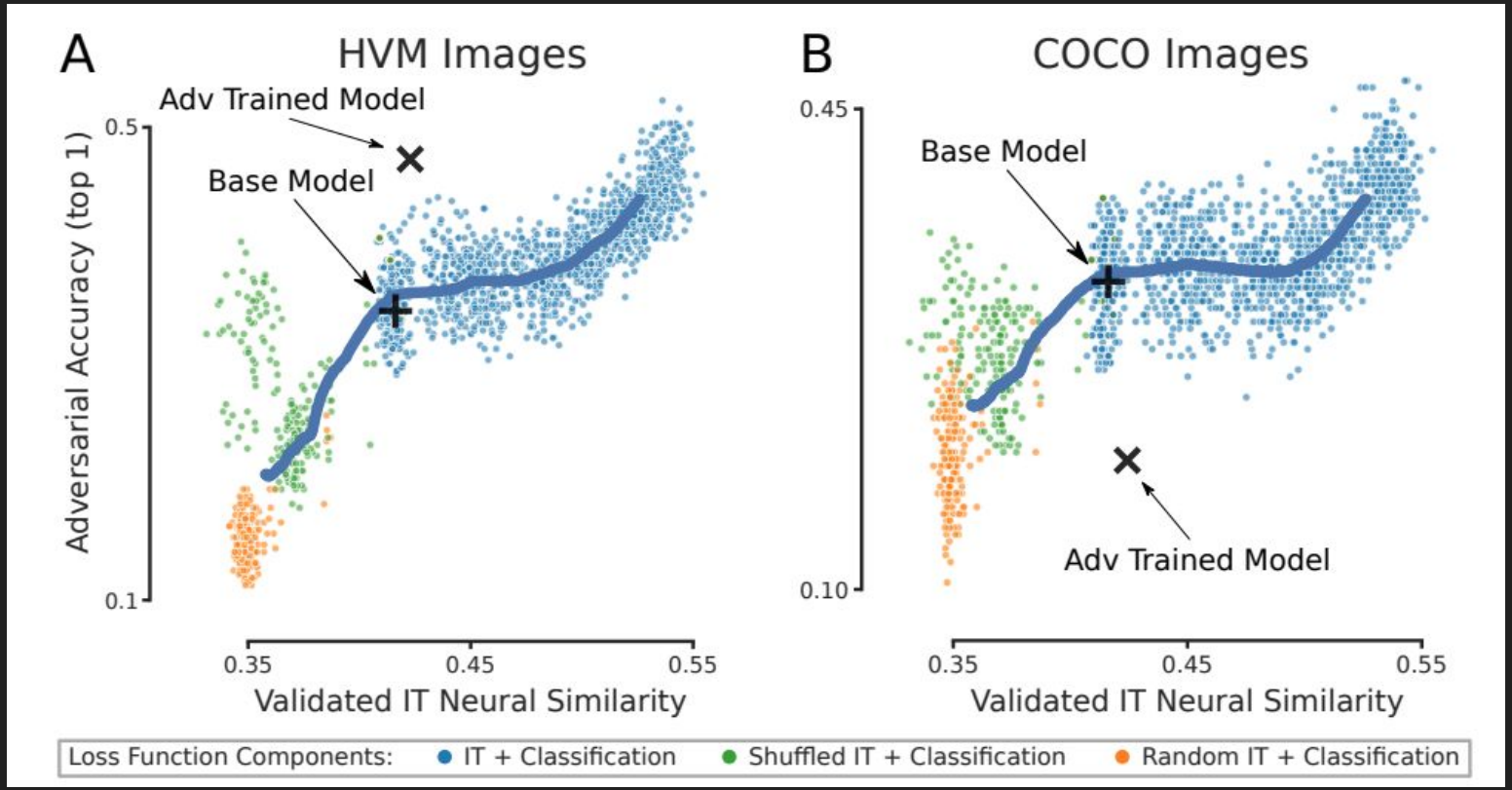
- The more the model is ‘corrected’ with the IT similarity loss, the more IT neural similarity it has
- Sanity check performed on held out Macaques and HVM images

Behavioural Alignment Results



Loss Function Components: IT IT + Classification Shuffled IT Shuffled IT + Classification Random IT Random IT + Classification

Adversarial Robustness Results



Implications

- It is possible to align the late stage ‘IT’ representations
- Representationally-aligned models also have better human behavioural alignment
- The aligned models are more robust to adversarial tasks even on unseen images
- Population geometry and not individual neuronal sensitivity might be playing a critical role in the robustness

Harmonizing the object recognition strategies of deep neural networks with humans

Thomas Fel[*1,2], Ivan Felipe[*1], Drew Linsley[*1,3], Thomas Serre[1,2,3]
{thomas_fel,ivan_felipe_rodriguez,drew_linsley}@brown.edu

Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex

Drew Linsley[*1,2], Iván F. Rodriguez[*1], Thomas Fel[2], Michael Arcaro[4], Saloni Sharma[5], Margaret Livingstone[5], Thomas Serre[1,2,3]
drew_linsley,ivan_felipe_rodriguez@brown.edu

5. The Neural Harmoniser (2022 & 2023)

- Introducing biological constraints could hurt performance of DNNs
- Neural Harmoniser: A way to align an DNN’s strategies with humans without hurting performance
- Behavioural alignment instead of neural alignment
- Introduces a loss that enforces alignment across feature importance maps

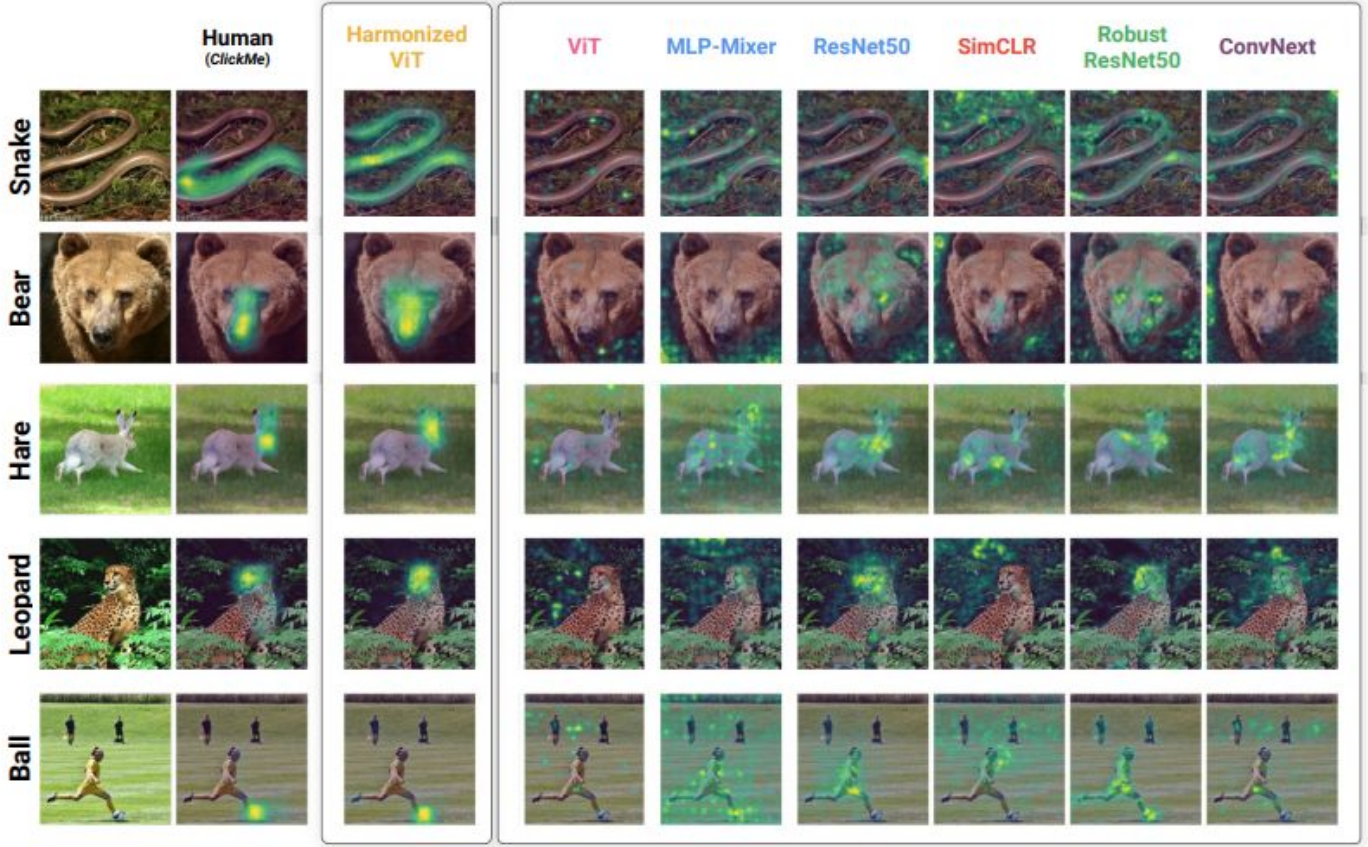
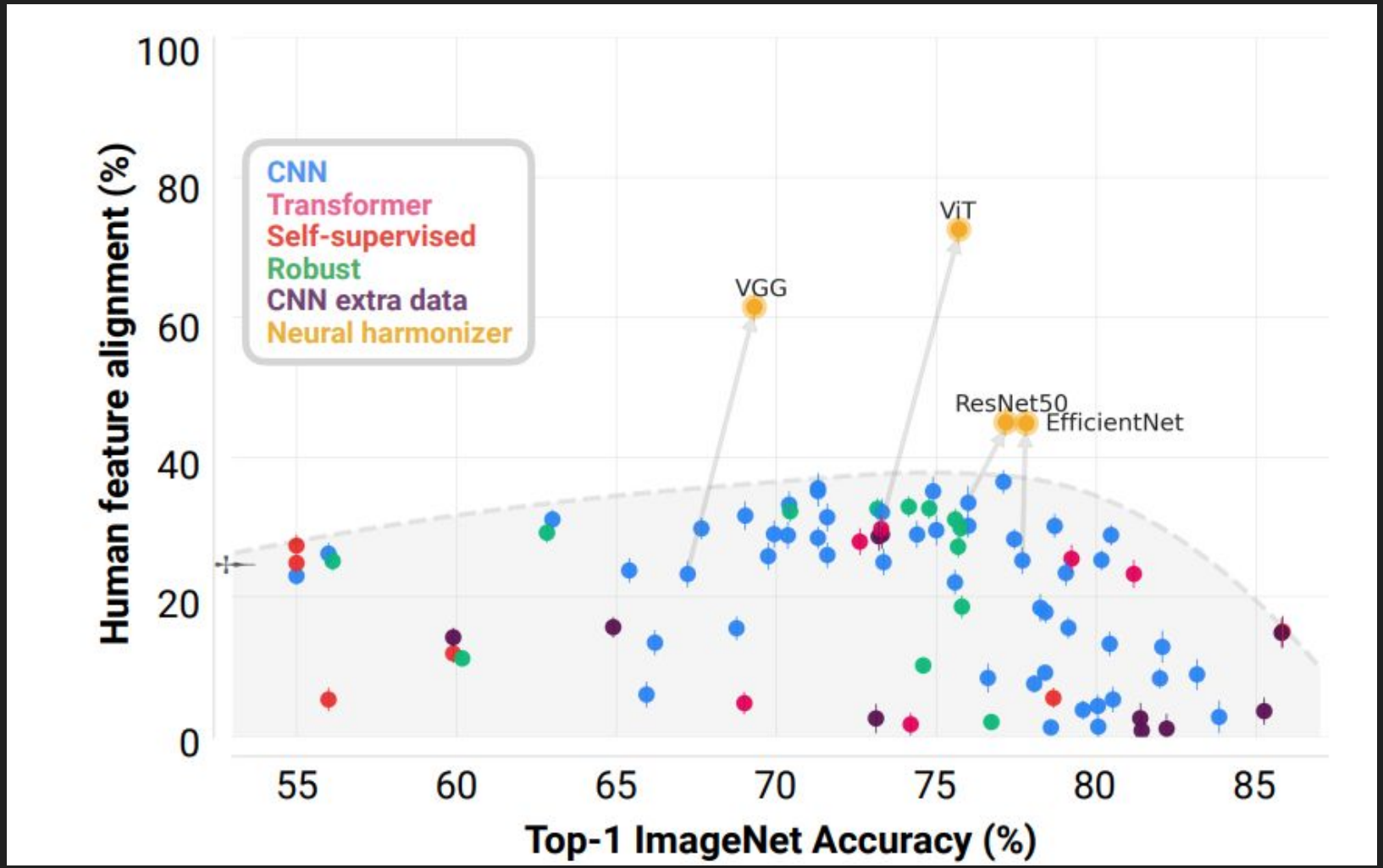
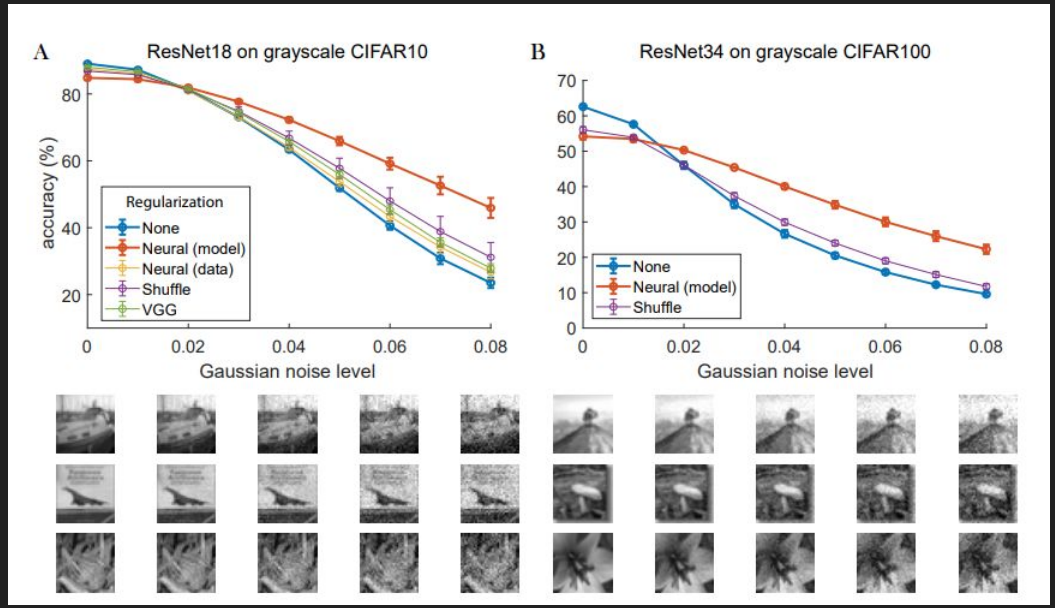


Figure 2: **Human and DNNs rely on different features to recognize objects.** In contrast, our neural harmonizer aligns DNN feature importance with humans. We smooth feature importance maps from humans (*ClickMe*) and DNNs with a Gaussian kernel for visualization.


CNN
Transformer
Self-supervised
Robust
CNN extra data
Neural harmonizer
VGG
ViT
ResNet50
EfficientNet
Human feature alignment (%)
100
80
60
40
20
0
55
60
65
70
75
80
85
Top-1 ImageNet Accuracy (%)

6. Learning From Brains How to Regularize Machines (2019)



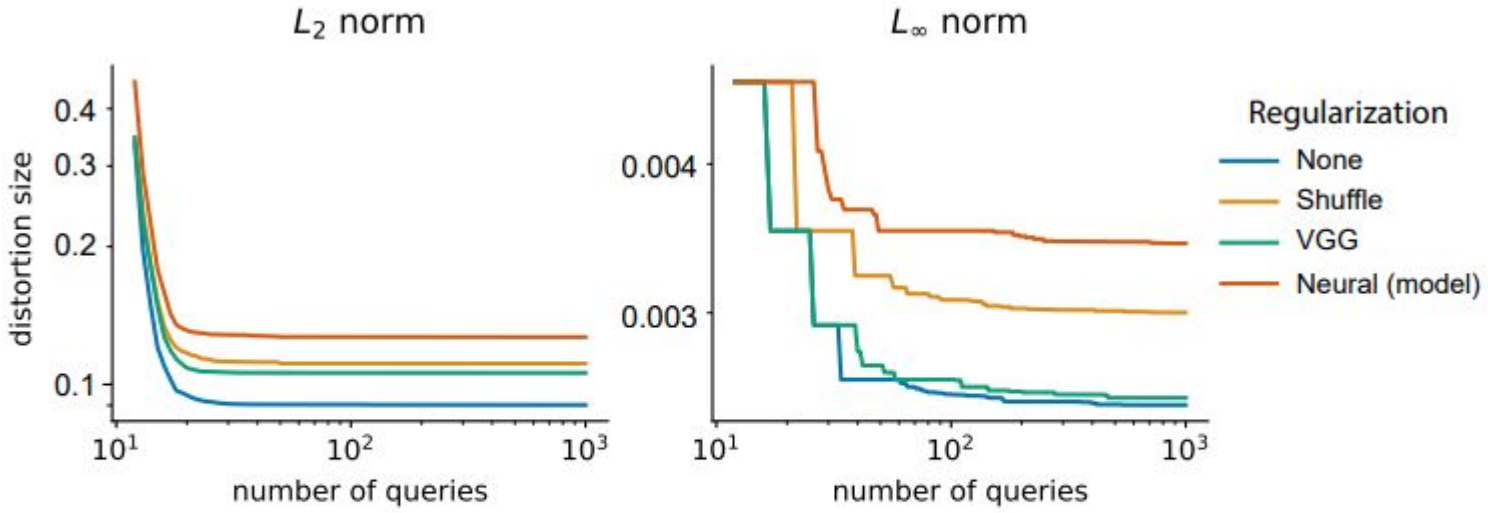


Figure 5: Adversarial robustness of classifier networks according to the L_2 and L_∞ norms. With more optimization queries for the attack, the minimal perturbation shrinks. Regularization improves adversarial robustness, with neural regularization providing the best defense throughout the attacker's optimization. 'Neural (data)' regularizer is not tested.

7. Adversarially-trained artificial neurons are more robust than biological neurons (2022)

- They find that representations of adversarially trained ANNs have exceeded that of corresponding biological neural representations in the single unit level in terms of robustness

This result confronts us with an apparent paradox: How is it that primate visual perception seems so robust yet its fundamental units of computation are far more sensitive than expected? One distinct possibility is that visual object recognition behavior in primate is actually not robust. This could be potentially explored with an iterative adversarial psychophysics experiment, similar to what we have done here for IT neurons. An alternative explanation is that there is an unknown error-correction mechanism at the population level in IT or in a down-stream area that decodes object identity. These hypotheses can be tested in subsequent experiments. We believe the current line of work could potentially lead to biologically-inspired solutions in ML robustness research, provide fundamental insights into the nature of adversarial phenomena in biological cognition, and perhaps provide new avenues to precisely modulate internal brain states without disrupting daily visual behavior.

8. Biologically Inspired Mechanisms for Adversarial Robustness (2020)

- These two features of primate vision help improve adversarial robustness:
  - Foveation due to non-uniform distribution of cones in the retina
  - Multiscale filtering because of receptive fields of different sizes in V1 at each eccentricity