FEATURE SEGREGATION BY SIGNED WEIGHTS IN NEURAL NETWORKS FOR ARTIFICIAL AND BIOLOGICAL VISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Signed connections are central to both artificial and biological intelligence, positive and negative weights in artificial networks, and excitatory and inhibitory synapses in the brain, yet their representational role remains unclear. Here, we investigate how signed weights shape visual representations in artificial and biological systems involved in object recognition. Using sign consistency as a proxy for biological Dale's law, which requires neurons to send either exclusively excitatory or inhibitory outputs, we found that accuracy of ImageNet trained networks positively correlated with the Dale index of their output layer. Ablation and feature visualization reveal a functional segregation: removing positive inputs disrupts object related, low frequency structure, while removing negative inputs mainly alters background textures. This segregation is more pronounced in adversarially robust models, persists with unsupervised learning, and vanishes with non-rectified activations. In intermediate layers, the most positive Dale-like channels encoded localized, object-like features, whereas the most negative ones captured dispersed, background features. We next performed *in vivo* feature visualization in monkey ventral visual cortex (V1, V4, and IT) and fitted linear models using the input layer to the neural networks classification units. These models reproduced features similar to those preferred by the biological neurons. In the model neurons, removing positive inputs altered representations more than removing negative ones. The most Dale-like positively projecting units exhibited localized features, while the negatively projecting units showed larger, more dispersed features, suited to carrying contextual input. Consistent with this, clearing the background around each neuron's preferred feature enhanced its response, likely by reducing inhibitory drive, supporting inhibition as a contextual modulation of the excitatory feature. Our results demonstrate that both artificial and biological vision systems segregate features by weight sign: positive weights emphasize objects and low frequencies, negative weights encode context. This shows the convergence of representational strategies in brains and machines, yielding predictions for visual neuroscience.

1 INTRODUCTION

Brains and artificial neural networks (ANNs) both rely on signed connections. In biological circuits, Dale's law states that neurons are either excitatory or inhibitory (Dale, 1935). Excitatory neurons are thought to compute core visual features, while inhibitory neurons sharpen selectivity, modulate context, and gate information flow (Isaacson & Scanziani, 2011). The primate ventral visual stream supports object recognition (DiCarlo et al., 2012), and representations from V1 to IT increase in complexity, paralleling hierarchies in convolutional neural networks (CNNs) (Yamins et al., 2014; Güçlü & van Gerven, 2017). While early mechanisms such as lateral inhibition and center-surround receptive fields are well charachterized, the excitatory–inhibitory organization of higher ventral-stream areas remains less understood (Tamura et al., 2004), motivating the question of whether information is segregated by connection sign in both artificial and biological vision.

ANNs also use positive and negative weights, loosely analogous to excitatory and inhibitory signals but unconstrained by Dale's law. Yet it remains unclear how deep networks divide visual information across signed weights, particularly in object classification models where each output unit represents

a category. Prior work suggests segregation by absolute weight strength (Li et al., 2023), , but whether feature types such as foreground objects versus background context systematically separate by weight sign is unknown. Here, we hypothesize that CNNs segregate visual information into positive and negative inputs.

We test this hypothesis across diverse ImageNet-trained CNNs by analyzing weight sign organization, ablating positive and negative inputs, and visualizing the resulting feature selectivity. We asses feature segregation across network depth by analyzing Dale-like channels. We examine a range of architectures and training regimes, including adversarially robust models, unsupervised models, and networks with nonrectified (Tanh) activations.

For biological comparison, we fit linear models from ANN features to neural responses recorded across the macaque ventral stream (V1, V4, IT) and use *in vivo* feature visualization and background manipulations to probe inhibitory contributions. These analyses relate ANN-derived sign structure to biological feature preferences.

Our results support an emerging principle: across artificial and biological vision systems, connection sign organizes feature representation. Positive weights emphasize object-related, localized features, whereas negative weights encode contextual, more dispersed structure. This connects classic ideas rooted in Dale's law with representational strategies in modern ANNs, generating mechanistic and testable predictions for visual neuroscience.

2 Related work

Mechanistic interpretability of computer and biological vision There has been progress in mechanistic interpretability in ANNs from work using perspectives adapted from neuroscience circuit dissection (Olah et al., 2020). This line of explainable AI research explains model behavior by dissecting smaller network subgraphs, revealing how relevant features arise from input weights and are composed hierarchically. Such work has uncovered motifs involving positive and negative connections between related features, reminiscent of early visual system organization. New approaches to address representations beyond single units rely on sparse dictionary learning, with early work in vision (Olshausen & Field, 1996), an approach that has recently regained popularity in language modeling (Cunningham et al., 2023), as well as in multimodal models (Pach et al., 2025). Some studies also characterized object shape and texture biases in feature visualizations by reconstructing images from sparse weight sets (Li et al., 2023). However, the systematic division between positive and negative inputs across the entire range of weight strengths, and its possible role in feature segregation, remains underexplored and is a focus of this study.

Feature visualization by closed-loop optimization Characterizing learned representations is foundational for both biological and artificial vision research. Feature visualization, i.e. optimizing for images that strongly activate target units, was originally pioneered in the brain by hand (Hubel & Wiesel, 1959), and later in silico by gradient ascent on pixels of neural networks (Erhan et al., 2009; Nguyen et al., 2016a;b; Olah et al., 2017). Because gradients are unavailable when recording in vivo, gradient-free black-box optimization techniques were developed for synthesizing preferred images of biological neurons in real-time (Ponce et al., 2019; Xiao & Kreiman, 2020; Wang & Ponce, 2022). These approaches constrain the search space via generative adversarial networks, promoting naturalistic solutions (Nguyen et al., 2016a). Further methods involve first fitting a predictive network to neural data and then using in silico gradient ascent (Bashivan et al., 2019; Walker et al., 2019). While most prior studies use grayscale images, our study applies gradient-free visualization to color images in both CNNs and primate recordings.

Robustness Neural networks are susceptible to adversarial attacks, where noise that is nearly imperceptible by humans can be added to natural images, changing output classification (Szegedy et al., 2014; Salman et al., 2020; Elsayed et al., 2018). Robust training, i.e. introducing adversarial perturbations during learning, improves resistance to such attacks and is hypothesized to align learned representations more closely with primate visual processing. Prior work does not assess how robustness impacts the organization of image representations after ablation of signed weights, which we systematically investigate here.

Nonlinearity influence on representations Beyond training objectives, the role of activation functions such as ReLU versus Tanh profoundly influences representational properties (Alleman et al.,

2023), with ReLU inducing representations better aligned with input features and Tanh inducing alignment with output features (labels). This prior work was done in small networks from a theoretical perspective; thus, the impact of rectification on the potential segregation of visual features at practical scales is unknown and addressed by our study.

3 METHODS

An extended methods section is in the Appendix A.1.

Networks We performed our ablation studies in CNNs pretrained on the ImageNet dataset: AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan & Zisserman, 2015), ResNet50 (He et al., 2015), and robust ResNet50 models with robustness radii specified by the L_∞ norm ($L_\infty \in \{0.5, 1, 2, 4, 8\}$; Salman et al. 2020). Here, L_∞ denotes the maximum-norm constraint used during adversarial training, i.e., $\|\delta\|_\infty \leq \epsilon$, which bounds the maximum per-pixel perturbation by ϵ. To reduce computing time, we used the *imagenette* dataset (noa, 2024) and the *ImageNet* macaque category. We also tested 100 classes sampled by k-means on the output of ResNet50 (Figs. 12,13) For all networks, we visualized the representations of the units in the fully-connected output layer (pre-softmax) matching those classes under different ablation conditions.

Dale index We quantified Dale like structure with a Dale index for each outgoing channel per layer, defined as the fraction of its weights that share the majority sign, $D = \max(p_+, p_-)$, where p_+ and p_- are the proportions of positive and negative outgoing weights. The index ranges from 0.5 to 1 and measures sign consistency.

Ablation For each layer, we ablated positive and negative weights separately. Given a layer's weight matrix W, we defined the sets of positive weights $P = \{w \in W : w > 0\}$ and negative weights $N = \{w \in W : w < 0\}$. For each set $S \in \{P, N\}$, we sorted its elements in decreasing order of absolute value. We then defined the ablation strength $\alpha \in [0, 1]$ as the fraction of the total magnitude of S to remove. Specifically, we identified the smallest k satisfying $\sum_{i=1}^{k} |w_i| \;/\; \sum_{w \in S} |w| \geq \alpha$, where $w_1, w_2, \ldots$ are the sorted weights in S, and set those k weights to zero. Because α is a normalized cumulative magnitude, it lies in $[0, 1]$, and sweeping α from 0 to 1 removes none to all of the positive (or negative) weights.

Feature visualization For each ablation condition, we performed feature visualization by optimizing a GAN latent code to create an activity-maximizing image. We used this closed-loop, zeroth-order-search approach to allow comparison with our neuronal experiments, where gradient ascent would not be possible. To increase the span of the stimulus space, we used two GANs: AlexNet fc6 DeePSiM (Dosovitskiy & Brox, 2016) which can render textures and objects, and BigGAN (Brock et al., 2019) that can render photo-realistic images with objects. For optimization, we used a variant of *covariance matrix adaptation evolutionary strategy* or CMAES (Wang & Ponce, 2022; Loshchilov, 2015). We optimized ten images per GAN, resulting in 20 feature visualizations per output unit and ablation condition. Diverse visualizations better capture the multifaceted high-level representations in CNNs (Nguyen et al., 2016b). For our examples, we show the best of the 20 visualizations, but used all for quantitative analyses. For visualizations of neural networks predicting biological neuron responses, due to experimental time restrictions, we used five visualizations per ablation condition, via DeePSim only. Our experiments are performed in a PC with Nvidia 4090 GPU, and each visualization takes about 3 mins.

Network training Both ResNet18 networks were trained using the FFCV library (Leclerc et al., 2023) for 16 epochs on the ImageNet1K dataset. The top-5 classification accuracy was 0.797 for the network with Tanh activations and 0.870 for the network with ReLU activations. Note that these models were trained for only 16 epochs rather than the standard 90, so their accuracy underperforms published benchmarks. However, they are suitable for our mechanistic analyses.

Visual cortex electrophysiology We recorded multi-unit (neuron microcluster) and occasional single-unit activity from chronically implanted multielectrode arrays in V1, V4, and PIT of two macaques. Animals fixated while 2–8° images were flashed briefly (100 ms ON, 150 ms inter-stimulus interval). Neurons were driven with a 160-image stimulus set (diverseSet), constructed to span a broad range of visual features derived from k-means clustering of AlexNet output layer over the IN1K validation set and typical visual neuroscience image sets. For each session, we modeled responses of a single neuron or multiunit using a one-component PLS regression between firing rates

and AlexNet penultimate-layer activations, and used the resulting model for *in silico* ablations and feature visualizations, validating visualizations *in vivo* within the same session.

4 RESULTS

4.1 PROXIMITY TO DALE'S LAW IN CNN OUTPUTS CORRELATES WITH ACCURACY

A key challenge in comparing artificial and biological circuits is that CNNs are not constrained by Dale's law. Therefore, we asked whether the Dale index (our measure of sign consistency) in the output layer of diverse CNNs has any relation to its performance. Dale index increased from random initialization with training (Fig 1A). Top 1 accuracy on ImageNet1K positively correlated with the mean Dale index of the output layer. Within a given architecture, the Dale index increased with network depth. And batchnorm training in VGGs produced output layers with higher Dale index. Thus, even without an explicit Dale constraint, these networks naturally developed more sign consistent output channels. This motivated us to examine the specific visual features carried by positive versus negative weights.

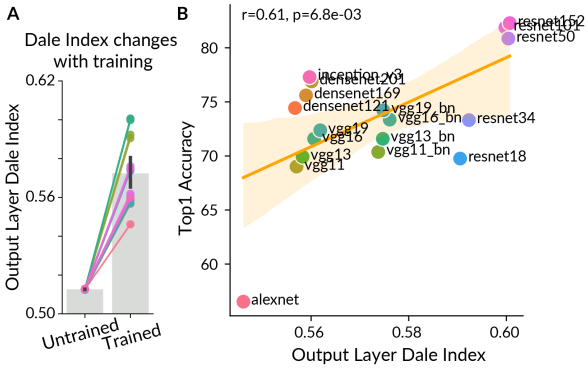

Figure 1: Dale index of output layer correlates with top 1 accuracy on ImageNet1K CNNs. **A.** Training increases Dale index of output layers. **B.** Network top 1 accuracy correlates with Dale index of the output layer.

4.2 OBJECT INFORMATION IS PREFERENTIALLY ENCODED BY POSITIVE WEIGHTS

Hypothesis Motivated by biological center-surround organization, where inhibitory surrounds convey contextual information, we hypothesized that CNN output units trained for object recognition segregate object features to positive weights and contextual features to negative weights.

Testing segregation by ablation and visualization. We examined this hypothesis in ImageNet-pretrained CNNs using ablation and feature visualization. The overall ratio of positive to negative input weights per unit was close to unity (Table 2), suggesting that both polarities may encode relevant information. We then selectively ablated positive or negative input weights to class units and visualized features across ablation strengths. Ablating positive weights greatly reduced the maximal achievable activation during feature visualization, whereas removing negative weights slightly increased it (appendix Fig. 11). Visually, positive-weight ablation disrupted recognizable object structure, while negative-weight ablation largely preserved object identity and instead altered background or color context (Figs. 2B, C). To quantify these effects, we compared image sets generated before and after ablation using mean pairwise cosine similarity across an ensemble of readout CNNs. Positive-weight ablation yielded substantially less similar representations, whereas negative-weight ablation produced only minor shifts (Fig. 2D). These results replicated across 100 ImageNet classes and with alternative metrics such as LPIPS (Zhang et al., 2018) (appendix Fig. 13), demonstrating robustness and generality.

To quantify to what extent objects disappear from the preferred images under ablations, we evaluated objectness using an object-detection network (YOLOv7 Wang et al. (2022)). Relative to baseline objectness scores computed from intact visualizations, ablating positive weights reduced objectness, whereas ablating negative weights had minimal impact (Fig. 2E). Analyzing spatial frequencies revealed positive ablations majorly affected low frequencies (app. Fig. 14,15), consistent with the objectness reduction. Together, these results show that in ImageNet-trained CNNs, removing positive input weights disrupts object features, while removing negative weights primarily alters context, indicating that positive weights preferentially encode object features.

4.3 SEGREGATION DEPENDS ON RELU BUT NOT ON UNSUPERVISED PRETRAINING

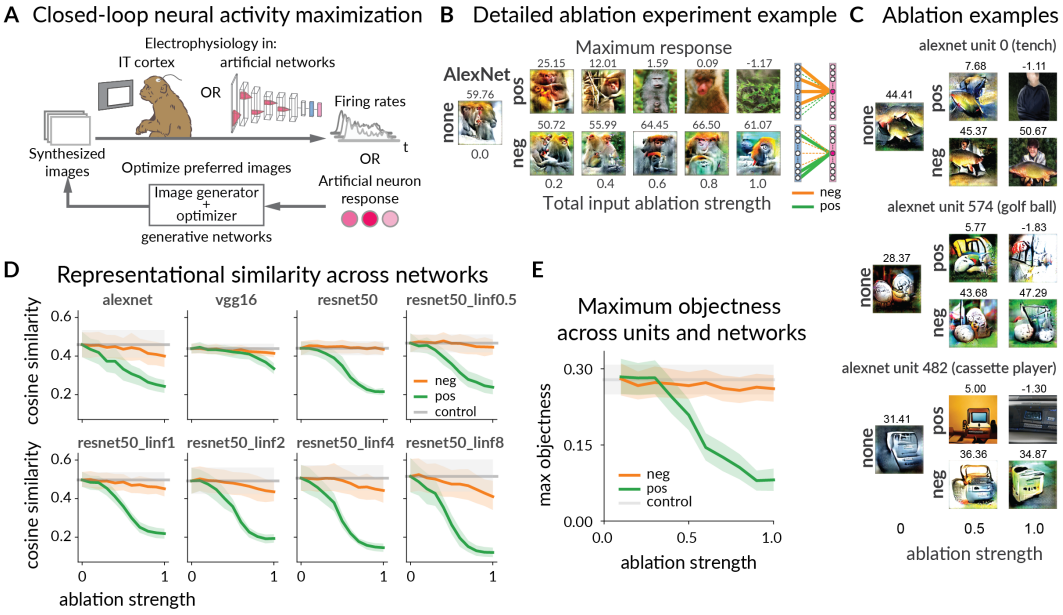

Figure 2: **A.** Schematic of feature visualization workflow in ANNs and brains. **B.** Preferred feature changes for different ablation strengths of input weights to the macaque 373 output unit of AlexNet (last fc layer of 1000 units before softmax). Images are the most activating images out of the 20 visualizations per ablation strength. Ablation strengths are below each image, and activation scores are above. **C.** Changes in preferred features to different ablations of example AlexNet output units: 0 tench, 574 golf ball, and 482 cassette player. Notice the large image changes for positive ablations. Same methods as in **B**. **D.** Representational similarity of intact vs input-ablated units across networks tested, measured by the pairwise cosine similarity of control vs ablation images over an ensemble of networks. Error bars are 95% confidence intervals over units, each unit is the mean of its 20 visualizations. The units correspond to the 10 imagenette categories ([0, 217, 482, 491, 497, 566, 569, 571, 574, 701]) plus the macaque category (373). **E.** Objectness scores across units per ablation condition. As in **D**, we tested 11 units from the 1000-unit fully-connected output layer (pre-softmax) of: AlexNet, VGG16, ResNet50, and robust ResNet50 ($L_\infty \in \{0.5, 1, 2, 4, 8\}$). For each network, we averaged over the objectness scores of 20 visualizations per unit and all units. The plot shows the mean over previously described network averages. Error bars are 95% confidence interval over network averages.

To examine whether functional segregation by weight sign relies on supervised learning, we analyzed a ResNet50 backbone trained without supervision (SimSiam from Chen & He (2020)). After unsupervised pretraining, the backbone was frozen and a linear classifier was trained on ImageNet-1K. Feature ablation and visualization revealed that this network still allocated object features to positive weights, although these features disappeared at lower ablation strengths than in fully supervised CNNs (Fig. 3A, appendix Fig. 18). Negative input ablation had only a minor

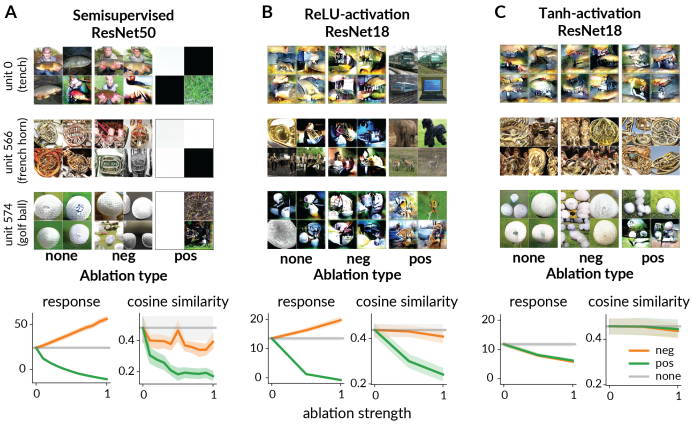

Figure 3: Ablation experiments (rows are features of three example units, and responses, and representational similarity for all 11 classes) in **A.** Semisupervised networks, **B.** Vanilla ReLU supervised network, and **C.** Network with a non-rectified activation function Tanh, replacing all ReLUs in **B**.

effect, suggesting that even unsupervised representations are organized so that positive weights convey most object-related features.

We hypothesized that this segregation depends on the rectifying ReLU activation, which forces unit outputs to be non-negative and makes weight signs determine the direction of each contribution. Prior work in toy networks shows that ReLU promotes alignment to the input space, whereas Tanh promotes alignment to the output space (Alleman et al., 2023). To test the role of rectification, we trained ResNet18 models with either ReLU or Tanh. As expected, the ReLU model showed clear segregation, with the strongest disruptions arising from ablation of positive weights. In contrast, the Tanh model showed similar representational changes across ablation types and preserved key features even when either sign was removed (Fig. 3B,C). Thus, rectified activations are required for strong segregation of object information into positive weights in CNNs.

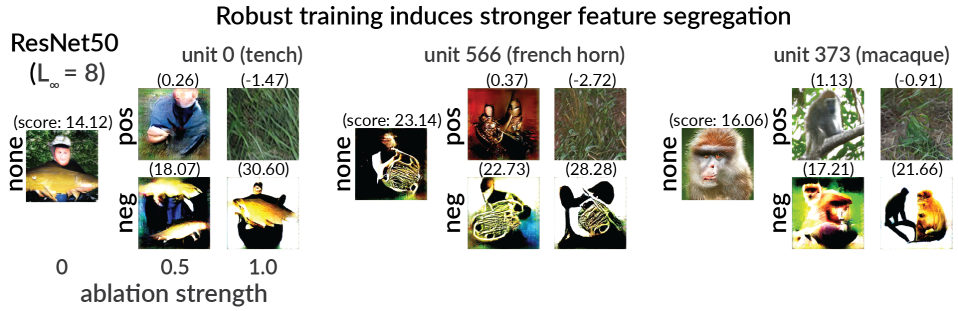

Figure 4: Robust network ResNet50 $L_\infty = 8$ shows a large change in preferred features upon input ablation. Notice the white background in the negative-weight ablation condition.

4.4 Adversarially robust networks show enhanced feature segregation

Having established that both unsupervised pretraining and rectified activations support the segregation of object and context information by weight sign, we next asked how this organization is affected by other salient properties of deep vision networks. In particular, adversarially robust networks, which are trained to resist small targeted image perturbations (Szegedy et al., 2014; Madry et al., 2019), are believed to better reflect aspects of biological vision and may therefore show distinctive patterns of feature segregation. We examined whether and how adversarial robustness influences the allocation of object and contextual information to positive and negative weights.

In robust ResNet50 networks, intact feature visualizations appeared more object-like, and ablation of negative input weights reliably altered the background color, often rendering it white (Fig. 4). This hinted at a stronger feature segregation than in vanilla networks. Quantitative analysis confirmed that as network robustness to adversarial attacks increased, so did the model's vulnerability to ablation, as measured by the difference in cosine similarity between control images and ablated images (see Δ(cosine similarity) in Fig. 5). For ablation strenght of 1 (yellow/light lines), the difference increased with network robustness, and slopes in Table 1 indicate this trend holds across ablation polarities and strengths. Moreover, the robustness effects translated to higher shape bias in the benchmark by Geirhos et al. (2022), with negative ablations affecting more the texture than the shape accuracy (Fig. 16). Overall, classification CNNs segregate object information to positive weights and texture or background information to negative weights, and that adversarially robust training further sharpens this sign-based segregation.

4.5 Feature segregation is not exclusive to class units

To determine whether feature segregation by weight sign is present beyond the output layer, we systematically analyzed intermediate and early layers of CNNs. We first searched for channels that approximated Dale's law, identifying those that predominantly provided positive or negative inputs to their downstream units in each layer. For each convolutional layer, we calculated the sign consistency of outgoing weights and ranked channels according to whether they sent mostly positive or mostly negative signals forward. We then visualized the preferred features of these channels using

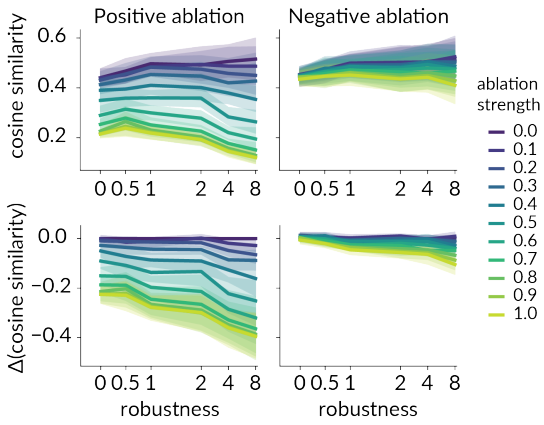

Figure 5: Representational changes under input ablation increase with robust training in ResNet50. Top: cosine similarities to control. Bottom: changes relative to control. Most ablation strengths show significant correlations with robustness.

Table 1: Spearman correlation of representational change upon ablation vs robustness (L_∞ norm)

type	Positive	Negative
α	ρ (pvalue)	ρ (pvalue)
0.1	-0.17 (2e-1)	-0.10 (4e-1)
0.2	-0.39 (1e-3)	-0.21 (8e-2)
0.3	-0.34 (4e-3)	-0.14 (3e-1)
0.4	-0.38 (1e-3)	-0.34 (5e-3)
0.5	-0.47 (6e-5)	-0.46 (9e-5)
0.6	-0.48 (3e-5)	-0.34 (5e-3)
0.7	-0.51 (9e-6)	-0.52 (6e-6)
0.8	-0.50 (2e-5)	-0.49 (2e-5)
0.9	-0.48 (4e-5)	-0.62 (2e-8)
1.0	-0.47 (6e-5)	-0.57 (5e-7)

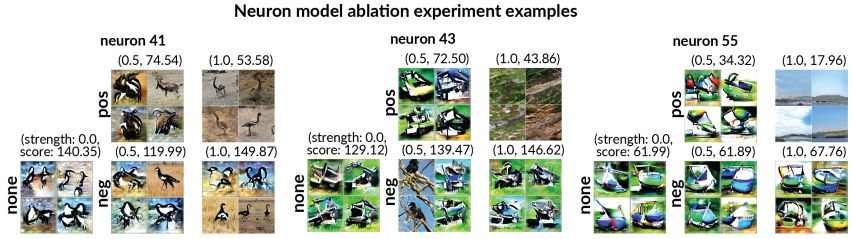

Figure 6: Preferred features of neuronal network models of visual neurons in the primate ventral stream. Pos: are positive ablations, neg are negative ablations, number indicates ablation strength. Shown are top 4 visualizations at 0, 0.5 and 1.0 ablation strenghts.

gradient-based methods with the Lucent library in PyTorch. Examining all five convolutional layers of AlexNet, we found that feature segregation by sign emerged throughout the network. In the first layer, channels with mostly positive weights responded to high-frequency achromatic edges, while those with mostly negative weights responded to lower-frequency, colored edges and spots. In the middle layers, positive channels tended to emphasize edges and detailed textures, whereas negative channels often represented broader, colored, or background-like patterns. By the final convolutional layer, channels with mostly negative weights produced features that resembled background elements such as sky or grass, while channels with positive weights highlighted sharp, localized object fragments like animal snouts and eyes (appendix Fig. 19). Altogether, our results show that feature segregation by weight sign is not restricted to the output layer, but gradually develops throughout the network. This pattern is reminiscent of Dale's law in biological circuits, suggesting that artificial neural networks can develop sign-consistent and functionally distinct representations across all layers, even in the absence of a biological constraint.

4.6 Biological models based on ImageNet networks segregate local feature information into positive weights

The ventral stream in primates is responsible for object recognition, and artificial networks are the best models of its function. We therefore wondered if the segregation of positive and negative inputs we observed in networks might also occur in the brain. However, it is not currently possible to selectively remove positive or negative synaptic inputs from real neurons the way we can in artificial networks. To address this limitation, we fit linear models mapping CNN features to macaque ventral stream neural responses, and applied feature visualization to both model units and in vivo data. This allowed us to test feature segregation in biological representations and to generate testable neuro-

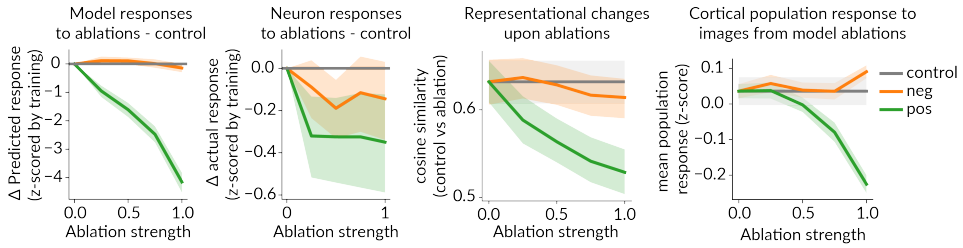

Figure 7: Left: predicted and actual neuron responses of model to ablations. Images obtained from positive ablations in the neuron models elicited a consistent activity drop on the biological neurons modeled. Right: Representational change of model to ablations measured by our cosine similarity metric on the neuron model feature visualizations upon ablation; and cortical population response to the images obtained from feature visualization from ablation of model units, neurons were z-scored before computing the population average. Plots show averages over 59 models, (35 for monkey C, and 24 for monkey D), shaded regions are the 95% C.I of the mean. For all plots the positive ablation condition was statistically different to the control.

science predictions. We recorded neural activity from V1, V4, and IT cortex in two monkeys, using a diverse set of images and modeled each neuron's response using partial least squares regression with activations from the penultimate layer of AlexNet (4096 units). [1] We then applied our ablation and visualization protocol to these neuron models (see appendix for validation A.4) and showed the resulting images to the monkey during the same session. First, images from intact models reliably drove biological neurons to firing rates more than one standard deviation above those observed during natural image presentations (Fig.21, left), indicating out-of-distribution generalization. For the subset of neurons in which we performed in vivo closed-loop feature visualization, we found that the model's preferred features often matched those of the neuron, providing additional validation (appendix Fig. 21, right). However, in vivo features were more spatially localized (procedure in A.1), whereas in silico features exhibited greater spatial variation (rotated, mirrored, or repeated versions). This likely reflects invariances, due to using a fully connected layer, that are not present in our recorded neurons. Moreover, unlike the images from recognition units, images from the neuron models did not resemble objects (appendix Fig. 21, Fig. 6).

Ablation experiments reveal sign-based segregation in neuron models. Ablation experiments on these neuron models showed that removing positive input weights led to a significant decrease in both predicted and observed firing rates, while removing negative weights had a smaller effect (Fig. 7). This pattern was consistent across individual neurons and at the population level in the ventral stream (Fig.7, rightmost). The population changes suggest that sign-based functional segregation in model predictions translates to measurable changes across the ventral stream population and perhaps perception.

Dale's law inspired analysis. To bring our models closer to Dale's law, we applied two approaches. First, we tested whether neural responses could be predicted using only positive input weights, corresponding to receiving input exclusively from excitatory artificial neurons. This constraint reduced both training and test accuracy relative to unconstrained models (appendix Fig. 22), indicating that neuron models require both positive and negative inputs. Second, we identified Dale-like artificial units that contributed mainly positive or negative weights to all output neuron models, defining putative excitatory and inhibitory inputs. Positive-weight units corresponded to smaller-scale edges and localized spots, cleanly separated from the background, whereas negative-weight units aligned with broader textures and larger patches (appendix Fig. 23). This pattern supports the idea that inhibitory-like artificial inputs preferentially encode contextual or background structure, paralleling inhibitory modulation in biological cortex.

Experimental manipulation of background as a test for inhibition. To further test this hypothesis, we experimentally manipulated image backgrounds in vivo. In a subset of experiments, we presented altered images in which the background was cleared around the neuron's preferred feature,

[1] Although using other layers could improve predictive accuracy, we selected the penultimate layer to directly test if inputs optimized for classification maintain weight sign-based segregation in biological neural predictions.

thereby reducing the putative inhibitory drive. As predicted, this manipulation resulted in increased neuronal responses (appendix Fig. 24), providing functional support for the idea that inhibitory or negative inputs are involved in contextual modulation and that their reduction can enhance feature selectivity in high-level visual cortex. Together, these results suggest that functional segregation by input sign extends to models of ventral stream neurons, providing concrete testable predictions for future experiments targeting excitation and inhibition in visual cortex.

5 LIMITATIONS

Our results hold in the last layer units of multiple networks. Due to limited computing time, we did not test all 1000 categories in as many networks as possible, our largest test consisted of 100 units. While larger scale simulations will provide exhaustive evidence, we are confident our main claims will stand. We limited our neuron recordings to a 160 image dataset for regressing neuron responses via CNNs. While we observed good fits and recovered relevant feature to the neurons, more images may improve the models, especially using larger-scale versions of our diverseSet. The neuroscience results would need to follow Dale's law to be mapped one-to-one to excitatory and inhibitory neurons, but we make no claim to a perfect mapping in this work. The fundamental question of shape vs texture, foreground vs background remains to be solved. We reconciled changes in frequency structure with changes in objectness and visual representations and LPIPS image similarity. However, more work remains to understand the full extent of the segregation reported here. Solving this problem for the visual cortex may provide better benchmarks for this task in AI.

6 DISCUSSION

Our study combined ablations with feature visualization guided by naturalistic image priors to reveal the functional segregation of class-level features in the output layer of ImageNet trained CNNs: positive weights contribute object/shape/low frequency information, while negative weights contribute background/contextual/texture information. This effect was enhanced in robust networks, it was present in networks with unsupervised pretraining, but was absent in network trained with Tanh instead of ReLU. Our results explain how the background contribution to classification observed in (Xiao et al., 2020) emerges, backgrounds are primarily encoded by the negative inputs.

Importantly for neuroscience, the observed functional segregation in neuron model units in CNNs hints at a functional segregation in the brain beyond the center-surround classically studied in V1. And we crafted a diverse dataset for visual neuroscience recordings that is scalable. Neuron responses to a smaller but diverse set of naturalistic, colored images, with complex foregrounds and backgrounds, led to models capturing relevant features obtained experimentally from the neuron. Thus, using both model-based and model-free approaches revealed richer neuronal representations. Preferred images from neuron models with positive input ablations elicited smaller average population responses of cortical neurons. This suggests that ablation in networks modeling neurons holds potential as a method to control the population activity in the brain. To relate ablation-induced changes in the images to the population responses is a future direction. This ablation based on the natural division of positive and negative weights can be easily extended into arbitrary layers, e.g., using gradients to define positive and negative contributions to any arbitrary unit. And our ablation approach proposes baselines for the functional differences between excitatory and inhibitory neurons in higher cortical visual areas. The functional segregation has consequences for neural coding and response selectivity. Our findings generate concrete predictions for future experiments using advanced genetic or optogenetic tools to dissect excitation and inhibition in primate cortex. Understanding the circuit mechanism of biological vision could aid further understanding and development of computer vision models. Interpretability is thus an important field for both AI and neuroscience.

REFERENCES

fastai/imagenette, May 2024. URL https://github.com/fastai/imagenette. original-date: 2019-03-06T01:58:45Z.

Matteo Alleman, Jack Lindsey, and Stefano Fusi. Task structure and nonlinearity jointly determine learned representational geometry. October 2023. URL https://openreview.net/forum?id=k9t8dQ30kU.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x. Publisher: Nature Publishing Group.

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019. doi: 10.1126/science.aav9436. URL https://www.science.org/doi/full/10.1126/science.aav9436. Publisher: American Association for the Advancement of Science.

Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, September 2008. doi: 10.1073/pnas.0803390105. URL https://www.pnas.org/doi/full/10.1073/pnas.0803390105. Publisher: Proceedings of the National Academy of Sciences.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis, February 2019. URL http://arxiv.org/abs/1809.11096. arXiv:1809.11096 [cs, stat].

Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning, November 2020. URL http://arxiv.org/abs/2011.10566. arXiv:2011.10566.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

Henry Dale. Pharmacology and Nerve-endings (Walter Ernest Dixon Memorial Lecture). *Proceedings of the Royal Society of Medicine*, 28(3):319–332, January 1935. ISSN 0035-9157. doi: 10.1177/003591573502800330. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2205701/.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434, February 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.01.010. URL https://www.sciencedirect.com/science/article/pii/S089662731200092X.

Alexey Dosovitskiy and Thomas Brox. Generating Images with Perceptual Similarity Metrics based on Deep Networks, February 2016. URL http://arxiv.org/abs/1602.02644. arXiv:1602.02644 [cs].

Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial Examples that Fool both Computer Vision and Time-Limited Humans. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/8562ae5e286544710b2e7ebe9858833b-Abstract.html.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. URL https://www.researchgate.net/profile/Aaron-Courville/publication/265022827_Visualizing_Higher-Layer_Features_of_a_Deep_Network/links/53ff82b00cf24c81027da530/Visualizing-Higher-Layer-Features-of-a-Deep-Network.pdf.

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, November 2023. ISSN 1546-1726.

doi: 10.1038/s41593-023-01442-0. URL https://www.nature.com/articles/s41593-023-01442-0. Publisher: Nature Publishing Group.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, November 2022. URL http://arxiv.org/abs/1811.12231. arXiv:1811.12231 [cs].

Umut Güçlü and Marcel A. J. van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, January 2017. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.12.036. URL https://www.sciencedirect.com/science/article/pii/S1053811915011490.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL https://arxiv.org/abs/1512.03385v1.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, October 1959. ISSN 0022-3751. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/.

Chou P. Hung, Gabriel Kreiman, Tomaso Poggio, and James J. DiCarlo. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866, November 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1117593. URL https://www.science.org/doi/10.1126/science.1117593.

Jaewon Hwang, Andrew R. Mitz, and Elisabeth A. Murray. NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB. *Journal of Neuroscience Methods*, 323:13–21, July 2019. ISSN 1872-678X. doi: 10.1016/j.jneumeth.2019.05.002.

Jeffry S. Isaacson and Massimo Scanziani. How Inhibition Shapes Cortical Activity. *Neuron*, 72(2):231–243, October 2011. ISSN 0896-6273. doi: 10.1016/j.neuron.2011.09.027. URL https://www.cell.com/neuron/abstract/S0896-6273(11)00879-8. Publisher: Elsevier.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, June 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0392-5. URL https://www.nature.com/articles/s41593-019-0392-5. Number: 6 Publisher: Nature Publishing Group.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. FFCV: Accelerating Training by Removing Data Bottlenecks, June 2023. URL http://arxiv.org/abs/2306.12517. arXiv:2306.12517.

Tianqin Li, Ziqi Wen, Yangfan Li, and Tai Sing Lee. Emergence of Shape Bias in Convolutional Neural Networks through Activation Sparsity. *Advances in Neural Information Processing Systems*, 36:71755–71766, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e31c16c7b3e0ccee5159ae5443154fac-Abstract-Conference.html.

Ilya Loshchilov. LM-CMA: an Alternative to L-BFGS for Large Scale Black-box Optimization, November 2015. URL http://arxiv.org/abs/1511.00221. arXiv:1511.00221 [cs, math].

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, September 2019. URL http://arxiv.org/abs/1706.06083. arXiv:1706.06083.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, November 2016a. URL http://arxiv.org/abs/1605.09304. arXiv:1605.09304 [cs].

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks, February 2016b. URL https://arxiv.org/abs/1602.03616v2.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2(11):e7, November 2017. ISSN 2476-0757. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL https://www.nature.com/articles/381607a0. Publisher: Nature Publishing Group.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models, June 2025. URL http://arxiv.org/abs/2504.02821. arXiv:2504.02821 [cs].

Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.e10, May 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.04.005. URL https://www.cell.com/cell/abstract/S0092-8674(19)30391-5. Publisher: Elsevier.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do Adversarially Robust ImageNet Models Transfer Better?, December 2020. URL http://arxiv.org/abs/2007.08489. arXiv:2007.08489 [cs, stat].

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. URL http://arxiv.org/abs/1409.1556. arXiv:1409.1556 [cs].

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014. URL http://arxiv.org/abs/1312.6199. arXiv:1312.6199 [cs].

Hiroshi Tamura, Hidekazu Kaneko, Keisuke Kawasaki, and Ichiro Fujita. Presumed Inhibitory Neurons in the Macaque Inferior Temporal Cortex: Visual Response Properties and Functional Interactions With Adjacent Neurons. *Journal of Neurophysiology*, 91(6):2782–2796, June 2004. ISSN 0022-3077. doi: 10.1152/jn.01267.2003. URL https://journals.physiology.org/doi/full/10.1152/jn.01267.2003. Publisher: American Physiological Society.

Edgar Y. Walker, Fabian H. Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G. Fahey, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22(12):2060–2065, December 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0517-x. URL https://www.nature.com/articles/s41593-019-0517-x. Publisher: Nature Publishing Group.

Binxu Wang and Carlos R. Ponce. High-performance evolutionary algorithms for online neuron control. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '22, pp. 1308–1316, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9237-2. doi: 10.1145/3512290.3528725. URL https://dl.acm.org/doi/10.1145/3512290.3528725.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, July 2022. URL https://arxiv.org/abs/2207.02696v1.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or Signal: The Role of Image Backgrounds in Object Recognition, June 2020. URL http://arxiv.org/abs/2006.09994. arXiv:2006.09994 [cs].

Will Xiao and Gabriel Kreiman. XDream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Computational Biology*, 16(6):e1007973, June 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007973. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007973. Publisher: Public Library of Science.

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification, May 2019. URL http://arxiv.org/abs/1905.00546. arXiv:1905.00546 [cs].

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, June 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/10.1073/pnas.1403112111. Publisher: Proceedings of the National Academy of Sciences.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, January 2018. URL https://arxiv.org/abs/1801.03924v2.

A Appendix

A.1 Extended methods

Networks The ablation studies were performed on CNNs pretrained on the ImageNet dataset: AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan & Zisserman, 2015), ResNet50 (He et al., 2015), and robustly-trained ResNet50 ($L_\infty \in \{0.5, 1, 2, 4, 8\}$, Salman et al. (2020)). All these networks end on a 1000-unit fully connected layer, each unit corresponding to one of the 1000 ImageNet categories. Neural networks were used in Pytorch.

ImageNet subsampling To reduce computing time, for most of the experiments, we used a subset of ImageNet, the *imagenette* dataset (noa, 2024) and the macaque category, 11 classes in total. These classes and their corresponding output units in each network trained on the 1000-class ImageNet dataset are as follows: (0, tench), (207, English Springer), (482, cassette player), (491, chain saw), (566, church), (569, French horn), (571, garbage truck), (574, gas pump), (701, golf ball), (970, parachute), and (373, macaque). We visualized the representations of the output layer units of those classes under different ablation conditions. For Fig. 12, to sample 100 diverse classes out of the 1000 ImageNet classes, the 50k validation images were first clustered into 100 clusters via agglomerative clustering of the L2 distance matrix from the 1000-d output features of ResNet50, which was pre-trained on ImageNet. Then, one new unique class is selected from each cluster.

Ablation We used two ablation conditions: we ablated weights that were (1) only positive or (2) only negative. We ablated weights cumulatively by first sorting the positive (or negative) weights by their (absolute) decreasing value. We defined the *ablation strength*, α, as a fraction of the total positive or total negative weights to a unit. We identified the top k weights necessary to reach the silencing strength, i.e., $\sum_{i=1}^{k} w_i \leq \alpha$, and set them to zero. We covered the range of ablations from 0 to 1. For most experiments with ANNs, we used silencing strengths in 0.1 steps, from 0 (intact) to 1 (complete ablation).

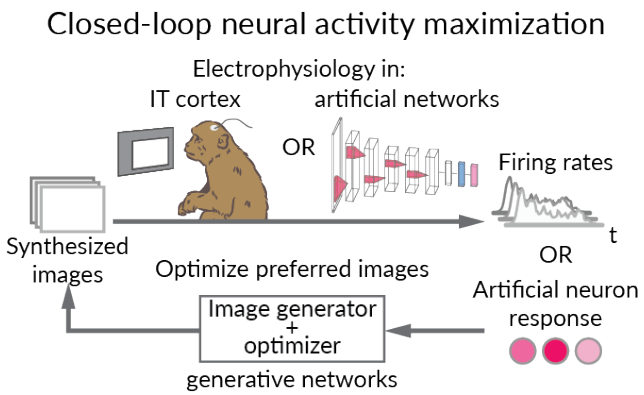

Figure 8: Schematic of feature visualization workflow in ANNs and brains. Optimizer is CMAES, image generators are DeePSim fc6 or BigGAN.

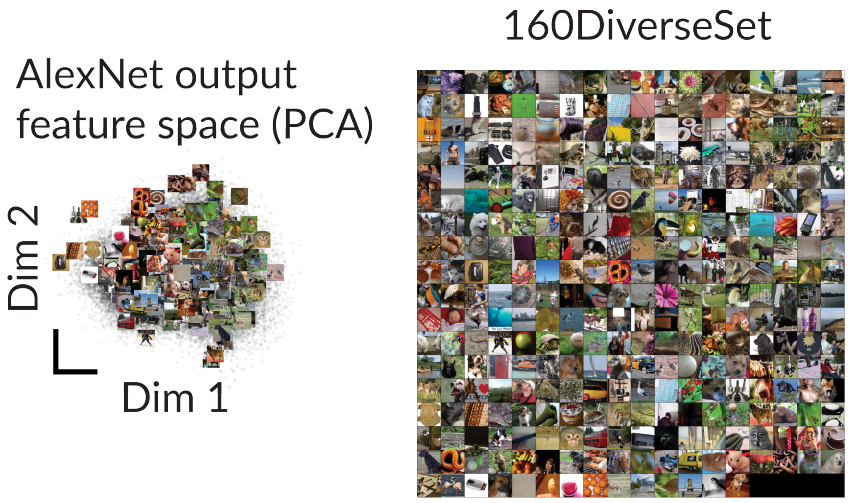

Figure 9: Illustration of a diverse dataset construction using AlexNet output feature space. The embedding is the output of the last layer before softmax of AlexNet, a vector space of 1000-dimensions. Left: PCA showing the coverage of the feature space by the diverseSet 160, only for illustration purposes. Right: images from diverseSet 160 used to fit neuron models.

Feature visualization For each ablation condition, we performed feature visualization by optimizing a GAN latent code to create an activity-maximizing image Fig. 8. We used this closed-loop, zeroth-order-search approach to allow comparison with our neuronal experiments, where gradient ascent would not be possible. To increase the span of the stimulus space, we used two GANs: AlexNet fc6 DeePSiM (Dosovitskiy & Brox, 2016) and BigGAN (Brock et al., 2019). For optimization, we used a variant of *covariance matrix adaptation evolutionary strategy* or CMAES (Wang & Ponce, 2022; Loshchilov, 2015). Initial conditions for the CMAES were given as standard deviation of 3.0 for DeePSim, and 0.2 for BigGAN. Initial images for the algorithm were small norm vectors for both GANs, close to the origin of the latent spaces. For BigGAN, we generated a fixed noise vector by scaling a 128-dimensional truncated noise sample (-1.4, 1.4), and concatenated it with a 128-dimensional zero vector of the class embedding, to form the required 256-dimensional input code. The remaining parameters are determined by the dimensionality of the search space of each GAN. We optimized ten images per GAN, resulting in 20 feature visualizations per output unit and ablation condition. Diverse visualizations better capture the multifaceted high-level representations in CNNs (Nguyen et al., 2016b). For our examples, we show the best of the 20 visualizations, but used all for quantitative analyses. For visualizations of neural networks predicting biological neuron responses, due to experimental time restrictions, we used five visualizations per ablation condition, via DeePSim only. Our experiments are performed in a PC with Nvidia 4090 GPU, and each visu-

alization running 100 iterations takes about 3 mins. For *in vivo* experiments, we ran from 20 to 60 iterations of the AlexNet fc6 DeePSiM with the CMAES algorithm implemented in Matlab, linked to our real-time spike-sorting data acquisition. The responses fed to the CMAES algorithm were the average firing rate on the window 70-170 ms from image onset.

Feature analysis We computed image similarity using an ensemble of CNNs, including AlexNet, ResNet50, and ResNet50 with robustness in $L_\infty \in \{0.5, 1, 2, 4, 8\}$, inspired by (Feather et al., 2023) And confirmed the results with LPIPS (Zhang et al., 2018) in the appendix. We computed their activations and defined similarity as the average pairwise cosine similarity (LPIPS) between control activity vs input-ablated activity. We averaged the results of the CNNs ensemble, resulting in one quantity per ablation condition. We computed *objectness* as the maximum bounding box score provided by YOLOv7 (Wang et al., 2022), this was averaged over visualizations per unit, units per network, and then across networks.

Visual cortex electrophysiology We collected data from two animals (monkey C and monkey D), each implanted chronically with floating multielectrode arrays (Microprobes for Life Sciences, MD) of 32 or 16 channels (monkey C, N = 96 electrodes, monkey D, 64), in areas V1, V4 and posterior inferotemporal cortex (PIT). All institutional procedures were followed. Channels were distributed as (V1, V4, PIT): monkey C (32, 32, 32), monkey D (16, 16, 32). Some electrodes captured the activity of single units, but most showed multi-unit activity (reflecting the pooled activity of micro-clusters of neurons). The animals performed a simple fixation task, which required them to keep their eyes on a 0.25-deg diameter spot at the center of the screen, within a square fixation window measuring 0.5–1° per side. Images were presented for 100 milliseconds ON, 150-ms off, 4-5 images per trial, after which the animal received water or juice. Images were presented to monkey C were 2 deg in size, and 4-8 deg for monkey D to match the receptive field centers of most channels in all cortical areas (V1, V4 and PIT). Image presentation and data acquisition (electrophysiology, eye tracking) were integrated by the MonkeyLogic2 software (Hwang et al., 2019) and OmniPlex Neural Recording Data Acquisition Systems (Plexon Inc.), interfaced through custom Matlab code. We performed online spike sorting using the PlexControl client based on waveforms. We used ViewPixx EEG monitors (ViewPixx Technologies), at a resolution of 1920x1080 pixels with 120 Hz refresh rate. Eye tracking used ISCAN cameras (ISCAN Inc.). And reward was delivered using the DARIS Control Module System (Crist Instruments).

Feature localization in vivo We conducted a perturbation-based localization to identify relevant image regions from a feature visualization performed in vivo, where gradient information from the animal brain is unavailable. We perturbed a circular region with a 50-pixel diameter within the 256-pixel image by randomly shuffling the pixels inside this circle, effectively disrupting the local image structure while maintaining local contrast. We selected 30 such regions for perturbation at random, excluding those that extended beyond the image boundaries. The modified images were then presented to the monkey. We hypothesized that perturbing regions crucial for driving the neuron response would lead to a decreased firing rate. To assess local image importance, we calculated the normalized response change: the difference between the firing rate response to the intact image and the firing rate response to the perturbed image, divided by the firing rate response to the intact image. A normalized response change of 0.5 indicates the neuron response decreased by half due to perturbation. To generate the localized response mask, we averaged the circular masks corresponding to each perturbed region, weighted by their response change. This response mask was further smoothed using a Gaussian kernel with a 30-pixel standard deviation. We defined relevant regions as those causing a normalized response change of 0.5 or greater. Finally, we applied this mask to the original feature visualization image to highlight the local features.

Image dataset We collected a reference image dataset to activate neurons in the monkey along the hierarchy of V1, V4, and PIT. Because neurons vary in their preferred features, we constructed a dataset spanning the image space as represented by the neural embedding of ImageNet-trained AlexNet. The embedding is the output of the last layer before softmax of AlexNet, a vector space of 1000-dimensions. The images from this dataset also spanned uniformly the 1000-dimensional output space of a semi-supervised trained network, trained on a billion images, ResNet50SS (Yalniz et al., 2019). To define this embedding space, we performed PCA on the output activations from AlexNet to the 50k ImageNet validation images, we kept the top 300 components (accounting for

about 95% of total explained variance). Then we partitioned the space into a defined number of clusters k, according to the desired dataset size, using batched k-means to reduce computational burden. After finding the k cluster centers, we could feed arbitrary images to the network, map them to the PCA space, and then pick the nearest neighbors to the cluster centers from the desired image space. In addition to the ImageNet validation set, we added other common neuroscience datasets (Brady et al., 2008; Kar et al., 2019; Allen et al., 2022; Hung et al., 2005) to form our image space. We selected $k = 160$ images, as a set that was diverse but small enough to be used in every experimental session. We called this image dataset *diverseSet* .

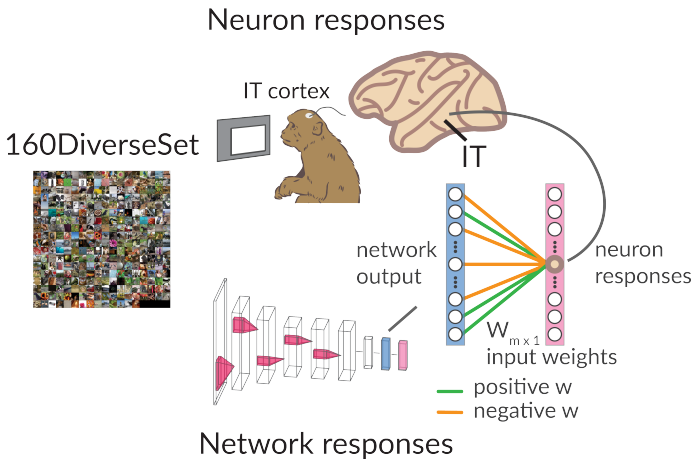

Figure 10: Schematic of model fitting using the dataset diverseSet. 160 images were split into train/test datasets (80/20).

Models fit on neuronal activity We recorded responses of neurons in the ventral stream to a 160 image dataset, our diverseSet Fig. 10. We relied on a small dataset to fit neuron responses and perform feature visualizations within the same experimental session. We performed partial least-squares linear (PLS) regression (80/20 train/test split) between the neuron responses to images and the activations of the penultimate layer of AlexNet. We used one component for the PLS regression. We selected one neuron or microcluster per experimental session, fitted a model, and performed the ablation and feature visualizations *in silico* for that model. We selected the best fitted neuron per session, based on the r^2 on the 20 % held out test set, usually in the range of 0.15 to 0.5. When time allowed, we also performed the feature visualization of the modeled neuron *in vivo* using a gradient-free approach (Ponce et al., 2019), within the same experimental session. To test whether features learned by the model were relevant to the biological neuron, we recorded the neuronal responses to the preferred images of the model. We then analyzed the representational similarity of the model features under ablations using ANNs. And analyzed the responses of the biological neuron populations from V1, V4 and IT.

A.2 Supporting results

Table 2: Ratio of positive to negative weights. We divided the sum of positive weights by the sum of the absolute values of the negative weights.

MODEL	**RATIO (MEAN ± STD)**
AlexNet	1.03 ± 0.08
VGG16	1.01 ± 0.09
ResNet50	1.00 ± 0.06
ResNet50 ($L_\infty = 0.5$)	1.00 ± 0.05
ResNet50 ($L_\infty = 1$)	0.99 ± 0.05
ResNet50 ($L_\infty = 2$)	1.00 ± 0.04
ResNet50 ($L_\infty = 4$)	1.00 ± 0.05
ResNet50 ($L_\infty = 8$)	1.01 ± 0.05

Table 3: Dale index of the final classification layer before and after training, and trained Top1 accuracy.

Model	Untrained DI	Trained DI	Δ DI	Top1 Acc
alexnet	0.5126	0.5461	+0.0335	56.52
densenet121	0.5125	0.5567	+0.0442	74.43
densenet169	0.5126	0.5590	+0.0464	75.60
densenet201	0.5126	0.5601	+0.0475	76.90
resnet18	0.5128	0.5905	+0.0778	69.76
resnet34	0.5130	0.5923	+0.0793	73.31
resnet50	0.5125	0.6004	+0.0879	80.86
resnet101	0.5126	0.5998	+0.0872	81.89
resnet152	0.5127	0.6007	+0.0880	82.28
vgg11	0.5123	0.5570	+0.0447	69.02
vgg11_bn	0.5123	0.5737	+0.0614	70.37
vgg13	0.5125	0.5583	+0.0457	69.93
vgg13_bn	0.5125	0.5747	+0.0622	71.59
vgg16	0.5126	0.5606	+0.0480	71.59
vgg16_bn	0.5126	0.5761	+0.0635	73.36
vgg19	0.5129	0.5619	+0.0490	72.38
vgg19_bn	0.5129	0.5748	+0.0619	74.22
inception_v3	0.5127	0.5597	+0.0470	77.29

A.3 Dale's Law inspired analysis of intermediate features

To determine if weight segregation of features occurs beyond the output layer, we visualized feature representations that predominantly provide negative or positive inputs to subsequent layers in AlexNet. We calculated sign consistency by averaging spatial weights and determining the frequency of positive and negative weights across output channels. The visualization of sign-consistent input features was conducted using the Lucent library in PyTorch, leveraging gradient-descent channel activity maximization. We focused on AlexNet's intermediate layers, examining the top and bottom sign-consistent features for each input channel.

Layer Details:

Conv1: Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
Conv2: Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
Conv3: Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
Conv4: Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
Conv5: Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))

Features that contributed mostly positive weights differed from the features contributing mainly negative weights, with object vs background arising with increasing depth. This positive vs negative

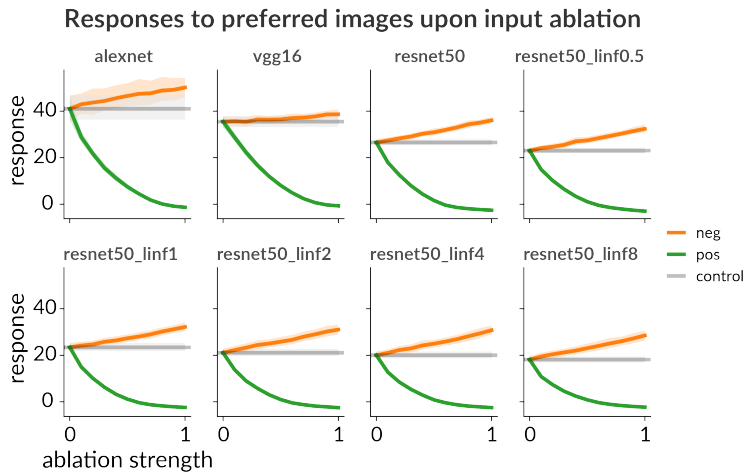

Figure 11: Mean activation scores of units used in ablation experiments. For all networks, units scores come from the last fully-connected layer, with 1000 units, before the softmax. The units correspond to the 10 imagenette categories ([0, 217, 482, 491, 497, 566, 569, 571, 574, 701]) plus the macaque category (373). Error bars are 95% confidence intervals over units (categories tested), where each unit response is the mean of its 20 visualizations. *Control* refers to the feature visualizations in the intact networks for the same units, we extended it as a horizontal line to ease visual comparisons to the different ablation strengths.

weight split is evident even in the first layer, where low-frequency color features are contrasted with high-frequency black-and-white features.

Figure 12: Functional segregation holds in a 10x larger dataset. 100 classes out of the 1000 ImageNet categories were selected by clustering the 50k validation images embedded in the 1000-d output space of ResNet50 picking one class per cluster. Thus, we now have 10x more data points that should span the representational space of the output layer we study. Consistent with the smaller dataset, the main object features degrade into more uniform background images upon positive ablation. Here we show examples from 10 of the 100 classes.

A.4 Biological neuron models

For each recording session, we selected the best model for further analysis, based on predictive accuracy (mean test $r^2 = 0.27 \pm 0.10$ SD across sessions). The fitted models included both positive and negative input weights, with a mean ratio of 1.17 for the sum of positive to negative absolute weights (Fig. 20). Our final dataset comprised (V1, V4, pIT): (7, 5, 23) neurons in Monkey C and (1, 5, 18) in Monkey D, with the majority of data from pIT cortex.

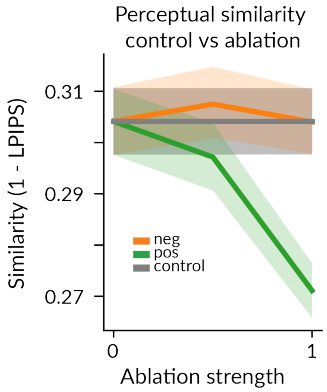

Figure 13: Functional segregation holds in a 10x larger dataset with LPIPS (Zhang et al., 2018) as representational similarity measure. We measured the representational similarity of the images as 1 - LPIPS among control images and between control images and ablation images (Fig. 12). We average results per class, and show the mean and 95% C.I. across the 100 classes. The representational similarity degrades upon positive input ablations, confirming results obtained from the imagenette dataset.

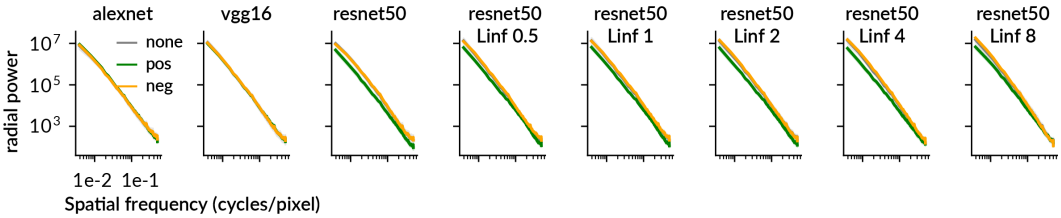

Figure 14: Functional segregation of frequency content. We measured the radial power spectrum for control and ablation images of different networks. We average results per class, and show the mean and 95% C.I. across the 11 classes. Low frequencies degrade upon positive input ablations, while negative ablations overlap with control spectra.

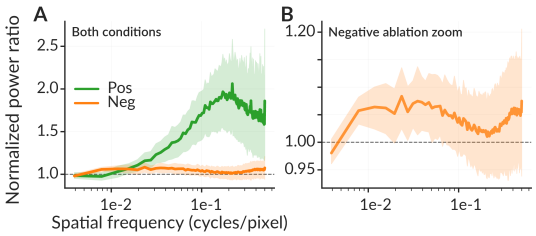

Figure 15: Functional segregation of frequency content. We measured the radial power spectrum for control and ablation images of different networks. We average results per condition and per network, and show the ratio of control to ablated spectra. The mean and 95% C.I. are across networks. Low frequencies degrade upon positive input ablations, while negative ablations overlap with control spectra but slightly enhance low frequencies.

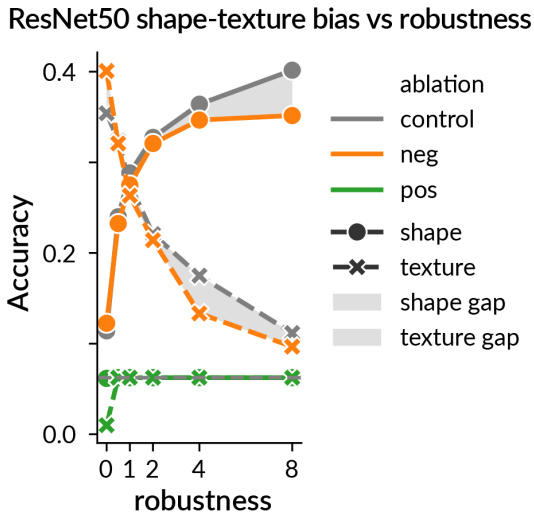

Figure 16: Functional segregation of shape vs texture. We measured robustness effect on shape vs texture encoding in ResNet50 using the shape/texture bias benchmark (Geirhos et al., 2022). Robustness induces a shift from texture to shape, and negative ablations disrupt more strongly the texture encoding of robust networks.

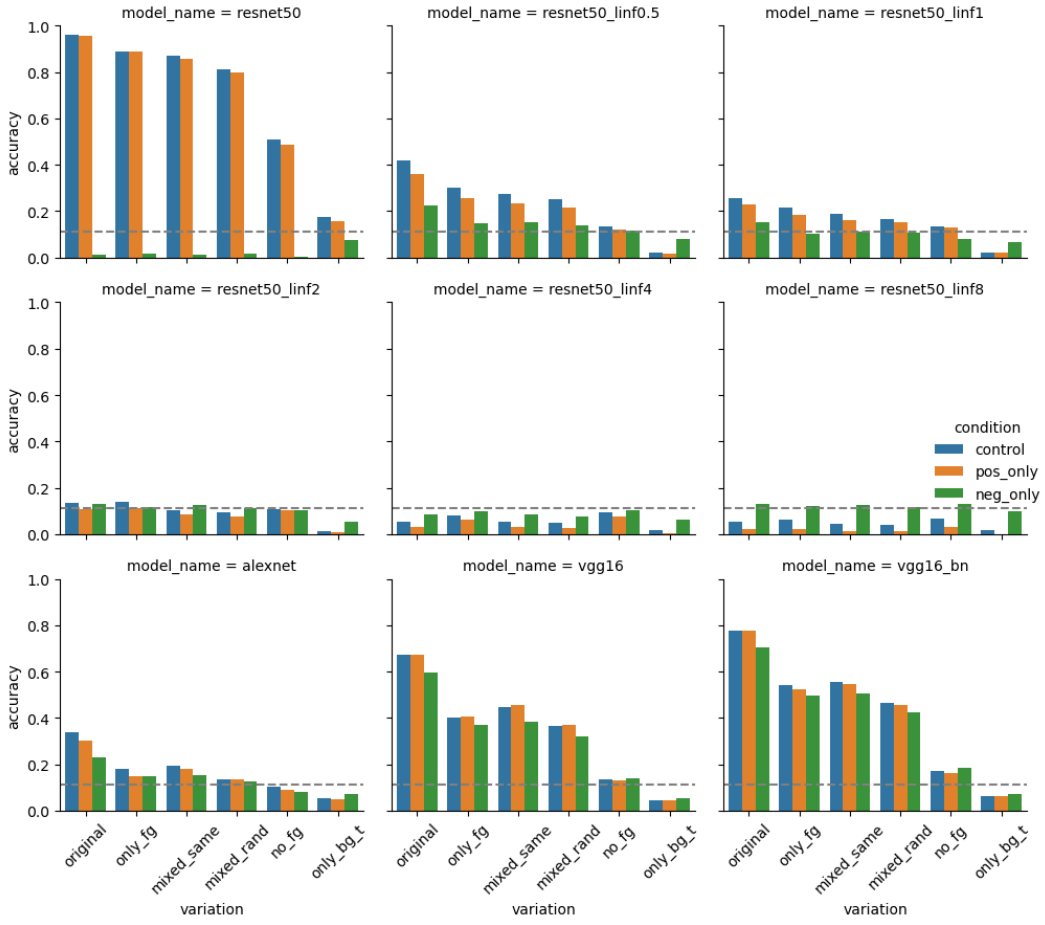

Figure 17: Functional segregation of background vs foreground. We measured the accuracy on the background challenge (Xiao et al., 2020). Robustness increases the role of negative weights in encoding backgrounds but below the chance level. Counterintuitive results show either sign can solve this challenge in AlexNet and VGG.

Figure 18: Feature visualizations of ablation experiments in a network pretrained with unsupervised learning. ResNet50SimSiam (Chen & He, 2020). The unsupervised network with frozen weights was coupled to a fully connected layer, only this layer was fine-tuned to classify ImageNet1000. Network units changed starting with small positive weight ablations, see unit 574 golf ball. Smaller changes are visible upon negative weight ablations, however object relevant features remain. Overall behavior is consistent with CNNs trained directly on ImageNet1000 classification.

Figure 19: Layer Conv5 from Conv4: Features contributing mainly negative weights resemble backgrounds, such as patches of sky and grass, and sometimes face-like features (e.g., in the tench class), highlighted in orange borders. Positive weights align with localized object-like fragments, such as snouts and eyes of animals, and sharp spotted textures vs the blurry spotted textures for negative weights. Layer Conv4 from Conv3: Negative features still incorporate some background elements like ground or grass textures (orange borders), together with some spiral, square and blurry textures. Positive features exhibit more heterogeneous textures and higher frequency details, without evident background-like textures. Layer Conv2 from Conv1: Positive weights carry high-frequency edges mostly without color, while negative weights include lower frequency edge features and spotted textures with color, overall more spatially coarse. Channel index from the visualized features is shown as a list below each panel.

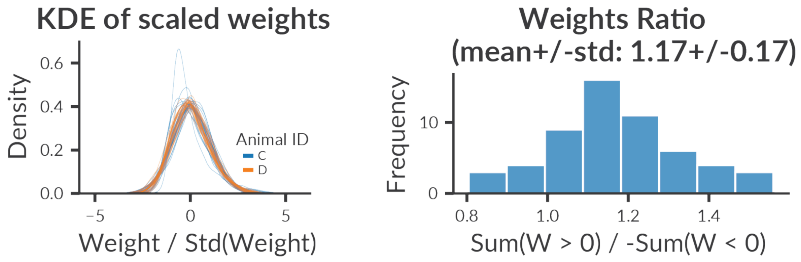

Figure 20: Left: Distribution of the model weights from neuronal fits with AlexNet penultimate layer features. Each model maps 4096 parameters from penultimate layer of AlexNet to the response of one biological neuron. Models use positive and negative weights. Model weights were normalized by their standard deviation to plot them on the same scale, for sake of visualization. Right: Ratio of total positive to total negative weights, per neuron model. Models use slightly larger positive weights with a mean of 1.17 and std of 0.17. Model numbers: 35 for monkey C, and 24 for monkey D.

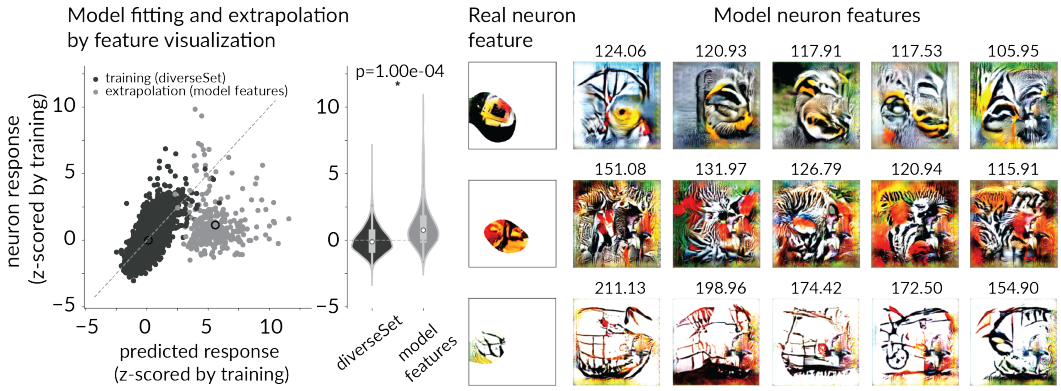

Figure 21: Neuron model units recover features relevant for the biological neurons. Left: Responses vs predicted responses of neurons to the training images, and the extrapolated features visualized from the intact models, which are extrapolations because the training data did not cover those high response ranges. Permutation t-test of neuron responses shows higher responses to images from model features than the natural images of the training dataset (diverseSet). Right: three neuron examples that show the feature visualization of the preferred feature of the neuron masked by the full-width at half-maximum obtained from perturbations to the image, and to their right the five feature visualizations of the intact model with the real neuron responses to those images on top.

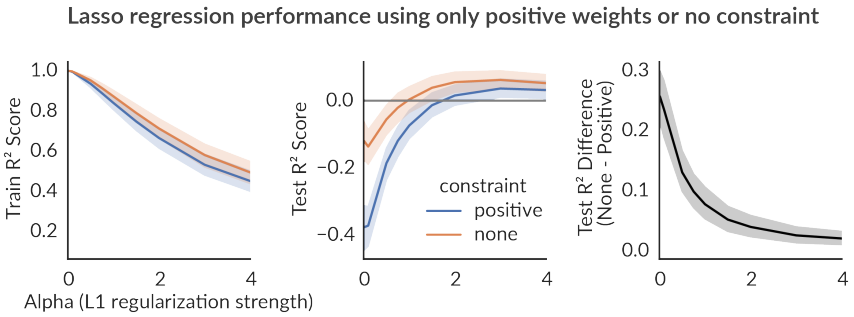

Figure 22: Using negative weights improves neuron models obtained via Lasso regression. Lasso regression models were fit with and without the positive constraint, over a 5-fold cross validation. Models were a linear regression from the 4096 features to a single neuron, over all neurons modeled from both animals. Left: performance on the training set measured by r^2 score. Middle: r^2 performance on the test set. Right: Model improvement by using positive and negative weights vs using only positive weights given by the difference in r^2 on the test set. Unconstrained models perform better than the positively constrained model, across the range of L1 penalties (sparseness penalty) tested, suggesting negative inputs from artificial network features are useful to predict biological neuron responses.

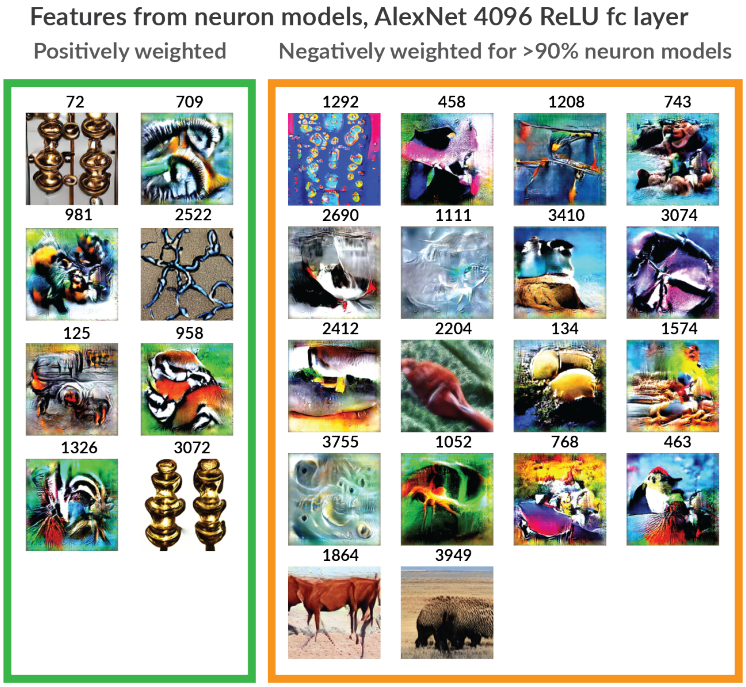

Figure 23: Features that had positive or negative weights in most of the neurons models (91% of the 56 neurons, binomial test $p = 5.09e^{-10}$). These features are the closest approximation to features respecting Dale's law from our models. Left: best of 20 feature visualizations for the features with positive weights across neurons, feature index is on top of the image. Features are from the penultimate fc layer post ReLU, containing 4096 units. Right: best feature visualization from the negatively weighted features across neurons. Positively weighted features contain more local features like curved edges, while negative features contain textures or larger image patches. Sign consistency tested for statistical significance against the Bernoulli distribution of 0.5 probability with Bonferroni correction for testing 4096 features.

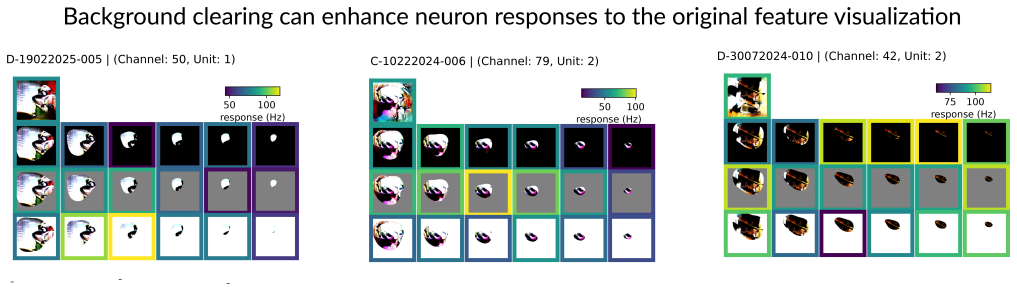

Figure 24: Clearing the background around the images obtained via closed-loop visualization can further boost responses in real-time recordings. Examples of 3 neurons in 2 monkeys.

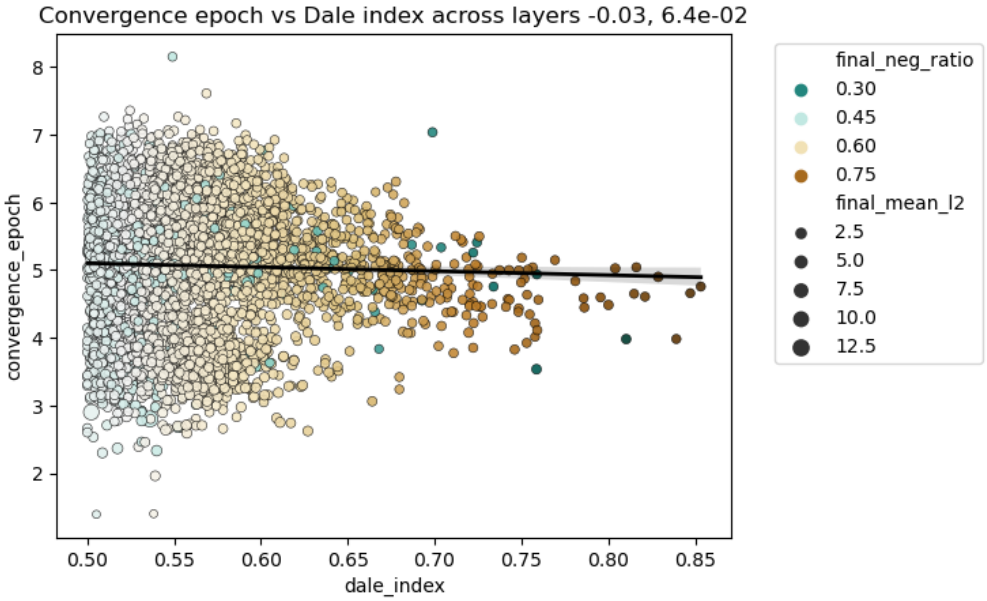

Figure 25: Weight dynamics vs Dale index. Convergence rate of outgoing weights measured by cosine similarity to final weights, as time to reach 90% of the final similarity. Each dot is one output channel for all layers in ResNet18 trained over 16 epochs. Color indicates proportion of negative signs. There is no obvious correlation. However, high Dale index channels mature within narrower time windows than more mixed channels.