Instance-level contrastive learning yields human brain-like representation without category-supervision

Talia Konkle *
Department of Psychology
Harvard University
Cambridge, MA 02478
talia_konkle@harvard.edu

George A. Alvarez
Department of Psychology
Harvard University
Cambridge, MA 02478
alvarez@wjh.harvard.edu

Abstract

Humans learn object categories without millions of labels, but to date the models with the highest correspondence to primate visual systems are all category-supervised. This paper introduces a new self-supervised learning framework: instance-prototype contrastive learning (IPCL), and compares the internal representations learned by this model and other instance-level contrastive learning systems to the structure of human brain responses. We present the first evidence to date showing that self-supervised systems can show more brain-like representation than category-supervised models. Further, we find that recent substantial gains in top-1 accuracy from instance-wise contrastive learning models do not result in more brain-like representation—instead we find the architecture and normalization scheme are critical. Finally, this dataset reveals substantial representational structure in intermediate and late stages of the human visual system that is not accounted for by any model, whether self-supervised or category-supervised. Considering both neuroscience and machine vision perspectives, these results provide promise for instance-level representation as a key objective of visual system encoding, and highlight the room to grow towards more robust, efficient, human-like object representation.

1 Introduction

A fundamental goal for machine vision is to learn useful, flexible, generalizable, robust visual representations, e.g. with comparable capacities to human vision. Drawing insights from biology has been valuable—the past decade's breakthroughs using deep convolutional neural networks have leveraged structural and algorithmic parallels to biological neural systems (LeCun et al., 2015). And, in a stunning act of reciprocity, these networks learn hierarchical visual representations that show an emergent match to the structure of visual brain responses to depicted objects (e.g. Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2018; Yamins et al., 2014), though certainly with room to improve (Geirhos et al. 2018; Xu and Vaziri-Pashkam 2020, see Serre 2019; Sinz et al. 2019 for recent reviews). In this way, comparing learned visual representations directly to brain representations can be an additional source of biological feedback for model development (Schrimpf et al., 2018).

Beyond the CNN architecture, further biological inspiration can be taken from the nature of the learning: humans do not learn from millions of labeled examples, and it has been argued that in the next decade of advances, neither should machines (e.g. LeCun et al., 2015). And, recently,

*

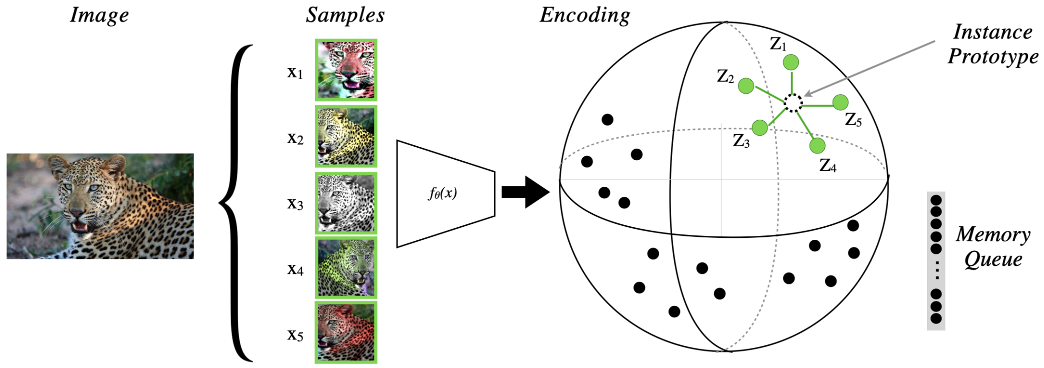

Figure 1: Schematic of our instance-prototype contrastive learning framework. We used a base convolutional neural network $f_\theta(x)$ to encode each image as a normalized 128-D feature vector (z_i). The feature embedding is learned by bringing each image projection z_i closer to its instance prototype $\bar{z}$, and separating it from the previously encountered z_i in the memory queue.

there has been substantial advances in self-supervised learning, particularly using instance-level contrastive learning (He et al., 2019; Hjelm et al., 2018; Misra and van der Maaten, 2019; Oord et al., 2018; Tian et al., 2019; Wu et al., 2018; Ye et al., 2019). For example, recent representation learning frameworks like SimCLR (Chen et al., 2020a) and Moco2 (Chen et al., 2020b) are now able to achieve dramatically higher overall categorization accuracy, approaching supervised model performance.

A key insight across these frameworks is that they operate over instances—individual images—where category-level representation develops as an emergent capacity of the learned representational space (see Wu et al., 2018). With a biological lens, this is a critical step towards a more plausible learning framework: categories do not need to be presupposed ahead of time but can be indexed from within a more generically useful representation that is learned solely from the structure of natural images and the architecture-induced prior of a hierarchical CNN. Do such models learn representations that are similar to those of supervised models, with similar matches to the brain, e.g. arrive at the same representation through different mechanisms? Or could these models be learning something different that is even more brain-like?

To examine these possibilities, this paper first presents a new self-supervised, contrastive-learning framework: *instance-prototype contrastive learning* (IPCL). The IPCL framework trains the model to build an online prototype representation of each instance from multiple samples (augmentations), and to represent each sample as dissimilar from previously viewed samples stored in an offline memory queue (non-indexed). Next, we compare this model and other instance-level contrastive learning systems to the representations measured in the human visual system, considering early, intermediate, and later hierarchical stages.

2 Instance-Prototype Contrastive Learning (IPCL)

In contrastive-learning frameworks, the goal is to learn an embedding function that maps images into a low-dimensional latent space, where visually similar images are close to each other, and visually dissimilar images are far apart. Learning proceeds by organizing the training data into similar pairs (positive samples) and dissimilar pairs (negative samples), where different frameworks make different choices of how positive and negative samples are computed and retained throughout the learning process (Doersch and Zisserman, 2017; Dosovitskiy et al., 2014; He et al., 2019; Ji et al., 2019; Misra and van der Maaten, 2019; Tian et al., 2019; Wu et al., 2018; Ye et al., 2019; Zhuang et al., 2019).

Our instance-prototype contrastive learning framework is depicted in Figure 1. We randomly augment the same image (x) multiple times (here, $n = 5$), then pass each augmented image $(x_i \ldots x_j)$ through an embedding function $f_\theta(x)$ to obtain a low-dimensional representation of each image $(z_i \ldots z_j)$. We then compute an instance prototype $\bar{z}$ by averaging the embeddings for all 5 samples:

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} f_\theta(x_i) \tag{1}$$

where n is the number of samples, $f_\theta(x)$ is the embedding function, and x_i is the i^{th} augmented sample of an image.

For each augmented instance, its prototype serves as its positive pair, and all stored representations serve as negative pairs (implemented with a lightweight, non-indexed memory queue storing the K=4096 most recent samples). The normalized temperature-scaled cross entropy loss for a positive pair $(z_i, \bar{z})$ would be defined as:

$$\ell_{z_i,\bar{z}} = -\log \frac{\exp(\text{sim}(z_i, \bar{z})/\tau)}{\exp(\text{sim}(z_i, \bar{z})/\tau) + \sum_{k=1}^{K} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2}$$

where the similarity function sim is the dot product between embeddings, τ is a temperature parameter that controls the dynamic range of the similarity function, and K is the total number of instances in the memory queue. In practice, we used Noise Contrastive Estimation (NCE, Gutmann and Hyvärinen, 2010) to approximate sampling from a larger memory store (see Wu et al. 2018; Appendix B.1), though recent work suggests the loss function in equation 2 may suffice (e.g. see Chen et al., 2020a). The final loss is computed across all positive pairs in a minibatch (128 images, 5 samples per image, yielding 640 positive pairs). The queue is updated after every minibatch with the current samples added to the queue, displacing the oldest samples.

Architecturally, we used an Alexnet as the base image encoder (Krizhevsky et al., 2012), replacing the 1000-dimensional output layer with a 128-dimensional fully-connected layer with an L2 norm, following Wu et al. (2018). Further, we modified the Alexnet to have group norm (gn) rather than batch norm (bn) layers, which enabled successful learning (see Appendix A). And, to preview our results, this slight modification had a consequential effect on emergent brain-like representation.

3 Related instance-level contrastive learning frameworks

Our IPCL model was inspired by Wu et al. (2018), where models were trained to perform instance-level discrimination, with a base CNN encoder, a normalized low-dimensional embedding space, and a ≈1.28M *indexed memory bank*. In their learning framework, the current representation of an image is compared with its prior representation which can be directly accessed from the indexed memory to form the positive pair. Negative pairings are obtained using other the other indexed memory representations, estimated using an empirically determined draw from the memory bank of 4096 negative samples. The category structure learned in the latent space of the L2 layer supported then state-of-the art top-1 ImageNet classification for a unsupervised system (e.g., Resnet50=42.5%), outperforming other leading self-supervised representation learning systems by a substantial margin (e.g. Jigsaw, Noroozi and Favaro 2016; SplitBrain autoencoder, Zhang et al. 2017).

From both biological and machine vision perspectives, the indexed memory bank is somewhat problematic, as the exact number of images to be represented is fixed, and during training this image index provides perfect memory access, a bit like an (external) supervised label. Our IPCL learning framework makes biologically-inspired modifications, by replacing the indexed memory bank with a non-indexed memory queue, and using multi-augmentation to produce positive pairings, rather than using a perfectly-indexed stored memory trace of the previous encounter with that item.

For comparison with IPCL, we trained a variety of networks using as Wu et al.'s indexed memory framework, varying both base encoding architecture and the dimensionality of the latent space, to examine whether these instance-level contrastive learning frameworks learn visual representations that are as brain-like as category-supervised models.

Concurrently, new instance-level contrastive learning frameworks like MoCo2 (Chen et al., 2020b) and SimCLR (Chen et al., 2020a) have emerged, which both avoid an indexed memory bank, and have dramatically increased emergent top-1 ImageNet classification accuracy compared to Wu's nets (MoCo2-Resnet50 = 71.1%; SimCLR-Resnet50-4x = 76.5%; Wusnet-Resnet50 = 42.5%). Like our IPCL framework, Moco2 uses a memory queue, though with a 16x larger queue (4096 vs. 65,536 items), and a completely different framework for generating positive samples, using a dual-network architecture. SimCLR, like our IPCL framework, uses augmentation for the positive pair, but with no instance-prototype, and a very large batch size for negative samples (4x more items than in IPCL, 4096 vs 16382). In the present work, we also included these pre-trained models in our model-brain comparisons.

4 Methods

Our goal is to compare the representations learned by these self-supervised vision systems with their category-supervised counterparts, assessing how they fit the representational structure of human brain responses in early, intermediate, and later hierarchical stages of the visual system. Below we outline the models, brain dataset, and evaluation metrics, with expanded detail in Appendix B.

4.1 Models

Five self-supervised models were trained with our IPCL framework, with an Alexnet-gn base architecture, and a 128-D latent space, with variations in training regimes (Appendix B.1). Twelve more self-supervised models were trained using the indexed-memory framework of Wu et al. (2018) (henceforth "Wusnets"; Appendix B.2), where we varied (1) the *base architecture* (Alexnet-bn, Alexnet-gn, Resnet18), including a biologically-based model architecture (Cornet-z; Kubilius et al., 2018), and (2) the *dimensionality of the latent space* ($d = 128, 256$, and 1000). The 1000-d latent space was selected to be more clearly matched to the category-supervised supervised models. We also trained a Resnet50 base architecture with a 128-d latent space. Categorization accuracy was assessed in these models using the weighted k-nearest neighbors (kNN) procedure used by Wu et al. (2018), see Appendix B.3.

For the critical comparisons, five category-supervised models were trained with matched Alexnet-bn, Alexnet-gn, Resnet18, Resnet50, and Cornet-z base architectures, with the standard output layer and cross entropy loss function using the same augmentation regime as their self-supervised counterparts.

All models were trained using the ImageNet database (Russakovsky et al., 2015), using the same augmentation policy as in Wu et al. (2018): images were randomly cropped (between $0.2 - 1.0x$ their original area, $\frac{3}{4}$ and $\frac{4}{3}$ their original aspect ratio), rescaled to 224×224 pixels, randomly horizontally flipped ($p = .5$), with random adjustments to hue, saturation, contrast, and brightness, and random grayscale conversion ($p = .2$).

4.2 Evaluating Similarity with Human Brain Responses

We used a new brain dataset from Magri and Konkle (2019), which used functional magnetic resonance imaging to obtain human brain responses to 72 individual images, depicting isolated inanimate objects (Appendix C.1; images shown in Supplementary Figure 1).

To compare the representations evident across the ventral visual pathway to those learned by the model layers, we used a standard representational similarity analysis (RSA; Kriegeskorte et al., 2008a), Appendix C.2). To overview, the visual system was divided into three large-scale sectors reflecting early cortical stages (areas V1-V3), intermediate processing stages (posterior occipitotemporal cortex, pOTC), and later processing stages (anterior occipitotemporal cortex, aOTC). In each brain region, a 72x72 representational similarity matrix was obtained, which reflects the pairwise similarity of the activation patterns across the voxels, computed as the pearson correlation between two images' activation profiles across voxels. Correspondingly, for each layer in each trained model, we measured activations in every unit to the same 72 images, and created layer-wise representational similarity matrices. Finally, the correlation between the model-layer representational similarity matrix and the brain sector's representational similarity was computed, as the key outcome measure. The reliability of each brain region was also computed, where this noise ceiling serves as an estimate of the best possible model fit.

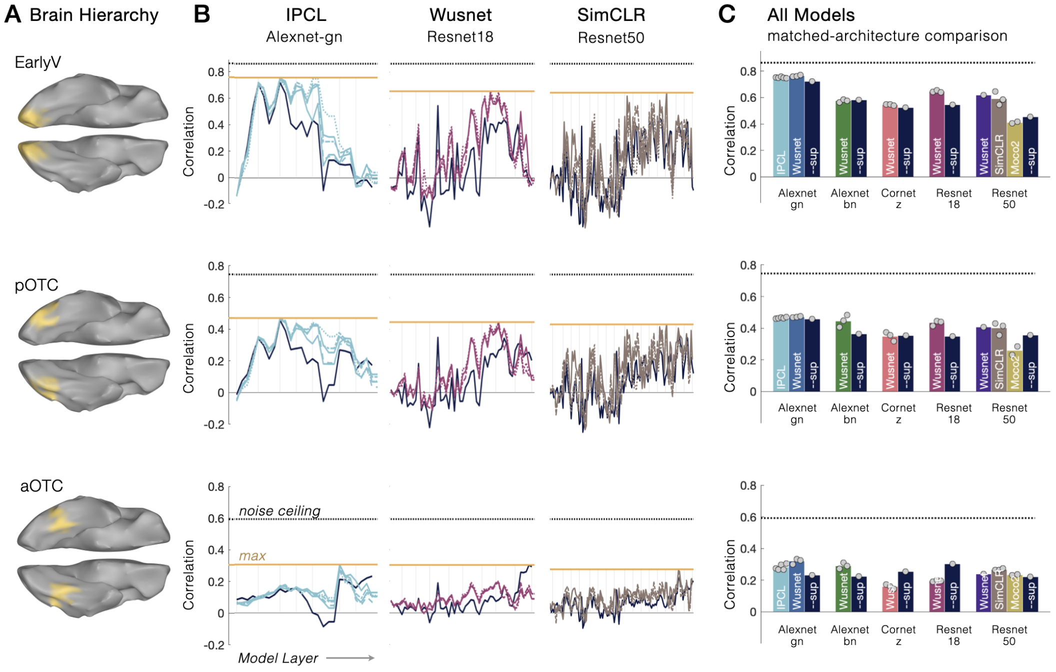

Figure 2: (A) Early, intermediate and late stages of the ventral visual stream hierarchy are depicted along the rows, shown in yellow on inflated cortical hemispheres. (B) Layerwise correlations with each brain sector. Colored lines are self-supervised, dark lines are supervised. Multiple colored lines indicate replications (IPCL), latent dimension variation (Wusnet), and Resnet50 width (1x, 2x, 4x; SimCLR). (C) Summary of layer with maximum correlation for all models tested. Bars indicate the average over replicates for that model framework, with individual models plotted in white dots, grouped by base encoding architecture.

Note that more complex RSA procedures can be used to improve fits between models and brains (e.g. see Storrs et al., 2020). Here we chose to use simple correlation between brain and model-layer representational geometries because it requires a fully emergent relationship, and thus is a more conservative bar (no brain-based fine-tuning or layer-feature-reweighting involved; e.g. Federer et al. 2019; Khaligh-Razavi and Kriegeskorte 2014). These other analytic techniques could be explored in future work.

This fMRI dataset has two distinctive features. First, the data have substantially more reliable image-level responses than other current datasets (e.g. Chang et al., 2019; Cichy et al., 2016; King et al., 2019), due to different experimental design choices. Second, the dataset reflects responses only from inanimate objects, which provides a different lens into neural representation than most other datasets that sample both animate and inanimate categories. The distinction between animates and inanimates is one of the strongest representational divisions for biological visual systems (e.g. Kriegeskorte et al., 2008b), evident in large-scale topographic organization (e.g. Grill-Spector and Weiner, 2014; Konkle and Caramazza, 2013), and also captured in the representations learned by category-supervised deep neural networks (e.g. Long et al. 2018, though see Bracci et al. 2019)–thus, capturing similarity relationship among *only* inanimate items is a finer-grained representational challenge.

5 Results

The results are shown in Figure 2. Early, intermediate, and later stage brain regions are plotted along the rows. Layer-by-layer correlations with each brain sector are shown for three selected models: IPCL-Alexnet-gn, Wusnet-Resnet18, and SimCLR-Resnet50 (see Appendix D, Supplementary Figure 2, for the other models). Finally, a summary of all models' most correlated layer is shown, grouped to highlight the self-supervised vs category-supervised comparison. There are several patterns in the data to highlight.

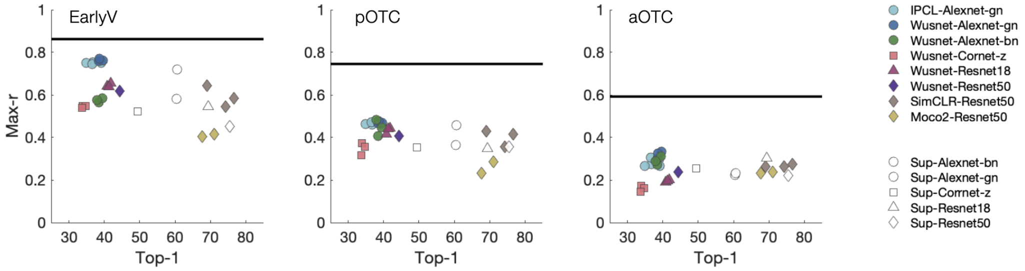

Figure 3: Relationship between Top-1 object categorization accuracy and max brain correlation, plotted for early, intermediate, and late stages of the visual system hierarchy across subplots. Colored markers are self-supervised models, open markers are supervised models, and each point is an individual model.

Self-supervised vs Category-supervised Comparison. The first key result is that the IPCL self-supervised model shows the highest general correspondence with the visual system hierarchy, along with the Wusnet model using the same Alexnet group norm architecture. Further, in nearly all cases, the self-supervised models showed comparable or higher correlation than category-supervised models with the same architecture (Figure 2c, colored vs dark bars). Interestingly, the main cases where a category-supervised model showed a stronger correlation than the self-supervised counterparts were the final model layers in some architectures (Resnet 18, Cornet-z) in the later stage brain sector (aOTC), hinting at a more category-like shift in the representational structure of this region. However, it is notable that, for all three brain sectors including the most anterior stage, the highest layer-brain correlation was a self-supervised (IPCL or Wusnet) Alexnet. These results provide the first empirical demonstration to our knowledge that self-supervised models can perform as well or better than category-supervised models in the degree to which the model layer representations directly correlate with brain representation.

Base Architecture Effects. The second result is that the convolutional neural network architecture is the single factor that matters most for brain fits in this dataset, surprisingly more than either the learning framework, the dimensionality of the projection head, or the emergent categorization accuracy. For example, Wusnet training over an Alexnet-gn base encoder produced more brain-like representations than the same training on an Alexnet-bn, Resnet18, Resnet50, or Cornet-z architecture (Figure 2c). In this dataset, the best matching model in each sector was always from the Alexnet model family. Further, the variations in the dimensionality of the latent space from 128, 256, to 1000 had a negligible effect on the layer-brain correlations (Figure 2b; dotted, dashed, solid lines; Figure 2c; open circles; see Appendix D.2). Finally, as evident in Figure 3, emergent top-1 classification accuracy was not predictive of brain similarity—indeed, our model variations clearly cluster by CNN backbone architecture. This effect was known for category-supervised systems (e.g. Kubilius et al., 2018, 2019; Schrimpf et al., 2018), and here we extend that result to self-supervised systems as well.

Hiearchical Representation. The third observation is related to the hierarchical brain stages of the human visual system and model architectures. Based on initial studies comparing internal CNN representations to the human brain, where we expect early model layers to fit early brain regions best, and later model layers to fit later brain regions best (e.g. Güçlü and van Gerven, 2015). Somewhat surprisingly, this correspondence between model layer hierarchy and brain hierarchy is not particularly clear in these data. The Alexnet family architectures show slightly clearer hierarchical correspondence than Resnet18 or Resnet50 architectures, with convolutional layers vs fully-connected layers correlating better with the earlier vs later brain regions, respectively. Further, the category-supervised models show low, or even negative, correlations at late convolutional stages. This pattern of data is somewhat contradictory to previous results (e.g. Long et al., 2018) that used the original split-channel Alexnet architecture (Krizhevsky et al., 2012). These negative correlations were also not evident when we tested the original Alexnet on our dataset (see Appendix D). These results generally support the conclusion that internal layer normalization and channeling (group-norm vs batch norm vs channel norm, splitting channels into groups) have substantial impact on the learned representation and how brain-like they are.

6 Discussion

The present work introduced Instance-Prototype Contrastive Learning, which is a fully self-supervised framework that takes the framework of Wu et al. (2018) in more biologically plausible directions, and shows emergent representations that match human brain responses equal to, or better than, category-supervised counterparts. Here we discuss the algorithmic choices of the IPCL through a biological lens, and the implications of brain-model comparisons for both neuroscience and machine vision communities.

IPCL. The most distinctive concept for self-supervised learning introduced by IPCL is the *instance prototype* based on multiple-augmentations, which can be mapped to the idea of constructing an online prototype of the current scene over multiple fixations. This learning scheme pushes each sample towards the central tendency of the instance representation across variation, effectively learning to encode each view with respect to a prototype. This encoding dovetails with classic prototype theory of object representation proposed by Rosch and Lloyd (1978), wherein category representation is not about necessary and sufficient features per se, but is probabilistic, with each exemplar standing in a more central or distant relationship with other exemplars of the category. In IPCL, we apply this logic at the instance level. This instance-prototype concept invites clear and interpretable variations, e.g. increasing the number of samples, and the kind of augmentation (e.g. simulating eye-movements, and optionally including biologically inspired “efference copy” signals that indicate the magnitude and direction of eye-movements between samples; e.g. Colby et al. 1992; Crapse and Sommer 2008).

The use of a non-indexed memory queue in IPCL also has biological undertones: the human and non-human primate ventral streams are effectively a highway to the hippocampus (Van Essen and Maunsell, 1983), a brain structure supporting long-term memory where more compositional-like operations can be carried out over usefully factorized visual representations. This suggests that to some extent the representations in the ventral visual stream maybe be optimized to interface with long-term memory systems. Through this lens, the recent memory queue of IPCL is a stand-in for the traces that would be accessible in a hippocampal memory system. Further extensions into the biological realm might draw on hippocampal models of memory (e.g. Schapiro et al., 2017). For example, our memory queue has a temporal order but no temporal decay, unlike biological long-term memory signatures (e.g. see Anderson and Schooler, 1991), inviting further modifications that vary the weight of the contrast with fading negative samples.

Brain-Model Comparisons. In this interdisciplinary intersection between deep learning and neuroscience, comparing the representations of different model layers to different brain sectors can be done towards (at least) two distinct ends. One aim is to find the single best model system with the most emergent brain-like representation. Brain-Score formalizes this endeavor, aggregating brain datasets and automating pipelines for scoring models along these brain-based and behavior-based benchmarks (Schrimpf et al., 2018). This fMRI dataset has clear value towards this endeavor, as there is reliable neural representation in mid- and late stages of the visual system that is not accounted for well by any model, whether category-supervised or not, and across models that vary substantially in object categorization accuracy. These results contribute the emerging picture that *object categorization* is not the right task to close this representational gap between models and brains (Schrimpf et al., 2018). Given that Magri and Konkle (2019) found that aOTC representation is well correlated with human judgments of 3-dimensional shape similarity, models that must learn more 3D-aware representations present a provocative alternative to categorization (e.g. Tung et al., 2019; Zamir et al., 2016).

However, finding the best model is only one reason to compare models to brain data. For the cognitive neuroscientist, these models also serve as computational existence proofs for learnability arguments: that is, what kind of representational structure can be learned from natural image inputs and architectural constraints, given specific representational goals (operationalized as loss functions; e.g. Richards et al. 2019). For example, a prominent theory of object representation in the visual system asserts that specialized category-level (or "domain-level") forces are critical for shaping visual category representation (Mahon and Caramazza e.g. 2011, see also de Beeck et al. 2019). The finding that instance-level contrastive learning can result in emergent categorical representation supports an alternative theoretical point of view, in which category-specialized learning mechanisms are not necessary. On this generalist account, visual mechanisms operate similarly over all kinds of input, and the goal is to learn hierarchical visual features that simply try to discriminate each view from every other view of the world, regardless of the visual content or domain. Here, we add that these instance-level contrastive learning systems can have representations that are as brain-like as category-supervised systems, increasing the plausibility of the generalist account.

Broader Impact

This work is aimed at advancing self-supervised vision systems as well as our understanding the nature of human visual representation. Both scientific communities stand to benefit. Biological systems can be used to help inspire and inform machine vision model development. Models can inform cognitive neuroscience theories, serving as computational existence proofs for learnability arguments, and as testbeds for exploring the links between targeted mechanisms and their representational consequences. Beyond basic science understanding, the work presented in this paper does not raise many ethical issues, or have consequences have impact at a societal level.

Our paper is part of the larger endeavor of building more human-like models, focusing specifically on the nature of visual representation. Ultimately, a successful model will emulate human perception, providing more flexible and robust machine vision systems and providing insight into the nature of human visual processing. To strive towards these goals, our work leverages a large-scale image database—any sampling bias in this dataset thus will permeate the representational structure learned by the model systems. Further, the goal of making brain-like representation is also worth examining, as human perceptual systems are not without bias (e.g. consider the *other-race effect*). To address these limitations, it may be possible to develop targeted stimulus sets which can be used in to assess both human perception and machine vision systems, which can be used to quantify bias effects (e.g. a visual *implicit attitudes test*, see Greenwald et al. 2003). This endeavor is particularly important for machine vision of people, actions, and interactions; it is less relevant for the current work focusing on isolated inanimate objects.

Acknowledgments and Disclosure of Funding

This research was supported by an AWS Machine Learning Research Award.

References

John R Anderson and Lael J Schooler. Reflections of the environment in memory. *Psychological science*, 2(6):396–408, 1991.

Stefania Bracci, J Brendan Ritchie, Ioannis Kalfas, and Hans P Op de Beeck. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, 39(33):6513–6525, 2019.

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol*, 10(12):e1003963, 2014.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8):3563–3579, 2016.

CL Colby, ME Goldberg, et al. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255(5040):90–92, 1992.

Trinity B Crapse and Marc A Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, 2008.

Hans P Op de Beeck, Ineke Pillet, and J Brendan Ritchie. Factors determining where category-selective areas emerge in visual cortex. *Trends in cognitive sciences*, 2019.

Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

Callie Federer, Haoyan Xu, Alona Fyshe, and Joel Zylberberg. Training neural networks to mimic the brain improves object recognition performance, 2019.

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. *arXiv preprint arXiv:1811.12231*, 2018.

Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197, 2003.

Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

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

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.

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

Marcie L King, Iris IA Groen, Adam Steel, Dwight J Kravitz, and Chris I Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, 2019.

Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008a.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in Neural Information Processing Systems*, pages 12785–12796, 2019.

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

Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

Caterina Magri and Talia Konkle. Comparing facets of behavioral object representation: implicit perceptual similarity matches brains and models. In *Cognitive Computational Neuroscience*, 2019.

Bradford Z Mahon and Alfonso Caramazza. What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, 15(3):97–103, 2011.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*, 2019.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

Eleanor Rosch and Barbara Bloom Lloyd. *Cognition and categorization*. Lawrence Erlbaum Associates Hillsdale, NJ, 1978.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Anna C Schapiro, Nicholas B Turk-Browne, Matthew M Botvinick, and Kenneth A Norman. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160049, 2017.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5: 399–426, 2019.

Fabian H Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, 2019.

Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human it well, after training and fitting. *bioRxiv*, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2595–2603, 2019.

David C Van Essen and John HR Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6:370–375, 1983.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

Yaoda Xu and Maryam Vaziri-Pashkam. Limited correspondence in visual representation between the human brain and convolutional neural networks. *bioRxiv*, 2020.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019.

Amir R Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3d representation via pose estimation and matching. In *European Conference on Computer Vision*, pages 535–553. Springer, 2016.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6002–6012, 2019.

Appendix

A. Model architecture Details

We used the AlexNet architecture as specified in Krizhevsky et al. (2012), with a few changes. Layers were not divided into groups (in the original AlexNet early layers were divided into groups that were split across GPUs and then merged at later layers, and some implementations maintain this grouping, e.g., MATLAB). AlexNet (bn) used BatchNorm layers (Ioffe and Szegedy, 2015) after each convolutional and fully-connected layer (except the final fully-connected layer) with $eps = 1e-05$ added to the numerator for numerical stability and momentum=0.1 for the running mean and variance. In our AlexNet (gn) we used GroupNorm layers (Wu and He, 2018) after each convolutional layer (using 32 groups; $eps = 1e-5$, and with learnable per-channel affine parameters initialized to ones for weights and zeros for biases), and BatchNorm1d layers after each fully-connected layer (except the final layer). The final 1000-way output layer was replaced with an N-dimensional latent space with an L2 norm operation (following Wu et al. 2018). The dimensionality of this space was varied for different models from 128, 256, to 1000. See Appendix Section E for the exact model architecture specification for Alexnet(bn) and Alexnet(gn).

B. Training Details

B.1 IPCL Model Training

All IPCL models used temperature τ = 0.07, a non-indexed memory queue of size 4096, and multiple augmentations per image ($N = 5$), and reduced the batch size to 128 (the reduced batch size was needed to fit 5x images on our GPUs). We found that this architecture would not learn with the standard learning rate schedule, so we made the following changes to the training protocol: (1) we accumulated gradients over 20 batches (i.e., performed the optimizer step every 20 batches), and (2) we reduced the number of epochs to 100, and (3) we varied the learning rate schedule using a one-cycle policy (Smith, 2017), using cosine annealing to vary the learning rate from $.03/1000$ to $.03$ over the first 40 epochs, and from $.03$ down to $.03/(1000 * 1e4)$. We replicated the model training three times (run1,2,3) using SGD, and once using the Ranger optimizer (RectifiedAdam + LookAhead). The resulting models achieved between $36.5\% - 39.1\%$ accuracy on subsequent ImageNet categorization, detailed in the table in Appendix Section D.

B.2 Wusnet Model Training

Each of the 12 model variants (6 architectures × 3 latent space dimensionalities) were trained with the same procedure as in Wu et al. (2018; `https://github.com/zhirongw/lemniscate.pytorch`), using temperature $\tau = 0.07$ and NCE sample size 4096. Models were trained for 200 epochs using SGD (momentum $= .9$; weight decay $= .0001$), and a batch-size of 256. The learning rate was initialized to 0.03 and scaled down by 0.1 at epoch 120 and 160.

We used Wu et al. (2018)'s implementation of Noise Contrastive Estimation to approximate sampling, with slight modifications to accommodate our prototype and queue:

$$\ell_{z_i,\bar{z}} = -(\log(Pos) + \log(Neg)) \tag{1}$$

$$Pos = \frac{\exp(\text{sim}(z_i, \bar{z})/\tau)/Z}{\exp(\text{sim}(z_i, \bar{z})/\tau)/Z + \frac{K}{N} + \epsilon} \tag{2}$$

$$Neg = \frac{\frac{K}{N}}{\sum_{k=1}^{K} [\exp(\text{sim}(z_i, q_k)/\tau)/Z + \frac{K}{N} + \epsilon]} \tag{3}$$

where z_i is the embedding for the i^{th} sample, $\bar{z}$ is its corresponding prototype, sim is the similarity function (dot-product between embeddings), τ is the temperature parameter, Z is a normalization constant (estimated based on the first mini-batch of 128*5 augmented samples), q_k is the embedding for the k^{th} item stored in the queue, and $\epsilon = 1e - 7$ is a constant added for numerical stability.

B.3 Evaluating Categorization Accuracy

We tested classification accuracy on the ImageNet validation set for the Wusnet and IPCL models, using the same weighted k-nearest neighbors (kNN) procedure used by Wu et al. (2018). To classify a test image x, it's embedding was compared to the embedding of each training image using cosine similarity. The top $k = 200$ nearest neighbors were used to make the prediction via cosine-similarity-weighted voting, where the class c would receive the total weight given by:

$$w_c = \sum_{i}^{N_k} \exp(s_i/\tau) \cdot 1(c_i = c) \tag{4}$$

Where N_k denotes the k-nearest neighbors, and s_i is the cosine similarity between the target and the neighbor, $k = 200$, and $\tau = 0.07$.

C. Brain Data and Analysis

C.1 Summary of fMRI Experimental Procedures

The stimulus set consisted of 72 images (Supplementary Figure 1), and was selected to span a range of categories and contexts (e.g. accessories, bags, bathroom items, bedroom items, clothing, food-processed, fruits and vegetables, furniture, household items, kitchen, musical instruments, office supplies, outdoor items, sporting goods, tools, vehicles).

These images were presented in a mini-block design, while participant ($N = 10$) underwent functional magnetic resonance imaging scanning. In each 8-min run, each image was flashed 4x in a row (600ms on 400ms off) in a 4s block, with all 72 images presented in a block in each run (randomly ordered), with 4x15-s rest periods interleaved throughout. Participants completed 6 runs. Their task was to pay attention to each image and complete a vigilance task (press a button when a red-frame appeared around an object, which happened 12 times in run).

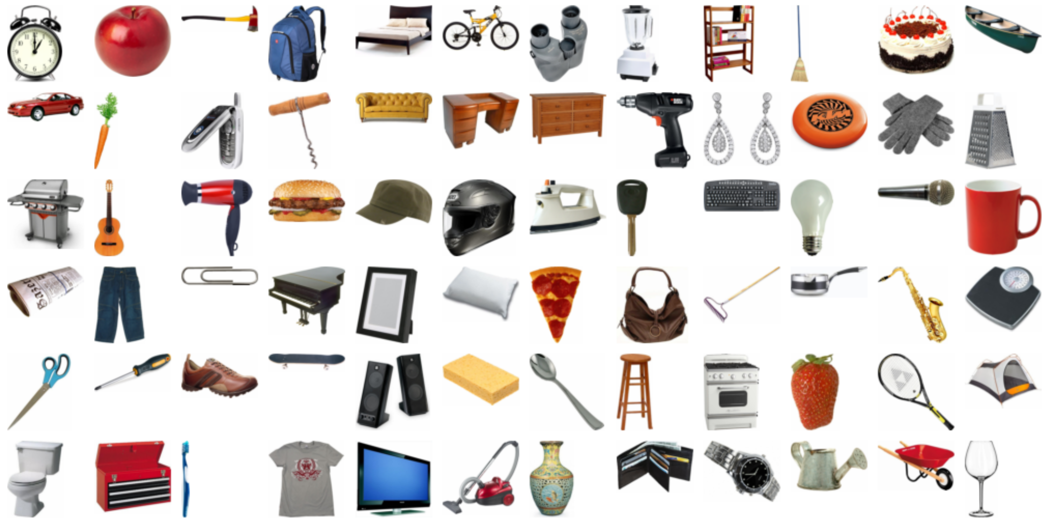

Figure 1: 72 inanimate object images used in the fMRI design.

Imaging data were acquired on a BioSpin MedSpec 4T scanner (Bruker) using an eight-channel head coil. Functional data were analyzed using Brain Voyager QX software and MATLAB. Standard preprocessing was performed and a general linear modeling framework was used to estimate the response of each voxel to each of the 72 images, using square-wave regressors for each image condition, convolved with a gamma function to approximate the hemodynamic response. Thus, this design allowed us to estimate voxel-level estimation to these 72 individual images.

All sectors were delineated by hand on the cortical surface of each hemisphere of each participant, using typical procedures(e.g. Cohen et al., 2017; Long et al., 2018). The early visual areas V1-V3 were delineated based on activations from a separate retinotopy protocol, using standard procedures. An occipitotemporal cortex mask was drawn on each hemisphere, within which the 1000-most active voxels were included, based on the contrast [All Objects > Rest] at the group-level. To divide this cortex into intermediate and later hierarchical stages, we used the analyses of these data done by Magri and Konkle (2019): in this dataset, the brain responses have a systematic dip in local-regional reliability at a particular anatomical location along the posterior to the anterior axis.

C.2. Representational Similarity Analysis

For each brain region and each participant, we computed a 72×72 representational similarity matrix (RSM), using the Pearson correlation between activation profiles between all pairs of images. A group-level RSM was created for each sector by averaging across individual participants. The lower diagonal of this symmetrical matrix was vectorized (representational similarity vector, RDV; 2556 pairs) and used as the target brain data to match. For each layer in each trained model, we presented the same images to the model, recorded activations from every unit in every layer, and computed a 72×72 RSM with the same method. The lower diagonal of these matrices were vectorized and correlated with the brain target RDV.

To compute the reliability of this RDV to contextualize the model fits, we split participants into two sets, computed the RDV in both sets, and repeated this procedure 1000 times, to estimate an average split-half reliability of the brain data.

D. Supplemental Results

D.1. Other trained model plots

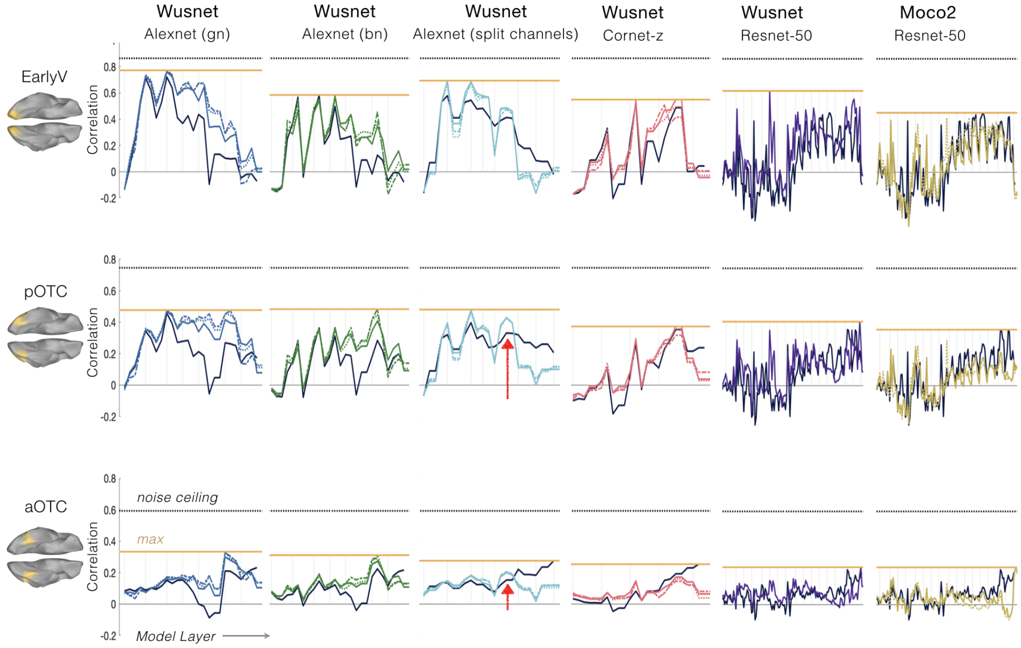

Figure 2: Layer-wise correlations for additional models (columns) for each brain region (rows). The noise ceiling of the brain data is indicated in dashed black; the peak correlation across layers is indicated with the gold horizontal line.

D.2. Summary Table

Table 1: Summary of Model Top-1 and Max-r

Framework	Architecture	Dim	Top-1	EarlyV	pOTC	aOTC
IPCL	Alexnet-gn	128	36.7	0.75	0.46	0.27
IPCL	Alexnet-gn	128	34.9	0.75	0.46	0.27
IPCL	Alexnet-gn	128	38.9	0.76	0.47	0.30
IPCL	Alexnet-gn	128	39.2	0.75	0.46	0.27
IPCL	Alexnet-gn	128	36.5	0.74	0.47	0.31
Wusnet	Alexnet-gn	128	38.4	0.76	0.47	0.30
Wusnet	Alexnet-gn	256	39.6	0.76	0.47	**0.33**
Wusnet	Alexnet-gn	1000	38.6	**0.77**	**0.48**	**0.33**
–Supervised	Alexnet-gn	-	**60.6**	0.72	0.46	0.23
Wusnet	Alexnet-bn	128	38.5	0.56	0.41	0.27
Wusnet	Alexnet-bn	256	39.4	0.58	0.45	0.31
Wusnet	Alexnet-bn	1000	37.9	0.58	**0.48**	0.29
–Supervised	Alexnet-bn	-	**60.4**	0.58	0.36	0.22
Wusnet	Cornet-z-bn	128	33.9	0.55	0.37	0.17
Wusnet	Cornet-z-bn	256	34.7	0.55	0.36	0.16
Wusnet	Cornet-z-bn	1000	33.7	0.54	0.32	0.15
–Supervised	Cornet-z-bn	-	**49.3**	0.52	0.35	0.25
Wusnet	Resnet18	128	40.8	0.64	0.42	0.19
Wusnet	Resnet18	256	41.9	0.65	0.44	0.20
Wusnet	Resnet18	1000	41.3	0.64	0.44	0.20
–Supervised	Resnet18	-	**69.4**	0.55	0.35	0.30
Wusnet	Resnet50	128	44.3	0.62	0.41	0.24
–Supervised	Resnet50	-	**75.4**	0.45	0.36	0.22
SimCLR	Resnet50x1	128	69.1	0.64	0.43	0.26
SimCLR	Resnet50x2	128	74.2	0.55	0.36	0.26
SimCLR	Resnet50x4	128	**76.6**	0.59	0.42	0.28
–Supervised	Resnet50	-	75.4	0.45	0.36	0.22
Moco2	Resnet50-ep200	128	67.7	0.40	0.23	0.23
Moco2	Resnet50-ep800	128	71.1	0.42	0.29	0.24
–Supervised	rResnet50	-	**75.4**	0.45	0.36	0.22

E. Architecture Specification

```
AlexNetBN(
  (conv_block_1): Sequential(
    (0): Conv2d(3, 96, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2), bias=False)
    (1): BatchNorm2d(96, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
    (3): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv_block_2): Sequential(
    (0): Conv2d(96, 256, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2), bias=False)
    (1): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
    (3): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv_block_3): Sequential(
    (0): Conv2d(256, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
    (1): BatchNorm2d(384, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
  )
  (conv_block_4): Sequential(
    (0): Conv2d(384, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
    (1): BatchNorm2d(384, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
  )
  (conv_block_5): Sequential(
    (0): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
    (1): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
    (3): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (ave_pool): AdaptiveAvgPool2d(output_size=(6, 6))
  (fc6): Sequential(
    (0): Linear(in_features=9216, out_features=4096, bias=True)
    (1): BatchNorm1d(4096, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
  )
  (fc7): Sequential(
    (0): Linear(in_features=4096, out_features=4096, bias=True)
    (1): BatchNorm1d(4096, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
  )
  (fc8): Sequential(
    (0): Linear(in_features=4096, out_features=128, bias=True)
  )
  (l2norm): Normalize()
)
```

```
AlexNetGN(
  (conv_block_1): Sequential(
    (0): Conv2d(3, 96, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2), bias=False)
    (1): GroupNorm(32, 96, eps=1e-05, affine=True)
    (2): ReLU(inplace=True)
    (3): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv_block_2): Sequential(
    (0): Conv2d(96, 256, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2), bias=False)
    (1): GroupNorm(32, 256, eps=1e-05, affine=True)
    (2): ReLU(inplace=True)
    (3): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv_block_3): Sequential(
    (0): Conv2d(256, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
    (1): GroupNorm(32, 384, eps=1e-05, affine=True)
    (2): ReLU(inplace=True)
  )
  (conv_block_4): Sequential(
    (0): Conv2d(384, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
    (1): GroupNorm(32, 384, eps=1e-05, affine=True)
    (2): ReLU(inplace=True)
  )
  (conv_block_5): Sequential(
    (0): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
    (1): GroupNorm(32, 256, eps=1e-05, affine=True)
    (2): ReLU(inplace=True)
    (3): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (ave_pool): AdaptiveAvgPool2d(output_size=(6, 6))
  (fc6): Sequential(
    (0): Linear(in_features=9216, out_features=4096, bias=True)
    (1): BatchNorm1d(4096, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
  )
  (fc7): Sequential(
    (0): Linear(in_features=4096, out_features=4096, bias=True)
    (1): BatchNorm1d(4096, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
  )
  (fc8): Sequential(
    (0): Linear(in_features=4096, out_features=128, bias=True)
  )
  (l2norm): Normalize()
)
```

References

Michael A Cohen, George A Alvarez, Ken Nakayama, and Talia Konkle. Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of neurophysiology*, 117(1):388–402, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

Caterina Magri and Talia Konkle. Comparing facets of behavioral object representation: implicit perceptual similarity matches brains and models. In *Cognitive Computational Neuroscience*, 2019.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.