Human alignment of neural network representations

Lukas Muttenthaler[*]
Machine Learning Group
Technische Universität Berlin
BIFOLD[†]

Lorenz Linhardt
Machine Learning Group
Technische Universität Berlin
BIFOLD[†]

Jonas Dippel
Machine Learning Group
Technische Universität Berlin
BIFOLD[†]

Robert A. Vandermeulen
Machine Learning Group
Technische Universität Berlin
BIFOLD[†]

Simon Kornblith
Google Brain, Toronto

Abstract

Today's computer vision models achieve human or near-human level performance across a wide variety of vision tasks. However, their architectures, data, and learning algorithms differ in numerous ways from those that give rise to human vision. In this paper, we investigate the factors that affect alignment between the representations learned by neural networks and human concept representations. Human representations are inferred from behavioral responses in an odd-one-out triplet task, where humans were presented with three images and had to select the odd-one-out. We find that model scale and architecture have essentially no effect on alignment with human behavioral responses, whereas the training dataset and objective function have a much larger impact. Using a sparse Bayesian model of human conceptual representations, we partition triplets by the concept that distinguishes the two similar images from the odd-one-out, finding that some concepts such as food and animals are well-represented in neural network representations whereas others such as royal or sports-related objects are not. Overall, although models trained on larger, more diverse datasets achieve better alignment with humans than models trained on ImageNet alone, our results indicate that scaling alone is unlikely to be sufficient to train neural networks with conceptual representations that match those used by humans.

1 Introduction

Representation learning is a fundamental part of modern computer vision systems, but the paradigm has its roots in cognitive science. When Rumelhart et al. [57] developed backpropagation, their goal was to find a method that could learn representations of concepts that are distributed across neurons, similarly to the human brain. The discovery that representations learned by backpropagation could replicate nontrivial aspects of human concept learning was a key factor in its rise to popularity in the late 1980s [65, 45]. A string of empirical successes has since shifted the primary focus of representation learning research away from its similarities to human cognition and toward practical applications. This shift has been fruitful. By some metrics, the best computer vision models now outperform the best individual humans on benchmarks such as ImageNet [60, 8, 69]. However, the extent to which the conceptual representations learned by these high-performing vision models align with those used by humans remains unclear.

[*] Also affiliated with the Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.
[†] BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at the Neural Information Processing Systems (NeurIPS) conference 2022. New Orleans.

Do models that are better at classifying images naturally learn more human-like conceptual representations? Prior work has investigated this question indirectly, by measuring models' error consistency with humans [18, 52, 21] and the ability of their representations to predict neural activity in primate brains [71, 23, 59], with mixed results. Here, we approach the question of alignment between human and machine representation spaces more directly, using human similarity judgments collected from an odd-one-out task, where humans saw triplets of images and selected the image most different from the other two [28]. These similarity judgments allow us to infer that the two images that were not selected are closer to each other in an individual's concept space than either is to the odd-one-out. We define the odd-one-out in the neural network representation space analogously, and measure neural networks' alignment with human similarity judgments in terms of their *odd-one-out accuracy*, i.e., the accuracy of their odd-one-out "judgments" with respect to humans', under a wide variety of settings. Based on these odd-one-out accuracies, we draw the following conclusions:

- Scaling ImageNet models improves ImageNet accuracy, but does not consistently improve alignment of their representations with human similarity judgments. Differences in alignment across ImageNet models appear to arise primarily from differences in objective functions rather than from differences in architecture or width/depth.
- Models trained on image/text data, or on larger, more diverse classification datasets than ImageNet, achieve substantially better alignment with humans.
- We use a sparse Bayesian model of human mental representations [44] to partition triplets by the concept that distinguishes the odd-one-out. While food and animal-related concepts can easily be recovered from neural net representations, human alignment is weak for dimensions that depict sports-related or royal objects, especially for ImageNet models.

We discuss related work more thoroughly in Appendix A.

2 Methods

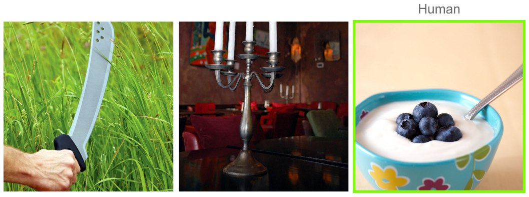

Figure 1: An example triplet from Hebart et al. [28], where neural nets choose a different odd-one-out than a human. The images in this triplet are copyright-free images from THINGS + [64].

Data The images used in this paper are taken from the THINGS database [27]. THINGS consists of a collection of 1,854 object categories, i.e., concrete nameable nouns in the English language, along with representative images for these categories. THINGS was curated to include categories that can be easily identified as a central object in a natural image. Hebart et al. [28] collected similarity judgments from human participants on categories in THINGS, which they then used to derive concept representations. These similarity judgments came in the form of responses to a *triplet task*. In a triplet task, images from three distinct categories are presented to a participant, from which the participant selects the image that is most different from the other two (or equivalently the pair of images that are most similar). The authors collected 1.46 million unique responses crowdsourced from 5,301 workers. See Figure 1 for an example triplet. For presentation purposes, we have replaced the images used in Hebart et al. [28] with images similar in appearance that are licensed under CC0 [64].

Models In our evaluation, we consider a diverse set of pretrained neural networks, including a wide variety of self-supervised and supervised models trained on ImageNet-1K and ImageNet-21K [13]; a Vision Transformer trained on the proprietary JFT-3B dataset [73]; and models that were trained on both image and text data such as CLIP [51], ALIGN [32], and BASIC [49]. See Table C.1 for a comprehensive list of all models. In our plots, we determine the ImageNet top-1 accuracy for networks not trained on ImageNet-1K by training a linear classifier on the network's penultimate layer using L-BFGS [41].

Zero-shot odd-one-out accuracy We examine the extent to which the odd-one-out can be identified directly from the similarities between images in models' representation spaces. Given representations $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ of the three images that make up the triplet, we first construct a similarity matrix $\boldsymbol{S} \in \mathbb{R}^{3\times 3}$ where $S_{i,j} := \boldsymbol{x}_i^T \boldsymbol{x}_j / (\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2)$, the cosine similarity between a pair of representations. We identify the closest pair of images in the triplet as $\arg\max_{i,j>i} S_{i,j}$; the remaining image is the odd-one-out. We define zero-shot odd-one-out accuracy as the proportion of triplets where the odd-one-out identified in this fashion matches the human odd-one-out response. When evaluating zero-shot odd-one-out accuracy, we report the better of the accuracies obtained from representations of the penultimate embedding layer and logits (if the network has a logits layer). As we show in Figure C.1, representations obtained from earlier network layers perform worse.

Probing In cases where a model's zero-shot accuracy is low, decoding the information necessary for downstream tasks may only require a linear transformation. Generally, linear probing yields insights into the information encoded in neural net's representation [61, 2].

To perform linear probing, we formulate the notion of the odd-one-out probabilistically, as in Hebart et al. [28]. Given similarity matrix $\boldsymbol{S}$ and a triplet $\{i, j, k\}$ (here the images are indexed by natural numbers), the likelihood of a particular pair, $\{a, b\} \subset \{i, j, k\}$, being most similar, and thus the remaining image being the odd-one-out, is modeled by the softmax of the image similarities,

$$p(\{a, b\}|\{i, j, k\}, \boldsymbol{S}) \coloneqq \exp(S_{a,b}) / \left(\exp(S_{i,j}) + \exp(S_{i,k}) + \exp(S_{j,k})\right). \tag{1}$$

We learn the linear transformation that maximizes the log-likelihood of the triplet odd-one-out judgments plus an ℓ_2 regularization term. Specifically, given triplet responses $(\{a_s, b_s\}, \{i_s, j_s, k_s\})_{s=1}^{n}$ we find a square matrix $\boldsymbol{W}$ yielding a similarity matrix $S_{ij} = (\boldsymbol{W}\boldsymbol{x}_i)^T(\boldsymbol{W}\boldsymbol{x}_j)$ that optimizes

$$\underset{\boldsymbol{W}}{\arg\min} \; -\frac{1}{n}\sum_{s=1}^{n} \log p\left(\{a_s, b_s\}|\{i_s, j_s, k_s\}, \boldsymbol{S}\right) + \lambda||\boldsymbol{W}||_2^2. \tag{2}$$

Here, we determine λ via grid-search during k-fold cross-validation (CV). To obtain a minimally biased estimate of the odd-one-out accuracy of a linear probe, we partition the m objects into two disjoint sets. Experimental details about the optimization process, k-fold CV, and how we partition the objects can be found in Appendix B.1 and in Algorithm 1 respectively.

VICE Several of our analyses make use of human concept representations obtained by Variational Interpretable Concept Embeddings (VICE), an approximate Bayesian method for finding concept representations from human odd-one-out responses in a triplet task [44]. VICE uses mean-field VI to yield a sparse representation for each image that best explains these responses. VICE does not consider the content of the images and cannot provide representations for novel images. VICE shows high reproducibility of representations across different random initializations, and has strong predictive power, achieving an odd-one-out accuracy of $\sim 64\%$ on THINGS, which is only marginally lower than the estimated ceiling accuracy of 67.22% [28].

3 Experiments

Here, we investigate how closely neural networks' representation spaces align with humans' concept spaces, and whether concepts can be recovered from a representation via a linear transformation.

3.1 Odd-one-out vs. ImageNet accuracy

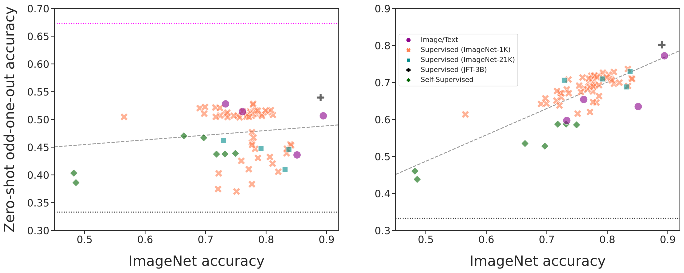

Figure 2: Zero-shot odd-one-out accuracy as a function of ImageNet accuracy for THINGS (*left*) and CIFAR-100 coarse (*right*). Dashed diagonal lines indicate a least-squares fit. Dashed horizontal lines reflect chance-level or ceiling accuracy respectively.

We begin by comparing zero-shot odd-one-out accuracy for THINGS with ImageNet accuracy for all models in Table C.1. ImageNet accuracy generally is a good predictor for transfer learning performance [35, 14, 17]. Thus, we evaluate zero-shot odd-one-out accuracy for all models in Table C.1 and compare it with their ImageNet top-1 accuracy. Figure 2 shows results for both the THINGS triplet task (left) as well as a triplet task constructed using the 20 coarse classes of the CIFAR-100 dataset (right). To generate CIFAR-100 triplets, we select two images from the same coarse class and one odd-one-out image from a different class; see Appendix D for further details. While ImageNet accuracy is highly correlated with odd-one-out accuracy for the CIFAR-100 coarse task ($r = 0.809$), its correlation with

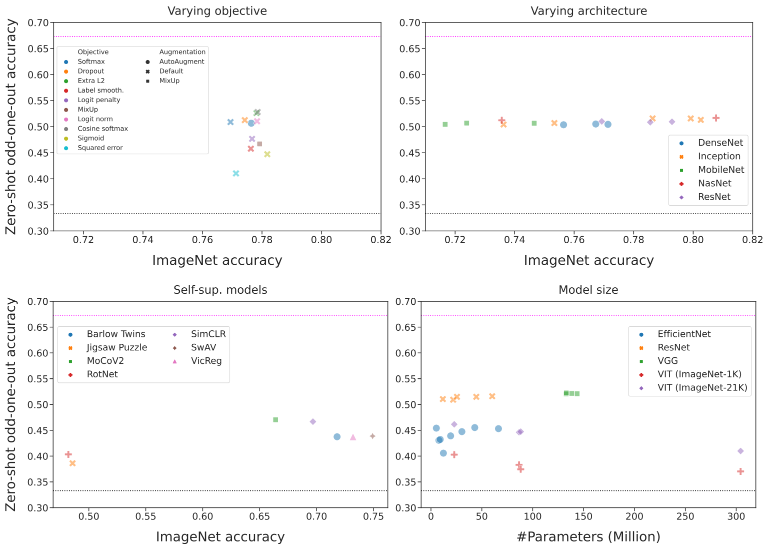

Figure 3: Zero-shot odd-one-out accuracy as a function of ImageNet accuracy or number of model parameters. **Top**: Models on the left have the same architecture (ResNet-50) but were trained with a different objective function or different data augmentation. Models on the right were trained with the same objective function but vary in architecture. **Bottom**: Performance for different SSL models on the left, and a subset of ImageNet models with their number of parameters on the right. Dashed horizontal lines reflect chance-level or ceiling accuracy respectively.

accuracy on human odd-one-out judgments is much weaker ($r = 0.131$). This raises the question whether there are model, task, or data characteristics that influence human alignment.

Architecture or objective? The top row of Figure 3 shows odd-one-out accuracy as a function of ImageNet performance for models from two recent studies that investigated the transferability of ImageNet pretrained representations that vary only in the architecture or training objective, with all other hyperparameters fixed [35, 36]. We find that models with the same architecture (ResNet-50) trained with different data augmentation or objective functions [36] yield substantially different zero-shot odd-one-out accuracies. Conversely, models with different architectures trained with the same objective function [35] - softmax cross-entropy -, achieve similar odd-one-out accuracies, although their ImageNet accuracies vary significantly. This suggests that architecture does not affect odd-one-out accuracy, while the objective function and the augmentation strategy have a significant impact.

Self-supervised learning The plot in the bottom left corner of Figure 3 compares zero-shot odd-one-out accuracy of different SSL models with their linear probing ImageNet performance. The non-Siamese models Jigsaw [46] and RotNet [22] show substantially worse alignment with human judgments than other SSL models. This is not surprising given their poor performance on ImageNet. For the Siamese methods SimCLR [11], MoCoV2 [26], Barlow Twins [72], SwAV [10], and VICReg [6], however, ImageNet performance does not correspond to alignment with human judgments.

Model capacity The graph in the bottom right corner of Figure 3plots zero-shot odd-one-out accuracy against the number of model parameters for a subset of ImageNet models. While one typically observes a positive correlation between model capacity and task performance in computer vision, we do not observe any relationship between model width/depth and odd-one-out accuracy.

3.2 How much alignment can a linear probe recover?

Probing and zero-shot odd-one-out accuracies are positively correlated, in the embedding (Figure 4; $r = 0.645$) and logits layer (Figure E.2; $r = 0.963$). However, there are models in Figure 4 that show poor zero-shot and strong linear probing odd-one-out accuracies, such as ALIGN, SWAV, EfficientNet B4 and ViT-L/16. ALIGN is probably the most interesting candidate. Although its zero-shot odd-one-out accuracy is average, this image/text model achieves the highest probing odd-one-out accuracy across all evaluated models.

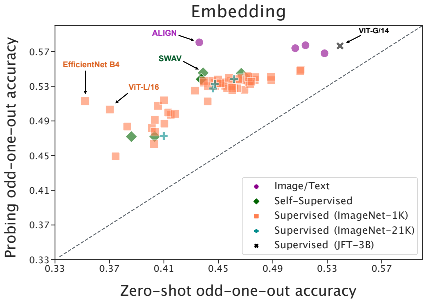

Figure 4: Zero-shot and probing odd-one-out accuracies for the embedding layer of all neural nets. Dashed line indicates $x = y$.

As we show in Appendix E, the relationship between probing odd-one-out accuracy and ImageNet accuracy is similar to the relationship between zero-shot odd-one-out accuracy and ImageNet accuracy described above. The correlation between ImageNet accuracy and probing odd-one-out accuracy is still weak ($r = 0.213$). Probing reduces the variance in odd-one-out accuracy among networks trained with different loss functions and Siamese self-supervised learning methods, but there is still no clear improvement in odd-one-out accuracy with better-performing architectures or larger model capacities.

3.3 Human alignment is concept-specific

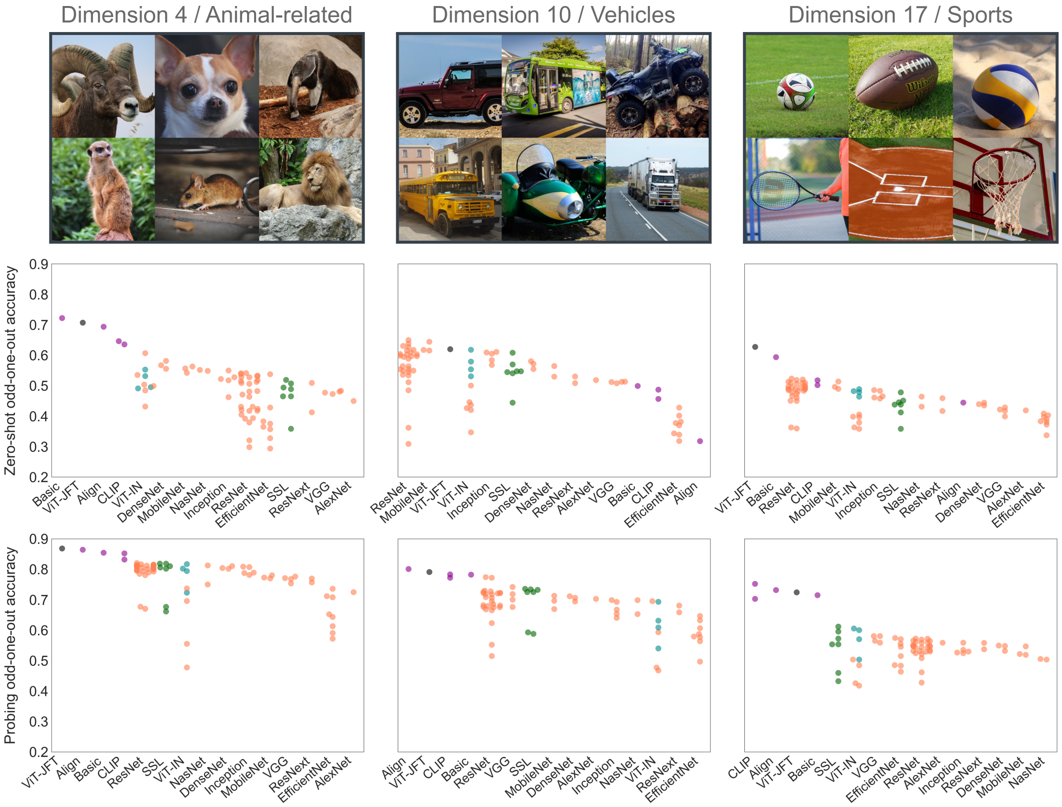

Figure 5: Zero-shot and linear probing odd-one-out accuracies for the embedding layer of all models for a subset of three of the 45 VICE dimensions. Color-coding was determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).

In the following analysis, we evaluate both zero-shot and linear probing odd-one-out accuracy for individual human concepts. We partitioned the original triplet dataset according to the VICE dimension shared between the two more similar images; see Appendix F for details. In Figure 5, we show zero-shot and linear probing odd-one-out accuracies for three VICE dimensions, for all models listed in Table C.1.

Although most image/text models and ViT-G/14 JFT showed a higher probing odd-one-out accuracy compared to self-supervised models or models trained on ImageNet, zero-shot odd-one-out accuracy was somewhat less consistent. For dimension 10, ResNets from Kornblith et al. [36], trained with a cosine softmax objective, were the best zero-shot performing models, whereas image/text models' zero-shot performance were among the worst. For dimension 4, an animal-related concept, models pretrained on ImageNet clearly showed the worst performance, whereas this concepts seems to be well represented in image/text models. After linear probing, results became less ambiguous. For almost every human concept, image/text models and ViT-G/14 JFT were the best human aligned models, whereas both AlexNet and EfficientNets achieved the lowest per-concept odd-one-out accuracies. This difference between image/text models and ViT-G/14 JFT and the other ImageNet-prerained models was particularly apparent for dimension 17 which summarizes sports-related objects. For this dimension, we observed a large performance gap after linear probing between image/text models and ViT-G/14 JFT and all remaining models. Analogously, in Appendix G we perform the same experiments using linear regression to predict representations from VICE. These experiments corroborate the results obtained from linear probing.

4 Discussion

In this work, we evaluated the alignment of neural network representation with human concept spaces through performance in an odd-one-out task. Before discussing our findings, we want to address limitations of our work. One obvious limitation is the fact that we did not consider non-linear transformations. It is possible that there exist families of simple non-linear transformations that can provide better alignment for the networks we investigate. We plan to investigate such transformations more thoroughly in future work. Another limitation relates to the use of pretrained models for our experiments. These models have been trained with a variety of objectives and regularization strategies. We have mitigated this limitation by comparing controlled subsets of models in Figure 3.

Nevertheless, we can draw the following conclusions from our findings. First, scaling ImageNet models does not lead to better alignment of their representations with human similarity judgments. Differences in human alignment across ImageNet models are mainly attributable to the objective function with which a model was trained, whereas architecture and model capacity are both insignificant. Second, models trained on image/text or more diverse data achieve much better alignment than ImageNet models. Albeit not consistent for zero-shot odd-one-out accuracy, this is clear in both linear probing and regression results. Third, good representations of concepts that are important to human similarity judgments can be recovered from neural network representation spaces. However, representations of less important concepts, such as `sports` and `royal` objects, are more difficult to recover.

How can we train neural networks that achieve better alignment with human concept spaces? Although our results indicate that large, diverse datasets improve alignment, all image/text and JFT models we investigate all attain probing accuracies of 57-58%. By contrast, VICE representations achieve 64%, and a Bayes-optimal classifier achieves 67%. Since our image/text models are trained on datasets of varying sizes (400M to 6.6B images) but achieve similar alignment, we suspect that further scaling of dataset size is unlikely to close this gap. To obtain substantial improvements, it may be necessary to incorporate additional forms of supervision when training the representation itself. Benefits of improving human/machine alignment may extend beyond accuracy on our triplet task, to transfer and retrieval tasks where it is important to capture human notions of similarity.

Acknowledgments

LM, LL, JD and RV acknowledge support by the Federal Ministry of Education and Research (BMBF) for the Berlin Institute for the Foundations of Learning and Data (BIFOLD) (01IS18037A). We thank Robert Geirhos and Andrew Lampinen for their helpful comments on earlier versions of the manuscript.

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In *International Conference on Learning Representations*, 2022.

[2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[3] Elissa M Aminoff, Shira Baror, Eric W Roginek, and Daniel D Leeds. Contextual associations represented both in neural networks and human behavior. *Scientific reports*, 12(1):1–12, 2022.

[4] Maria Attarian, Brett D Roads, and Michael C Mozer. Transforming neural network visual representations to predict human judgments of similarity. In *NeurIPS 2020 Workshop SVRHM*, 2020.

[5] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):1–43, 12 2018. doi: 10.1371/journal.pcbi.1006613.

[6] Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

[7] Alexander Berardino, Valero Laparra, Johannes Ballé, and Eero Simoncelli. Eigen-distortions of hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[8] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020.

[9] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10231–10241, October 2021.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020.

[12] Troy Chinen, Johannes Ballé, Chunhui Gu, Sung Jin Hwang, Sergey Ioffe, Nick Johnston, Thomas Leung, David Minnen, Sean O'Malley, Charles Rosenberg, et al. Towards a semantic perceptual image metric. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 624–628. IEEE, 2018.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[14] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16458–16468, June 2021.

[15] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–7, 2017. doi: 10.1109/ICCCN.2017.8038465.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[17] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5414–5423, June 2021.

[18] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[19] 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. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

[20] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13890–13902. Curran Associates, Inc., 2020.

[21] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23885–23899. Curran Associates, Inc., 2021.

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[23] Umut Güçlü and Marcel A. J. 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. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5023-14.2015.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24, 2019. doi: 10.1371/journal.pone.0223792.

[28] Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, October 2020. doi: 10.1038/s41562-020-00951-3.

[29] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, October 2021.

[30] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.

[31] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In Xuewen Chen, Bo Luo, Feng Luo, Vasile Palade, and M. Arif Wani (eds.), *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, pp. 352–358. IEEE, 2017. doi: 10.1109/ICMLA.2017.0-136.

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021.

[33] Kamila M. Jozwik, Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. Deep Convolutional Neural Networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, 8, 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.01726.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.

[35] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2661–2671. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00277.

[36] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why Do Better Loss Functions Lead to Less Transferable Features? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, volume 34, pp. 28648–28662, 2021.

[37] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv e-prints*, art. arXiv:1404.5997, April 2014.

[38] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[39] Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin D Cubuk. Do better ImageNet classifiers assess perceptual similarity better? *Transactions on Machine Learning Research*, 2022.

[40] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan L. Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

[41] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1): 503–528, 1989.

[42] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R. Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. Words are all you need? capturing human sensory similarity with textual descriptors, 2022.

[43] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7721–7735. PMLR, 18–24 Jul 2021.

[44] Lukas Muttenthaler, Charles Y. Zheng, Patrick McClure, Robert A. Vandermeulen, Martin N. Hebart, and Francisco Pereira. VICE: Variational Interpretable Concept Embeddings. *arXiv e-prints*, art. arXiv:2205.00756, 2022.

[45] Andrew Ng and Geoffrey E. Hinton. Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton, 2017.

[46] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pp. 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4_5.

[47] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12 (Oct):2825–2830, 2011.

[48] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Evaluating (and improving) the correspondence between Deep Neural Networks and Human Representations. *Cogn. Sci.*, 42(8):2648–2669, 2018. doi: 10.1111/cogs.12670.

[49] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for open-vocabulary image classification. *arXiv e-prints*, art. arXiv:2111.10050, 2022.

[50] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[52] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0388-18.2018.

[53] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019.

[54] Brandon RichardWebster, Samuel E. Anthony, and Walter J. Scheirer. PsyPhy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, 2019. doi: 10.1109/TPAMI.2018. 2849989.

[55] Brett D. Roads and Bradley C. Love. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3547–3557, June 2021.

[56] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4036–4054. PMLR, 2022.

[57] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323 (6088):533–536, 1986.

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[59] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2020. doi: 10.1101/407007.

[60] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8634–8644. PMLR, 2020.

[61] Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural MT learn source syntax? In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1526–1534. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1159.

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[63] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research*, 2022.

[64] Laura M Stoinski, Jonas Perkuhn, and Martin N Hebart. THINGS+: New norms and metadata for the THINGS database of 1,854 object concepts and 26,107 natural object images, Jul 2022.

[65] Stuart Sutherland. Cognition: Parallel distributed processing. *Nature*, 323(6088):486–486, Oct 1986. ISSN 1476-4687. doi: 10.1038/323486a0.

[66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308.

[67] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019.

[68] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, volume 33, pp. 18583–18599. Curran Associates, Inc., 2020.

[69] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *CoRR*, abs/2205.04596, 2022. doi: 10.48550/arXiv.2205.04596.

[70] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[71] 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, 2014. doi: 10.1073/pnas.1403112111.

[72] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021.

[73] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, June 2022.

[74] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

A Related Work

Most work comparing neural networks with human behavior has focused on the errors made during image classification. Although ImageNet-trained models appear to make very different errors than humans [52, 20, 21], models trained on larger datasets produce more consistent errors [21], consistent with our findings here. Compared to humans, ImageNet-trained models perform worse on distorted images [54, 15, 31, 18] and rely more heavily on texture cues and less on object shapes [19, 5], although reliance on texture can be mitigated through data augmentation [19, 30, 40], adversarial training [21], or larger datasets [9].

Previous work has also compared human and machine semantic similarity judgments, generally using smaller sets of images and models than we explore here. Jozwik et al. [33] measured the similarity of AlexNet and VGG-16 representations to human similarity judgments of 92 object images inferred from a multi-arrangement task. Peterson et al. [48] compared representations of five neural networks to similarity judgments for six different sets of 120 images, obtained by asking subjects to rate the similarities of pairs from 0 to 10. They report results both with and without rescaling of features. Attarian et al. [4] learned constrained linear transformations of representations to improve the fit of VGG-16 representations to similarity judgments for bird images, but found that unconstrained transformations perform best. Aminoff et al. [3] found that, across 11 networks, representations of contextually associated objects (e.g., bicycles and helmets) were more similar than those of non-associated objects; similarity correlated with both human ratings and reaction times. Roads & Love [55] collect human similarity judgments for the ImageNet validation set and evaluate triplet accuracy on these similarity judgments using 12 ImageNet networks. Most closely related to our work, Marjieh et al. [42] compare similarity of representations of 611 models to cardinal pairwise human similarity judgments. They find a weak correlation between parameter count and models' similarities with humans, and find that incorporating embeddings of both image and text models can further improve the correlation. However, they do not attempt to systematically examine factors that affect alignment between image models and human similarity judgments.

Other studies have focused on perceptual similarity rather than semantic similarity, where the task measures perceived similarity between a reference image and a distorted version of that reference image [50, 74], rather than between distinct images as in our task. Whereas the representations best aligned with human perceptual similarity are obtained from intermediate layers of small architectures [7, 74, 12, 39], the representations best aligned with our odd-one-out judgments are obtained at final model layers, and architecture has little impact.

Our work fits into a broader literature examining relationships between in-distribution accuracy of image classification and other model quality measures, such as accuracy on out-of-distribution data and downstream accuracy when transferring the model. Out-of-distribution accuracy correlates nearly linearly with accuracy on the training distribution [53, 68, 43], although certain forms of data augmentation can improve accuracy under some distribution shifts without an accompanying improvement in in-distribution accuracy [29]. When comparing the transfer learning performance of different architectures trained with similar settings, accuracy on the pretraining task correlates well with accuracy on the transfer tasks [35], although differences in regularization, training objective, and hyperparameters can have a substantial impact on linear transfer accuracy even if the impact on pretraining accuracy is small [35, 36, 1]. In our study, we find that the training objective has a significant impact, as it does for linear transfer. However, in contrast to previous observations regarding out-of-distribution generalization and transfer, we find that better-performing architectures do not achieve greater human alignment.

B Experimental details

B.1 Linear probing

Initialization We initialized the transformation matrix $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ used in Equation 2 with a temperature scaled identity matrix $\tau \boldsymbol{I} \in \mathbb{R}^{p \times p}$ such that $\boldsymbol{W} := \tau \boldsymbol{I}$ at the beginning of the optimization process. τ is model-specific and was found via grid search, minimizing the expected calibration error (ECE) [24]. Details on temperature scaling are described in B.2.

Training We optimized the transformation matrix $\boldsymbol{W}$ via gradient descent, using Adam [34] with a learning rate of $\eta = 0.001$. We performed a grid-search over the learning rate η, where $\eta \in \{0.0001, 0.001, 0.01\}$ and found 0.001 to work best for all models in Table C.1. We trained the linear probe for a maximum of 100 epochs and stopped the optimization process early whenever the generalization performance did not change by a factor of 0.0001 for $T = 10$ epochs.

Cross-validation To obtain a minimally biased estimate of the odd-one-out accuracy of a linear probe, we performed k-fold CV over objects rather than triplets. We partitioned the m objects into two disjoint sets for train and test triplets. Algorithm 1 demonstrates how object partitioning was performed for each of the k folds.

Note that the number of train objects that is sampled uniformly at random without replacement from the set of all objects is dependent on k. We performed a grid-search search over k, where $k \in \{2, 3, 4, 5\}$, and observed that 3-fold and 4-fold CV lead to the best linear probing results. Since objects between train and test triplets were not allowed to overlap, loss of data was inevitable (see Algorithm 1). One can easily see that minimizing the loss of triplet data, comes at the cost of disproportionally decreasing the size of the test set. We decided to proceed with 3-fold CV in our final experiments since using $2/3$ of the objects for training and $1/3$ for testing resulted in a proportionally larger test set than using $3/4$ for training and $1/4$ for testing ($\sim$ 433k train and $\sim$ 54k test triplets for 3-fold CV vs. $\sim$ 616k train and $\sim$ 23k test triplets for 4-fold CV). In general, the larger a test set, the more accurate the estimate of a model's generalization performance. To find the optimal strength of the ℓ_2 regularization for each linear probe, we performed a grid-search over λ for each k value individually. The optimal λ varied between models, where $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1\}$.

B.2 Temperature scaling

It is widely known that classifiers trained to minimize cross-entropy tend to be overconfident in their predictions [66, 24, 56], which is in stark contrast to the high-entropy predictions of VICE. We found it helpful to initialize the transformation matrices for the probing experiments using a temperature parameter, as described in Appendix B.1. For this purpose, we performed temperature scaling [24] on the model outputs for THINGS and searched over the scaling parameter τ for each model. In particular, we considered temperature-scaled predictions

$$p(\{a, b\} | \{i, j, k\}, \tau \boldsymbol{S}) = \frac{\exp(\tau S_{a,b})}{\exp(\tau S_{i,j}) + \exp(\tau S_{i,k}) + \exp(\tau S_{j,k})},$$

where we multiply $\boldsymbol{S}$ in Equation 1 by a constant $\tau > 0$ and $S_{i,j}$ is the inner product of the model representations for images i and j, i.e. the zero-shot similarities. There are several conceivable criteria that could be minimized to find the optimal scaling parameter τ from a set of candidates. For our analyses we considered the following,

Algorithm 1 Algorithm for object partitioning during k-fold CV

```
Input: $(\mathcal{D}, m)$    ▷ Here, $\mathcal{D} := (\{a_s, b_s\}, \{i_s, j_s, k_s\})_{s=1}^{n}$ and $m$ is the number of objects
  $[m] = \{1, \ldots, m\}$    ▷ $|[m]| = m$
  $\mathbb{O}_{\text{train}} \sim \mathcal{U}([m])$    ▷ Sample a number of train objects uniformly at random without replacement
  $\mathbb{O}_{\text{test}} := [m] \setminus \mathbb{O}_{\text{train}}$    ▷ Test objects are the remaining objects
  $\mathcal{D}_{\text{train}} := \{\}$    ▷ Initialize an empty set for the train triplets
  $\mathcal{D}_{\text{test}} := \{\}$    ▷ Initialize an empty set for the test triplets
  for $s \in \{1, \ldots, n\}$ do
      assignments $\triangleq$ list( )    ▷ For each triplet initialize an empty list to control object assignments
      for $x \in \{i_s, j_s, k_s\}$ do
          if $(x \in \mathbb{O}_{\text{train}})$ then
              assignment $\triangleq$ "train"
          else
              assignment $\triangleq$ "test"
          end if
          assignments $\leftarrow$ assignment    ▷ Append current assignment to the list of assignments
      end for
      if $(\mathrm{len}(\mathrm{set}(\text{assignments})) \neq 1)$ then
          continue    ▷ If not all objects in a triplet belong to the same set of objects, discard triplet
      else
          assignment $\triangleq$ $\mathrm{pop}(\mathrm{set}(\text{assignments}))$    ▷ Get object set assignment of current triplet
          if (assignment is "train") then
              $\mathcal{D}_{\text{train}} := \mathcal{D}_{\text{train}} \cup \mathcal{D}_s$    ▷ Assign current triplet to the train set
          else
              $\mathcal{D}_{\text{test}} := \mathcal{D}_{\text{test}} \cup \mathcal{D}_s$    ▷ Assign current triplet to the test set
          end if
      end if
  end for
Output: $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$    ▷ Return both train and test triplet sets
```

- Average Jensen-Shannon (JS) distance between model zero-shot probabilities and VICE probabilities over all triplets
- Average Kullback-Leibler divergence (KLD) between model zero-shot probabilities and VICE probabilities over all triplets
- Expected Calibration Error (ECE) [24].

The ECE is defined as follows. Let $\mathcal{D} = (\{a_s, b_s\}, \{i_s, j_s, k_s\})_{s=1}^{n}$ be the set of triplets and human responses from Hebart et al. [28]. For a given triplet $\{i, j, k\}$ and similarity matrix $\boldsymbol{S}$ we define confidence as

$$\mathrm{conf}\left(\{i, j, k\}, \boldsymbol{S}\right) := \max_{\{a,b\} \subset \{i,j,k\}} p\left(\{a, b\} \mid \{i, j, k\}, \boldsymbol{S}\right).$$

This corresponds to the expected accuracy of the Bayes classifier for that triplet according to the probability model from $\boldsymbol{S}$ with Equation 1. We define $B_m(\boldsymbol{S})$ to be those training triplets where

$$\mathrm{conf}\left(\{i_s, j_s, k_s\}, \boldsymbol{S}\right) \in \left[\frac{m-1}{10}, \frac{m}{10}\right].$$

For a similarity matrix, $\boldsymbol{S}$, and a set of triplets with responses, $\mathcal{D}' \subset \mathcal{D}$, we define $\mathrm{acc}\left(\mathcal{D}', \boldsymbol{S}\right)$ to be the portion of triplets in $\mathcal{D}'$ correctly classified according to the highest similarity according to $\boldsymbol{S}$. Finally for a set of triplets $\mathcal{D}' \subset \mathcal{D}$ and similarity matrix $\boldsymbol{S}$ we define $\mathrm{conf}(\mathcal{D}')$ to be the average confidence over that set (triplet responses are simply ignored). The ECE is defined as

$$\mathrm{ECE}\left(\tau, \boldsymbol{S}\right) = \sum_{m=1}^{10} \frac{\left|B_m\left(\tau \boldsymbol{S}\right)\right|}{n} \left|\mathrm{acc}(B_m\left(\tau \boldsymbol{S}\right)) - \mathrm{conf}\left(B_m\left(\tau \boldsymbol{S}\right)\right)\right|.$$

Intuitively, the ECE is low if for each subset $B_m(\tau \boldsymbol{S})$ the model's accuracy and its confidence are near each other. A model will be well-calibrated if its confidence in predicting the odd-one-out in a triplet corresponds to the probability that this prediction is correct.

Of the three considered criteria, ECE resulted in the clearest optima when varying τ, whereas KLD plateaued with increasing τ and JS distance was numerically unstable, most probably because the model output probabilities were near zero for some pairs, which may result in very large JS distance. For all models, we performed a grid-search over $\tau \in \{1 \cdot 10^0, 7.5 \cdot 10^{-1}, 5 \cdot 10^{-1}, 2.5 \cdot 10^{-1}, 1 \cdot 10^{-1}, 7.5 \cdot 10^{-2}, 5 \cdot 10^{-2}, 2.5 \cdot 10^{-2}, 1 \cdot 10^{-2}, 7.5 \cdot 10^{-3}, 5 \cdot 10^{-3}, 2.5 \cdot 10^{-3}, 1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-5}, 1 \cdot 10^{-5}\}$.

B.3 Linear regression

Cross-validation We used ridge regression, that is ℓ_2-regularized linear regression, to find the transformation matrix $\boldsymbol{A}_{j,:}$ and bias b_j that result in the best fit. We employed nested k-fold CV for each of the d VICE dimensions. For the outer CV we performed a grid-search over k, where $k \in \{2, 3, 4, 5\}$, similarly to how k-fold CV was performed for linear probing (see B.1). For our final experiments, we used 5-fold CV to obtain a minimally biased estimate for the R^2 score of the regression fit. For the inner CV, we leveraged leave-one-out CV to determine the optimal α for Equation 3 using `RidgeCV` from Pedregosa et al. [47]. We performed a grid search over α, where $\alpha \in \{0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 1000000\}$.

C Models

First, we evaluate supervised models trained on ImageNet [58], such as AlexNet [37], various VGGs [62], ResNets [25], EfficientNets [67], ResNext models [70], and Vision Transformers (ViTs) trained on ImageNet-1K[16] or ImageNet-21K [63] respectively. Second, we analyze recent state-of-the-art models trained on image/text data, CLIP-RN & CLIP-ViT [51], ALIGN [32] and BASIC-L [49]. Third, we evaluate self-supervised (SSL) models that were trained with a contrastive learning objective such as SimCLR [11] and MoCo [26], recent SSL models that were trained with a non-contrastive learning objective (no negative examples), BarlowTwins [72], SwAV [10], and VICReg [6], as well as earlier SSL, non-Siamese models, Jigsaw [46], and Rotnet [22]. Last, we evaluate the largest available ViT [73], trained on the proprietary JFT-3B image classification dataset, which consists of approximately three billion images belonging to approximately 30,000 classes [73]. See Table C.1 for further details regarding the models used. Figure C.1 shows the odd-one-out accuracy as a function of layer depth in a neural network for a few different network architectures. Later layers generally perform better which is why we performed our analyses exclusively for the logits or penultimate/embedding layers of the models in Table C.1.

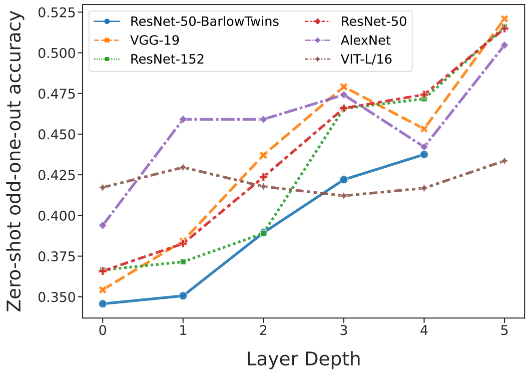

Figure C.1: Zero-shot odd-one-out accuracy at different layers for a subset of selected models.

D CIFAR-100 triplet task

In a similar vein to the THINGS triplet task, we constructed a reference triplet task from the CIFAR-100 dataset [38]. To show pairs of images that are similar to each other, but do not depict the same object, we leverage the 20 coarse classes of the dataset rather than the original fine-grained classes. For each triplet, we sample two images from the same and an one odd-one-out image from a different coarse class. We

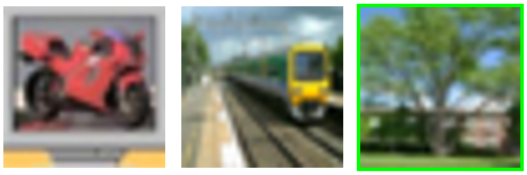

Figure D.1: An example triplet from the CIFAR-100 coarse dataset. The left two images are from one of the two CIFAR-100 "vehicle" superclasses, so the right-most image is the odd-one-out.

Model	Source	Architecture	Dataset	Objective	ImageNet Acc.
AlexNet	[37]	AlexNet	ImageNet-1K	Supervised (softmax)	56.52%
ALIGN	[32]	EfficientNet	ALIGN dataset	Image/Text (contrastive)	85.11%
Basic-L	[49]	CoAtNet	ALIGN + JFT-5B	Image/Text (contrastive)	89.45%
CLIP ResNet-50	[51]	ResNet	CLIP dataset	Image/Text (contrastive)	73.30%
CLIP ViT-B/32	[51]	ViT	CLIP dataset	Image/Text (contrastive)	76.10%
DenseNet-121	[35]	DenseNet	ImageNet-1K	Supervised (softmax)	75.64%
DenseNet-169	[35]	DenseNet	ImageNet-1K	Supervised (softmax)	76.73%
DenseNet-201	[35]	DenseNet	ImageNet-1K	Supervised (softmax)	77.14%
EfficientNet B0	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	77.69%
EfficientNet B1	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	78.64%
EfficientNet B2	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	80.61%
EfficientNet B3	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	82.01%
EfficientNet B4	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	83.38%
EfficientNet B5	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	83.44%
EfficientNet B6	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	84.01%
EfficientNet B7	[67]	EfficientNet	ImageNet-1K	Supervised (softmax)	84.12%
Inception-ResNet V2	[35]	Inception	ImageNet-1K	Supervised (softmax)	80.26%
Inception-V1	[35]	Inception	ImageNet-1K	Supervised (softmax)	73.63%
Inception-V2	[35]	Inception	ImageNet-1K	Supervised (softmax)	75.34%
Inception-V3	[35]	Inception	ImageNet-1K	Supervised (softmax)	78.64%
Inception-V4	[35]	Inception	ImageNet-1K	Supervised (softmax)	79.92%
MobileNet-V1	[35]	MobileNet	ImageNet-1K	Supervised (softmax)	72.39%
MobileNet-V2	[35]	MobileNet	ImageNet-1K	Supervised (softmax)	71.67%
MobileNet-V2 (1.4)	[35]	MobileNet	ImageNet-1K	Supervised (softmax)	74.66%
NASNet-L	[35]	NASNet	ImageNet-1K	Supervised (softmax)	80.77%
NASNet-Mobile	[35]	NASNet	ImageNet-1K	Supervised (softmax)	73.57%
ResNet-50-BarlowTwins	[72]	ResNet	ImageNet-1K	Self-sup. (non-contrastive)	71.80%
ResNet-50-Jigsaw	[46]	ResNet	ImageNet-1K	Self-sup. (non-Siamese)	48.57%
ResNet-50-MoCo-v2	[26]	ResNet	ImageNet-1K	Self-sup. (contrastive)	66.40%
ResNet-50-RotNet	[22]	ResNet	ImageNet-1K	Self-sup. (non-Siamese)	48.20%
ResNet-50-SimCLR	[11]	ResNet	ImageNet-1K	Self-sup. (contrastive)	69.68%
ResNet-50-SWAV	[10]	ResNet	ImageNet-1K	Self-sup. (non-contrastive)	74.92%
ResNet-50-VICReg	[6]	ResNet	ImageNet-1K	Self-sup. (non-contrastive)	73.20%
ResNet-18	[25]	ResNet	ImageNet-1K	Supervised (softmax)	69.76%
ResNet-34	[25]	ResNet	ImageNet-1K	Supervised (softmax)	73.31%
ResNet-50	[25]	ResNet	ImageNet-1K	Supervised (softmax)	76.13%
ResNet-101	[25]	ResNet	ImageNet-1K	Supervised (softmax)	77.37%
ResNet-152	[25]	ResNet	ImageNet-1K	Supervised (softmax)	78.31%
ResNet-101	[35]	ResNet	ImageNet-1K	Supervised (softmax)	78.56%
ResNet-152	[35]	ResNet	ImageNet-1K	Supervised (softmax)	79.29%
ResNet-50	[35]	ResNet	ImageNet-1K	Supervised (softmax)	76.93%
ResNet-50	[36]	ResNet	ImageNet-1K	Supervised (softmax)	77.42%
ResNet-50 (extra weight decay)	[36]	ResNet	ImageNet-1K	Supervised (softmax+)	77.82%
ResNet-50 (label smoothing)	[36]	ResNet	ImageNet-1K	Supervised (softmax+)	77.63%
ResNet-50 (logit penality)	[36]	ResNet	ImageNet-1K	Supervised (softmax+)	77.67%
ResNet-50 (mixup)	[36]	ResNet	ImageNet-1K	Supervised (softmax+)	77.92%
ResNet-50 (AutoAugment)	[36]	ResNet	ImageNet-1K	Supervised (softmax)	77.64%
ResNet-50 (logit norm)	[36]	ResNet	ImageNet-1K	Supervised (softmax+)	77.83%
ResNet-50 (cosine softmax)	[36]	ResNet	ImageNet-1K	Supervised (softmax+)	77.86%
ResNet-50 (sigmoid)	[36]	ResNet	ImageNet-1K	Supervised (sigmoid)	78.18%
ResNet-50 (softmax)	[36]	ResNet	ImageNet-1K	Supervised (softmax)	76.94%
ResNet-50 (squared error)	[36]	ResNet	ImageNet-1K	Supervised (squared error)	77.13%
ResNeXt-101 32x8d	[70]	ResNeXt	ImageNet-1K	Supervised (softmax)	79.32%
ResNeXt-50 32x4d	[70]	ResNeXt	ImageNet-1K	Supervised (softmax)	81.11%
VGG-11	[62]	VGG	ImageNet-1K	Supervised (softmax)	69.02%
VGG-13	[62]	VGG	ImageNet-1K	Supervised (softmax)	69.93%
VGG-16	[62]	VGG	ImageNet-1K	Supervised (softmax)	71.59%
VGG-19	[62]	VGG	ImageNet-1K	Supervised (softmax)	72.38%
ViT-B/16 I1K	[63]	ViT	ImageNet-1K	Supervised (sigmoid)	77.66%
ViT-B/16 I21K	[63]	ViT	ImageNet-21K	Supervised (sigmoid)	83.77%
ViT-B/32 I1K	[63]	ViT	ImageNet-1K	Supervised (sigmoid)	72.08%
ViT-B/32 I21K	[63]	ViT	ImageNet-21K	Supervised (sigmoid)	79.16%
ViT-L/16 I1K	[63]	ViT	ImageNet-1K	Supervised (sigmoid)	75.11%
ViT-L/16 I21K	[63]	ViT	ImageNet-21K	Supervised (sigmoid)	83.13%
ViT-S/32 I1K	[63]	ViT	ImageNet-1K	Supervised (sigmoid)	72.18%
ViT-S/32 I21K	[63]	ViT	ImageNet-21K	Supervised (sigmoid)	72.93%
ViT-G/14 JFT	[73]	ViT	JFT-3B	Supervised (sigmoid)	89.01%
ViT-B-16	[16]	ViT	ImageNet-1K	Supervised (softmax)	81.07%
ViT-B-32	[16]	ViT	ImageNet-1K	Supervised (softmax)	75.91%

Table C.1: Pretrained neural networks that we considered in our analyses.

restrict ourselves to examples from the CIFAR-100 train set and exclude the validation set. We randomly sample a total of 50,000 triplets which is equivalent to the size of the original train set. Figure D.1 shows an example triplet for this task.

E Linear probing

In the left plot of Figure E.1, we show probing odd-one-out accuracy as a function of ImageNet accuracy for all models in Table C.1. Similarly to the findings depicted in Figure 2, we observe a low Pearson correlation coefficient ($r = 0.241$) between ImageNet accuracy and probing odd-one-out accuracy. As a reference, here we show again ImageNet accuracy as a function of zero-shot odd-one-out accuracy on the CIFAR-100 coarse triplet task. In Figure E.2 we compare probing odd-one-out accuracy with zero-shot odd-one-out accuracy for models pretrained on ImageNet-1K or ImageNet-21K. We observe a strong positive correlation of $r = 0.963$ between probing odd-one-out and zero-shot odd-one-out accuracy.

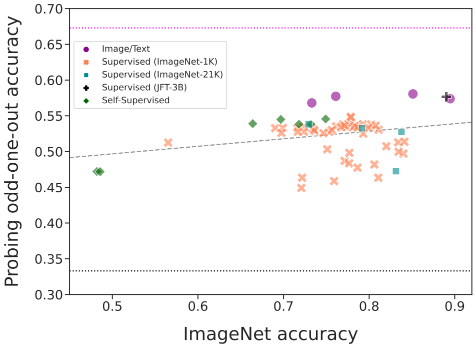

Figure E.1: Probing odd-one-out accuracy as a function of ImageNet accuracy. Dashed diagonal line indicate a least-squares fit. Dashed horizontal lines reflect chance-level or ceiling accuracy respectively.

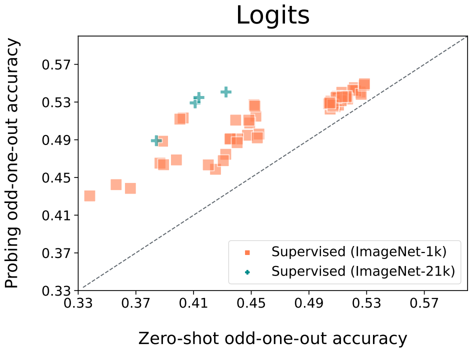

Figure E.2: Probing odd-one-out accuracy as a function of zero-shot odd-one-out accuracy for the logits layer of all ImageNet models in Table C.1. Dashed line indicates $x = y$ line.

F Human alignment is concept specific

To examine how well neural nets represent human concepts, we partitioned the original triplet dataset $\mathcal{D}$ into two sets $\mathcal{D}^*$ and $\mathcal{D}^\dagger$, with $\mathcal{D}^*$ containing triplets correctly predicted by VICE and $\mathcal{D}^\dagger$ containing those which are not. The triplets in $\mathcal{D}^\dagger$ mostly have high entropy, i.e., chosen odd-one-out is not consistent for humans. The triplets in $\mathcal{D}^\dagger$ are not used in the following analysis. We further partitioned $\mathcal{D}^*$ into 45 subsets according to the 45 VICE dimensions, $\mathcal{D}^*_1, \ldots, \mathcal{D}^*_{45}$. A triplet belongs to $\mathcal{D}^*_j$ when the sum of the VICE representations for the two most similar objects in the triplet, $\boldsymbol{x}_a$, $\boldsymbol{x}_b$, attains its maximum in dimension j, i.e. $j = \arg\max_{j'} x_{a,j'} + x_{b,j'}$.

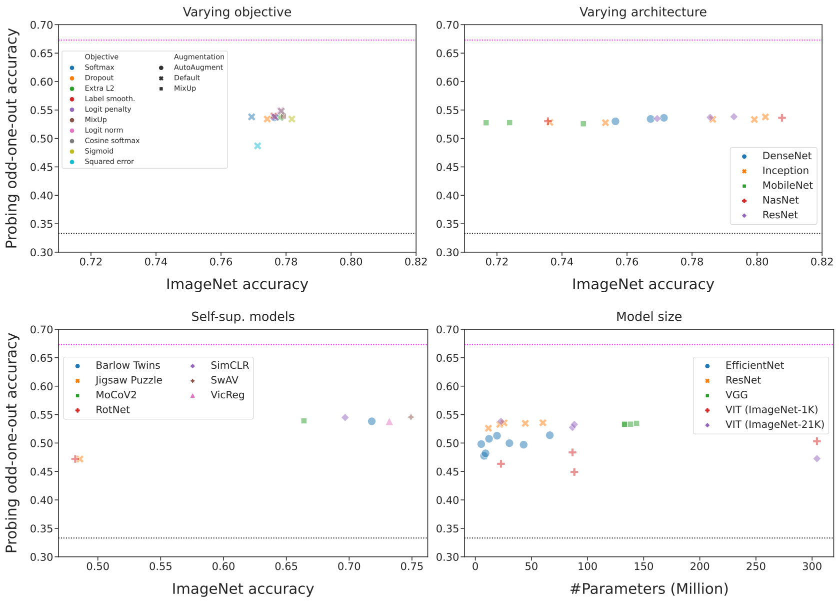

Figure E.3: Probing odd-one-out accuracy for THINGS as a function of ImageNet accuracy or number of model parameters. **Top**: Models on the left have the same architecture (ResNet-50) but were trained with a different objective function [36]. Models on the right were trained with the same objective function but vary in architecture [35]. **Bottom**: Performance for different SSL models on the left, and a subset of ImageNet models with their number of parameters on the right. Dashed horizontal lines reflect chance-level or ceiling accuracy respectively.

G Linear regression

G.1 Overall performance

In Figure G.1, we compare odd-one-out accuracies after linear probing with zero-shot odd-one-out accuracies and probing odd-one-out accuracies for logits vs. embedding layers of ImageNet models. The results are consistent with the results from linear probing shown in Figure 4.

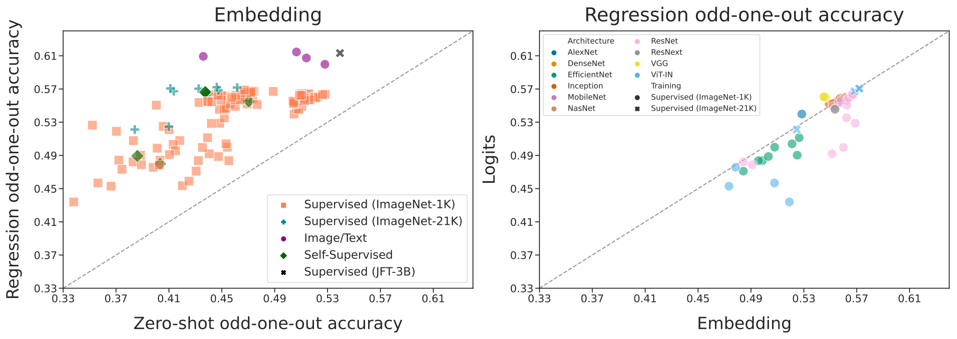

Figure G.1: **Left**: Zero-shot and regression odd-one-out accuracies for the embedding layer of all neural nets. **Right**: Regression odd-one-out accuracy for the embedding and logits layer for all supervised models trained on ImageNet-1K or ImageNet-21K. Dashed line indicates $x = y$.

G.2 Can human concepts be recovered via linear regression?

In addition to linear probing, we performed ℓ_2-regularized linear regression to examine models' ability to predict VICE dimension. This analysis helped us to further understand whether human concepts can be recovered from a neural network's representation space. Here, for each of the 45 representation dimensions, j, from VICE, we minimized the following least-squares objective

$$\arg\min_{\boldsymbol{A}_{j,:},b_j} \sum_{i=1}^{m} \left(Y_{i,j} - (\boldsymbol{A}_{j,:}\boldsymbol{x}_i + b_j)\right)^2 + \alpha_j \|\boldsymbol{A}_{j,:}\|_2^2, \tag{3}$$

where $Y_{i,j}$ is the value of the j^{th} VICE dimension for image i, $\boldsymbol{x}_i$ is the neural network representation of image i, and $\alpha_j > 0$ is a regularization hyperparameter. Each dimension was optimized separately with α_j selected via CV using grid search (details are in Appendix B.3).

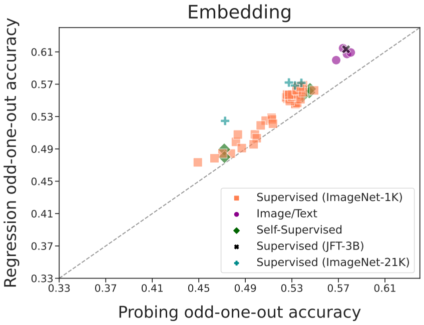

Figure G.2: Regression as a function of probing odd-one-out accuracies for all models in Table C.1

The results from this analysis corroborate the findings from § 3.2: models trained on image/text data and ViT-G/14 JFT consistently provided the best fit for VICE dimensions, while AlexNet and EfficientNets showed the poorest regression performance. Furthermore, we investigated whether the recovered VICE dimensions show better alignment than the original network embeddings. All models were evaluated on the THINGS triplet task using a similarity matrix $\boldsymbol{S}$ with $S_{ij} := (\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x}_j + \boldsymbol{b})$, where $\boldsymbol{A}$ and $\boldsymbol{b}$ are obtained by stacking the optimizers from Equation 3, so $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is a linear regression from a neural network representation to the VICE representation. In Figure G.2, we compare odd-one-out accuracies after linear probing and regression respectively. The two performance measures are highly correlated for both the embedding ($r = 0.960$) and logits ($r = 0.966$) layers. Note that odd-one-out accuracies are slightly higher for regression. We hypothesize that this is due to VICE being trained on all objects in the data so the transformation matrix learned in linear regression indirectly has access to all objects opposed to the transformation matrix learned during probing. Moreover, 2/3 of the objects were used for training the linear probe, whereas 4/5 of the objects were used to fit linear regression.

Figure G.3 shows the R^2 score for fitting a the same subset of VICE dimensions used in Figure 5 from embedding-layer representations.

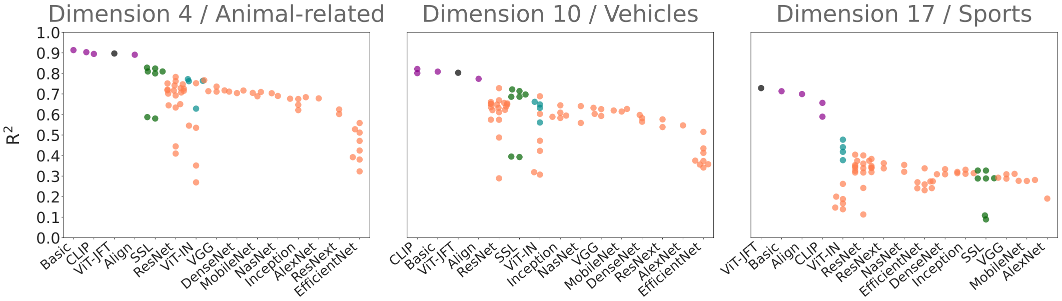

Figure G.3: R^2 scores for all models in Table C.1 after fitting an ℓ_2-regularized linear regression to predict individual VICE dimensions from the embedding-layer representation of the images in THINGS. Color-coding was determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).