HUMAN ALIGNMENT OF NEURAL NETWORK REPRESENTATIONS

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen
Machine Learning Group, Technische Universität Berlin & BIFOLD
Berlin, Germany

Simon Kornblith
Google Research, Brain Team

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 the alignment between the representations learned by neural networks and human mental representations inferred from behavioral responses. We find that model scale and architecture have essentially no effect on the alignment with human behavioral responses, whereas the training dataset and objective function both have a much larger impact. These findings are consistent across three datasets of human similarity judgments collected using two different tasks. Linear transformations of neural network representations learned from behavioral responses from one dataset substantially improve alignment with human similarity judgments on the other two datasets. In addition, we find that some human concepts such as food and animals are well-represented by neural networks 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. (1986) 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 (Sutherland, 1986; Ng & Hinton, 2017). 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 (Shankar et al., 2020; Beyer et al., 2020; Vasudevan et al., 2022). As computer vision systems become increasingly widely used outside of research, we would like to know if they see the world in the same way that humans do. However, the extent to which the conceptual representations learned by these systems align with those used by humans remains unclear.

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 (Geirhos et al., 2018; Rajalingham et al., 2018; Geirhos et al., 2021) and the ability of their representations to predict neural activity in primate brains (Yamins et al., 2014; Güçlü & van Gerven, 2015; Schrimpf et al., 2020), with mixed results. Networks trained on more data make somewhat more human-like errors (Geirhos et al., 2021), but do not necessarily obtain a better fit to brain data (Schrimpf et al., 2020). Here, we approach the question of alignment between human and machine representation spaces more directly. We focus primarily on 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 (Hebart et al., 2020). 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. We confirm our findings on two independent datasets collected using the multi-arrangement task, in which humans arrange images according to their similarity Cichy et al. (2019); King et al. (2019). Based on these analyses, 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 arise primarily from differences in objective functions and other hyperparameters 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.
- A linear transformation trained to improve odd-one-out accuracy on THINGS substantially increases the degree of alignment on held-out THINGS images as well as for two human similarity judgment datasets that used a multi-arrangement task to collect behavioral responses.
- We use a sparse Bayesian model of human mental representations (Muttenthaler et al., 2022) 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.

2 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 (Rajalingham et al., 2018; Geirhos et al., 2020; 2021), models trained on larger datasets than ImageNet exhibit greater error consistency (Geirhos et al., 2021). Compared to humans, ImageNet-trained models perform worse on distorted images (RichardWebster et al., 2019; Dodge & Karam, 2017; Hosseini et al., 2017; Geirhos et al., 2018) and rely more heavily on texture cues and less on object shapes (Geirhos et al., 2019; Baker et al., 2018), although reliance on texture can be mitigated through data augmentation (Geirhos et al., 2019; Hermann et al., 2020; Li et al., 2021), adversarial training (Geirhos et al., 2021), or larger datasets (Bhojanapalli et al., 2021).

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. (2017) 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. (2018) compared representations of five neural networks to pairwise similarity judgments for six different sets of 120 images. Aminoff et al. (2022) 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 (2021) collect human similarity judgments for ImageNet and evaluate triplet accuracy on these similarity judgments using 12 ImageNet networks. Most closely related to our work, Marjieh et al. (2022) measure aligment between representations of networks that process images, videos, audio, or text and the human pairwise similarity judgments of Peterson et al. (2018). They report a weak correlation between parameter count and alignment, but do not systematically examine factors that affect this relationship.

Other studies have focused on perceptual rather than semantic similarity, where the task measures perceived similarity between a reference image and a distorted version of that reference image (Ponomarenko et al., 2009; Zhang et al., 2018), 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 (Berardino et al., 2017; Zhang et al., 2018; Chinen et al., 2018; Kumar et al., 2022), the representations best aligned with our odd-one-out judgments are obtained at final model layers, and architecture has little impact. Jagadeesh & Gardner (2022) compared human odd-one-out judgments with similarities implied by neural network representations and

brain activity. They found that artificial and biological representations distinguish the odd one out when it differs in category, but do not distinguish natural images from synthetic scrambled images.

Our work fits into a broader literature examining relationships between in-distribution accuracy of image classification and other model quality measures, including accuracy on out-of-distribution (OOD) data and downstream accuracy when transferring the model. OOD accuracy correlates nearly linearly with accuracy on the training distribution (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021), although data augmentation can improve accuracy under some shifts without improving in-distribution accuracy (Hendrycks et al., 2021). When comparing the transfer learning performance across different architectures trained with similar settings, accuracy on the pretraining task correlates well with accuracy on the transfer tasks (Kornblith et al., 2019b), but differences in regularization, training objective, and hyperparameters can affect linear transfer accuracy even when the impact on pretraining accuracy is small (Kornblith et al., 2019b; 2021; Abnar et al., 2022). 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 OOD generalization and transfer, we find that better-performing architectures do not achieve greater human alignment.

3 METHODS

3.1 DATA

Our primary analyses use images and corresponding human odd-one-out triplet judgments from the THINGS dataset (Hebart et al., 2019). THINGS consists of a collection of 1,854 object categories, concrete nameable nouns in the English language that can be easily identified as a central object in a natural image, along with representative images for these categories. For presentation purposes, we have replaced the images used in Hebart et al. (2020) with images similar in appearance that are licensed under CC0 (Stoinski et al., 2022). We additionally consider two datasets of images with human similarity judgments obtained from a multi-arrangement task (Cichy et al., 2019; King et al., 2019). We briefly describe the procedures that were used to obtain these datasets below.

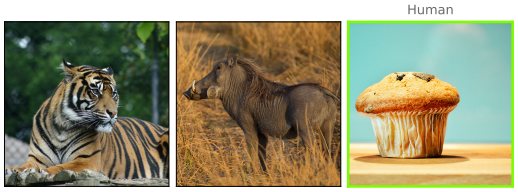


Figure 1: An example triplet from Hebart et al. (2020), where neural nets choose a different odd-one-out than a human. The images in this triplet are copyright-free images from THINGS + (Stoinski et al., 2022).

THINGS triplet task Hebart et al. (2020) collected similarity judgments from human participants on images in THINGS in the form of responses to a *triplet task*. In this task, images from three distinct categories are presented to a participant, and the participant selects the image that is most different from the other two (or equivalently the pair that are most similar). The triplet task has been used to study properties of human mental representation for many decades (e.g., Fukuzawa et al., 1988; Robilotto & Zaidi, 2004; Hebart et al., 2020). Compared to tasks involving numerical/Likert-scale pairwise similarity judgments, the triplet task does not require different subjects to interpret the scale similarly and does not require that the degree of perceived similarity is cognitively accessible.

Hebart et al. (2020) collected 1.46 million unique responses crowdsourced from 5,301 workers. See Figure 1 for an example triplet. Some triplets offer an obvious answer to the triplet task, e.g. "cat", "dog", "candelabra", whereas others can be ambiguous, e.g. "knife", "table", "candelabra." To estimate the consistency of triplet choices among participants Hebart et al. (2020) collected 25 responses for each triplet in a randomly selected set of 1,000 triplets. From these responses, Hebart et al. (2020) determined that the maximum achievable odd-one-out accuracy is $67.22\% \pm 1.04\%$.

Multi-arrangement task The multi-arrangement task is another task commonly used to measure human similarity judgments (Kriegeskorte & Mur, 2012). In this task, subjects arrange images on a computer screen so that the distances between them reflect their similarities. We use multi-arrangement task data from two recent studies. Cichy et al. (2019) collected similarity judgments from 20 human participants for 118 natural images from ImageNet (Deng et al., 2009), and King et al. (2019) collected similarity judgments from 20 human participants for two natural image sets with 144 images per image set. The 144 images correspond to 48 object categories, each with three images. For simplicity, we report results based on only one of these two sets of images.

3.2 METRICS

Zero-shot odd-one-out accuracy To measure alignment between humans and neural networks on the THINGS triplet task, 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 comprise the triplet, we first construct a similarity matrix $\boldsymbol{S} \in \mathbb{R}^{3\times3}$ where $S_{i,j} \coloneqq \boldsymbol{x}_i^\top \boldsymbol{x}_j / (\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2)$, the cosine similarity between a pair of representations.[1] 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 of supervised ImageNet models, we report the better of the accuracies obtained from representations of the penultimate embedding layer and logits; for self-supervised models, we use the projection head input; for image/text models, we use the representation from the joint image/text embedding space; and for the JFT-3B model, we use the penultimate layer. In Figure B.1 we show that representations obtained from earlier layers performed worse than representations from top layers, as in previous work (Montavon et al., 2011).

Probing In addition to measuring zero-shot odd-one-out accuracy on THINGS, we also learn a linear transformation of each neural network's representation that maximizes odd-one-out accuracy and then measure odd-one-out accuracy of the transformed representation on triplets comprising a held-out set of images. Following Alain & Bengio (2017), we refer to this procedure as linear probing. To learn the linear probe, we formulate the notion of the odd-one-out probabilistically, as in Hebart et al. (2020). Given image 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 object 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 $S_{ij} = (\boldsymbol{W}\boldsymbol{x}_i)^\top (\boldsymbol{W}\boldsymbol{x}_j)$ that optimizes

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

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 A.1 and Algorithm 1 respectively.

RSA To measure the alignment between human and neural net representation spaces on multi-arrangement datasets, following previous work, we perform representational similarity analysis (RSA; Kriegeskorte et al. (2008)) and compute correlation coefficients between neural network and human representational similarity matrices (RSMs) for the same sets of images (Kriegeskorte & Kievit, 2013; Cichy et al., 2019). We construct RSMs using a Pearson correlation kernel and measure the Spearman correlation between RSMs. We measure alignment on multi-arrangement datasets in a zero-shot setting as well as after applying the linear probe $\boldsymbol{W}$ learned on THINGS.

3.3 MODELS

Vision 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; three models trained on EcoSet (Mehrer et al., 2021), which is another natural image dataset; a "gigantic" Vision Transformer trained on the proprietary JFT-3B dataset (ViT-G/14 JFT) (Zhai et al., 2022); and image/text models CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BASIC (Pham et al., 2022). A comprehensive list of models that we analyze can be found in Table B.1. To obtain image representations, we use `thingsvision`, a Python library for extracting activations from neural nets (Muttenthaler & Hebart, 2021). 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 (Liu & Nocedal, 1989).

[1] We use cosine similarity rather than dot products because it nearly always yields similar or better zero-shot odd-one-out accuracies, as shown in Figure B.2.

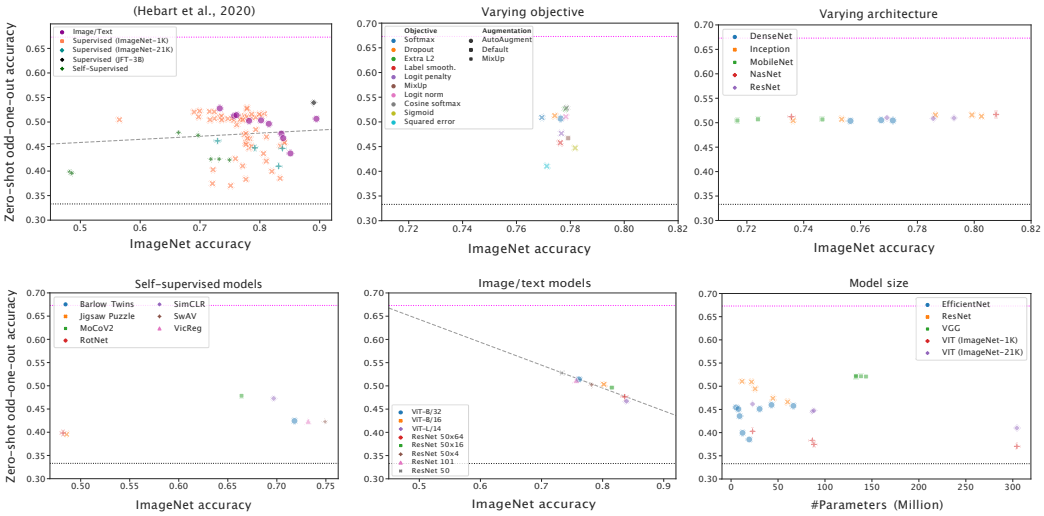


Figure 2: Zero-shot odd-one-out accuracy on THINGS only weakly correlates with ImageNet accuracy and varies with training objective but not with model architecture. **Top left**: Zero-shot accuracy as a function of ImageNet accuracy for all models. Diagonal line indicates least-squares fit. **Top center**: Models with the same architecture (ResNet-50) trained with a different objective function or different data augmentation. Since MixUp alters both inputs and targets, it is listed under both objectives and augmentations. **Top right**: Models trained with the same objective (softmax cross-entropy) but with different architectures. **Bottom left**: Performance of different SSL models. **Bottom center**: Zero-shot accuracy is negatively correlated with ImageNet accuracy for image/text models. **Bottom right**: A subset of ImageNet models with their number of parameters, colored by model family. Note that, in this subplot, models that belong to different families come from different sources and were trained with different objectives, hyperparameters, etc.; thus, models are only directly comparable within a family. In all plots, horizontal lines reflect chance-level or ceiling accuracy. See also Table B.1.

VICE Several of our analyses make use of human concept representations obtained by Variational Interpretable Concept Embeddings (VICE), an approximate Bayesian method for finding representations that explain human odd-one-out responses in a triplet task (Muttenthaler et al., 2022). VICE uses mean-field VI to learn a sparse representation for each image that explains the associated behavioral responses. VICE achieves an odd-one-out accuracy of $\sim 64\%$ on THINGS, which is only marginally below the ceiling accuracy of 67.22% (Hebart et al., 2020). The representation dimensions obtained from VICE are highly interpretable and thus give insight into properties humans deem important for similarity judgments. We use the representation dimensions to analyze the alignment of neural network representations with human concept spaces. However, VICE is not a vision model, and can only predict odd-one-out judgments for images included in the training triplets.

4 EXPERIMENTS

Here, we investigate how closely the representation spaces of neural networks align with humans' concept spaces, and whether concepts can be recovered from a representation via a linear transformation. Odd-one-out accuracies are measured on THINGS, unless otherwise stated.

4.1 ODD-ONE-OUT VS. IMAGENET ACCURACY

We begin by comparing zero-shot odd-one-out accuracy for THINGS with ImageNet accuracy for all models in Figure 2 (top left). ImageNet accuracy generally is a good predictor for transfer learning performance (Kornblith et al., 2019b; Djolonga et al., 2021; Ericsson et al., 2021). However, while ImageNet accuracy is highly correlated with odd-one-out accuracy for a reference triplet task that uses the CIFAR-100 superclasses (see Appendix C), its correlation with accuracy on human odd-one-out judgments is very weak ($r = 0.099$). This raises the question of whether there are model, task, or data characteristics that influence human alignment.

Architecture or objective? We investigate odd-one-out accuracy as a function of ImageNet accuracy for models that vary in the training objective/final layer regularization, data augmentation, or architecture with all other hyperparameters fixed. Models with the same architecture (ResNet-50) trained with different objectives (Kornblith et al., 2021) yield substantially different zero-shot odd-one-out accuracies (Figure 2 top center). Conversely, models with different architectures trained

with the same objective (Kornblith et al., 2019b) achieve similar odd-one-out accuracies, although their ImageNet accuracies vary significantly (Figure 2 top right). Thus, whereas architecture does not appear to affect odd-one-out accuracy, training objective has a significant effect.

Training objective also affects which layer yields the best human alignment. For networks trained with vanilla softmax cross-entropy, the logits layer consistently yields higher zero-shot odd-one-out accuracy than the penultimate layer, but among networks trained with other objectives, the penultimate layer often provides higher odd-one-out accuracy than the logits (Figure E.2). The superiority of the logits layer of networks trained with vanilla softmax cross-entropy is specific to the odd-one-out task and RSA and does not hold for linear transfer, as we show in Appendix D.

Self-supervised learning Jigsaw (Noroozi & Favaro, 2016) and RotNet (Gidaris et al., 2018) show substantially worse alignment with human judgments than other SSL models (Figure 2 bottom left). This is not surprising given their poor performance on ImageNet. Jigsaw and RotNet are the only SSL models in our analysis that are non-Siamese, i.e., they were not trained by connecting two augmented views of the same image. For Siamese networks, however, ImageNet performance does not correspond to alignment with human judgments. SimCLR (Chen et al., 2020) and MoCo-v2 (He et al., 2020), both trained with a contrastive learning objective, achieve higher zero-shot odd-one-out accuracy than Barlow Twins (Zbontar et al., 2021), SwAV (Caron et al., 2020), and VICReg (Bardes et al., 2022)—of which all were trained with a non-contrastive learning objective—although their ImageNet performances are reversed. This indicates that contrasting positive against negative examples rather than using positive examples only improves alignment with human similarity judgments.

Model capacity Whereas one typically observes a positive correlation between model capacity and task performance in computer vision (Tan & Le, 2019; Kolesnikov et al., 2020; Zhai et al., 2022), we observe no relationship between model parameter count and odd-one-out accuracy (Figure 2 bottom right). Thus, scaling model width/depth alone appears to be ineffective at improving alignment.

4.2 Consistency of results across different datasets

Although the multi-arrangement task is quite different from the triplet odd-one-out task, we observe similar results for both human similarity judgment datasets that leverage this task (see Figure 3). Again, ImageNet accuracy is not correlated with the degree of human alignment, and objective function and training data, but not architecture or model size, have a substantial impact.

4.3 How much alignment can a linear probe recover?

We next investigate human alignment of neural network representations after linearly transforming the representations to improve odd-one-out accuracy, as described in §3. In addition to evaluating probing odd-one-out accuracies, we perform RSA after applying the transformation matrix obtained from linear probing on the triplet odd-one-out task to a model's raw representation space. Note that the linear probe was trained exclusively on a subset of triplets from Hebart et al. (2020) (see Appendix E), without access to human responses from the two other human similarity judgments datasets (Cichy et al., 2019; King et al., 2019).

Across all three datasets, we observe that the transformation matrices obtained from linear probing substantially improve the degree of alignment with human similarity judgments. The probing odd-one-out accuracies are correlated with the zero-shot odd-one-out accuracies for both the embedding (Figure 4 left; ρ = 0.774) and the logit layers (Figure E.1; ρ = 0.880). Similarly, we observe a strong correlation between the human alignment of raw and transformed representation spaces for the embedding layer for both multi-arrangement task datasets from Cichy et al. (2019) (Figure 4 center; ρ = 0.749) and King et al. (2019) (Figure 4 right; ρ = 0.519) respectively. After applying the transformation matrices to neural nets' representations, we find that image/text models and ViT-G /14 JFT are better aligned than ImageNet or EcoSet models for all datasets and metrics.

As we discuss further in Appendix E, the relationship between probing odd-one-out accuracy and ImageNet accuracy is generally similar to the relationship between zero-shot odd-one-out accuracy and ImageNet accuracy. The same holds for the relationship between Spearman's ρ and ImageNet accuracy and Spearman's ρ (+ transform) and ImageNet accuracy. The correlation between ImageNet accuracy and probing odd-one-out accuracy remains weak ($r = 0.222$). Probing reduces the variance in odd-one-out accuracy or Spearman's ρ among networks trained with different loss functions, self-supervised learning methods, and image/text models, yet we still fail to see improvements in probing accuracy with better-performing architectures or larger model capacities. However,

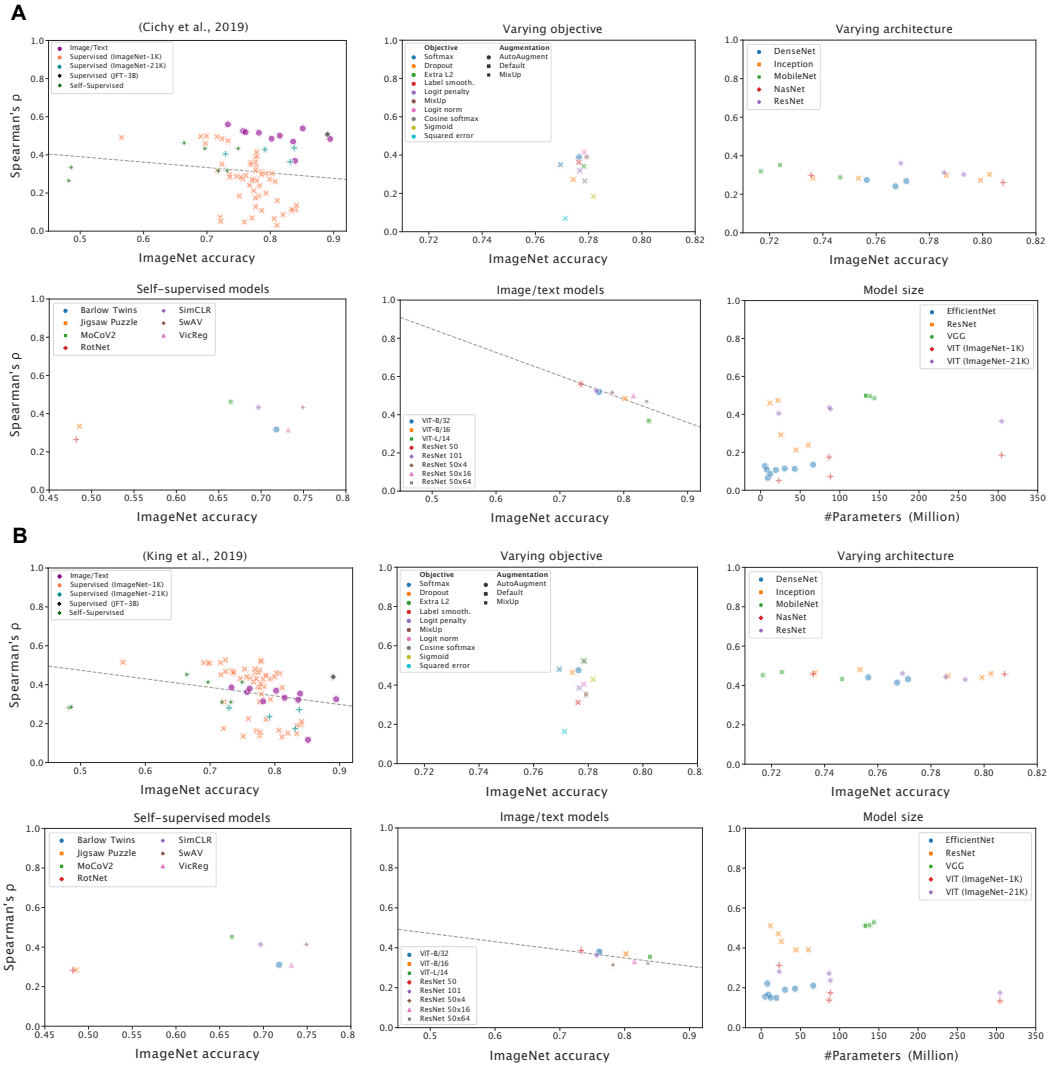


Figure 3: Spearman correlation between human and neural network representational similarity matrices is not correlated with ImageNet accuracy for ImageNet models and is negatively correlated for image/text models. Alignment varies with training objective but not with model architecture or number of parameters for both similarity judgment datasets (Cichy et al., 2019; King et al., 2019). See caption of Figure 2 for further description of panels. Diagonal lines indicate least-squares fits.

whereas image/text models exhibit a negative correlation between ImageNet accuracy and zero-shot odd-one-out accuracy is negative in Figures 2 and 3, the correlation between ImageNet accuracy and probing odd-one-out accuracy is small but positive.

Interestingly, for EcoSet models, transformation matrices do not improve alignment as much as they do for architecturally identical ImageNet models. Although one goal of EcoSet was to provide data that yields better alignment with human perception than ImageNet (Mehrer et al., 2021), we find that models trained on EcoSet are less aligned with human similarity judgments than ImageNet models.

4.4 How well do pretrained neural nets represent human concepts?

Below, we examine zero-shot and linear probing odd-one-out accuracies for individual human concepts. To investigate how well neural nets represent these concepts, we filter the original dataset $\mathcal{D}$ to produce a new dataset $\mathcal{D}^*$ containing only triplets correctly predicted by VICE. Thus, the best attainable odd-one-out accuracy for any model is 1 as opposed to the upper-bound of 0.6722 for the full data. We further partition $\mathcal{D}^*$ 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, $j = \arg\max_{j'} x_{a,j'} + x_{b,j'}$.

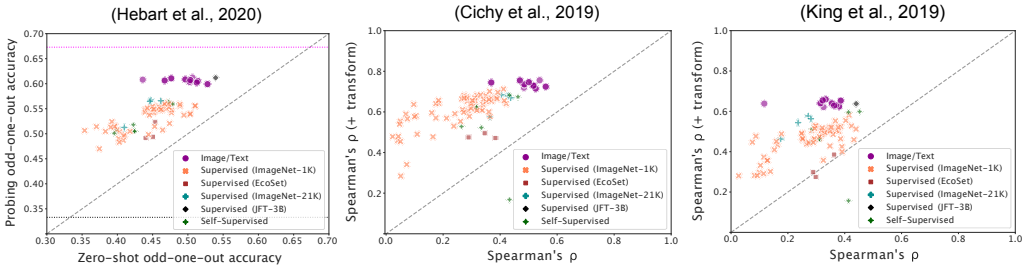


Figure 4: Left panel: Zero-shot and probing odd-one-out accuracies for the embedding layer of all neural nets. Right panels: Spearman rank correlation coefficients with and without applying the transformation matrix obtained from linear probing to a model's raw representation space. Dashed lines indicate $y = x$.

4.4.1 Human alignment is concept-specific

In Figure 5, we show zero-shot and linear probing odd-one-out accuracies for a subset of three of the 45 VICE dimensions for a large subset of the models listed in Table B.1. Zero-shot and probing odd-one-out accuracies for a larger set of dimensions can be found in Appendix J. Since the dimensions found for THINGS are similar both in visual appearance and in their number between Muttenthaler et al. (2022) and Hebart et al. (2020), we infer a labeling of the human dimensions from Hebart et al. (2020) who have evidenced the interpretability of these dimensions through human experiments.

Although models trained on large datasets — image/text models and ViT-G/14 JFT — generally show a higher zero-shot odd-one-out accuracy compared to self-supervised models or models trained on ImageNet, the ordering of models is not entirely consistent across concepts. For dimension 10 (`vehicles`), ResNets trained with a cosine softmax objective were the best zero-shot models, whereas image/text models were among the worst. For dimension 4, an animal-related concept, models pretrained on ImageNet clearly show the worst performance, whereas this concept is well represented in image/text models. Differences in the representation of the animal concept across models are additionally corroborated by the t-SNE visualizations in Appendix H.

Linear probing yields more consistent patterns than zero-shot performance. For almost every human concept, image/text models and ViT-G/14 JFT have the highest per-concept odd-one-out accuracies, whereas AlexNet and EfficientNets have the lowest. This difference is particularly apparent for dimension 17, which summarizes sports-related objects. For this dimension, image/text models and ViT-G/14 JFT perform much better than all remaining models. As shown in Appendix G, even for triplets where VICE predicts that human odd-one-out judgments are very consistent, ImageNet models make a substantial number of errors. By contrast, image/text models and ViT-G/14 JFT achieve a near-zero zero-shot odd-one-out error for these triplets.

4.4.2 Can human concepts be recovered via linear regression?

To further understand the extent to which human concepts can be recovered from neural networks' representation spaces, we perform ℓ_2-regularized linear regression to examine models' ability to predict VICE dimensions. The results from this analysis – which we present in detail in Appendix F – corroborate the findings from §4.3: models trained on image/text data and ViT-G/14 JFT consistently provide the best fit for VICE dimensions, while AlexNet and EfficientNets show the poorest regression performance. We compare odd-one-out accuracies after linear probing and regression respectively. The two performance measures are highly correlated for the embedding ($r = 0.982$) and logit ($r = 0.972$; see Figure F.3) layers, which additionally supports our observations from linear probing. Furthermore, we see that the leading VICE dimensions, which are the most important for explaining human triplet responses, could be fitted with an R^2 score of > 0.7 for most of the models – the quality of the regression fit declines with the importance of a dimension (see Figure F.4).

5 Discussion

In this work, we evaluated the alignment of neural network representations with human concepts spaces through performance in an odd-one-out task and by performing representational similarity analysis. Before discussing our findings we will address limitations of our work. One limitation is

Figure 5: Zero-shot and linear probing odd-one-out accuracies differ across VICE concepts. Results are shown for the embedding layer of all models for three of the 45 VICE dimensions. See Appendix J for additional dimensions. Color-coding is determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).

that we did not consider non-linear transformations. It is possible that simple non-linear transformations could provide better alignment for the networks we investigate. We plan to investigate such transformations further in future work. Another limitation stems from our use of pretrained models for our experiments, since they have been trained with a variety of objectives and regularization strategies. We have mitigated this by comparing controlled subsets of models in Figure 2.

Nevertheless, we can draw the following conclusions from our analyses. 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. These conclusions hold for all three datasets we have investigated, indicating that they are true properties of human/machine alignment rather than idiosyncrasies of the task. Finally, 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 60-61.5%. 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.

ACKNOWLEDGEMENTS

LM, LL, and RV acknowledge support from the Federal Ministry of Education and Research (BMBF) for the Berlin Institute for the Foundations of Learning and Data (BIFOLD) (01IS18037A). This work is in part supported by Google. We thank Klaus-Robert Müller, Robert Geirhos, Katherine Hermann, and Andrew Lampinen for their helpful comments on the manuscript.

REFERENCES

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

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.

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.

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.

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.

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.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2019–2026. IEEE, 2014.

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

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.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

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.

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.

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.

Radoslaw M. Cichy, Nikolaus Kriegeskorte, Kamila M. Jozwik, Jasper J.F. van den Bosch, and Ian Charest. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.03.031.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613. IEEE, 2014.

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.

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.

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.

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.

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.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Generative-Model Based Vision*, 2004.

Kazuyoshi Fukuzawa, Motonobu Itoh, Sumiko Sasanuma, Tsutomu Suzuki, Yoko Fukusako, and Tohru Masui. Internal representations and the conceptual operation of color in pure alexia with color naming defects. *Brain and Language*, 34(1):98–126, 1988. ISSN 0093-934X. doi: https://doi.org/10.1016/0093-934X(88)90126-5.

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.

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.

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.

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.

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.

Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. VISSL, 2021.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Akshay V. Jagadeesh and Justin L. Gardner. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119, 2022. doi: 10.1073/pnas.2115302119.

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.

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.

Marcie L. King, Iris I.A. 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. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.04.079.

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.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 491–507, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58558-7.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019a.

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, 2019b. doi: 10.1109/CVPR.2019.00277.

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.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization*, 2013.

Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2013.06.007.

Nikolaus Kriegeskorte and Marieke Mur. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3, 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00245.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

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

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

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

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.

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

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

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.

Johannes Mehrer, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021. doi: 10.1073/pnas.2011417118.

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.

Grégoire Montavon, Mikio L. Braun, and Klaus-Robert Müller. Kernel Analysis of Deep Networks. *Journal of Machine Learning Research*, 12:2563−2581, 2011. doi: 10.5555/1953048.2078188.

Lukas Muttenthaler and Martin N. Hebart. THINGSvision: A Python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics*, 15:45, 2021. ISSN 1662-5196. doi: 10.3389/fninf.2021.679838.

Lukas Muttenthaler, Charles Yang Zheng, Patrick McClure, Robert A Vandermeulen, Martin N Hebart, and Francisco Pereira. VICE: Variational Interpretable Concept Embeddings. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pp. 722–729. IEEE, 2008.

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.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505. IEEE, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019.

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.

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.

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.

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.

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.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021.

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.

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.

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.

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.

Rocco Robilotto and Qasim Zaidi. Limits of lightness identification for real objects under natural viewing conditions. *Journal of Vision*, 4(9):9–9, 09 2004. ISSN 1534-7362. doi: 10.1167/4.9.9.

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.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

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.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492. IEEE, 2010.

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.

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.

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.

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.

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 EXPERIMENTAL DETAILS

A.1 LINEAR PROBING

Initialization We initialized the transformation matrix $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ used in Eq. 2 with values from a tight Gaussian centered around 0, such that $\boldsymbol{W} \sim \mathcal{N}(0, 10^{-3}\boldsymbol{I})$ at the beginning of the optimization process.

Training We optimized the transformation matrix $\boldsymbol{W}$ via gradient descent, using Adam (Kingma & Ba, 2015) 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 B.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. For most of our evaluation and linear probing experiments, we use `PyTorch` (Paszke et al., 2019).

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.

Algorithm 1 Algorithm for object partitioning during k-fold CV

Input: $(\mathcal{D}, m)$ ▷ Here, $\mathcal{D} \coloneqq (\{a_s, b_s\}, \{i_s, j_s, k_s\})_{s=1}^{n}$ 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
  $\mathbb{O}_{\text{test}} \coloneqq [m] \setminus \mathbb{O}_{\text{train}}$ ▷ Test objects are the remaining objects
  $\mathcal{D}_{\text{train}} \coloneqq \{\}$ ▷ Initialize an empty set for the train triplets
  $\mathcal{D}_{\text{test}} \coloneqq \{\}$ ▷ 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 $(\text{len}(\text{set}(\text{assignments})) \neq 1)$ **then**
      continue ▷ If not all objects in a triplet belong to the same set of objects, discard triplet
    else
      assignment $\triangleq \text{pop}(\text{set}(\text{assignments}))$ ▷ Get object set assignment of current triplet
      if (assignment **is** "train") **then**
        $\mathcal{D}_{\text{train}} \coloneqq \mathcal{D}_{\text{train}} \cup \mathcal{D}_s$ ▷ Assign current triplet to the train set
      else
        $\mathcal{D}_{\text{test}} \coloneqq \mathcal{D}_{\text{test}} \cup \mathcal{D}_s$ ▷ 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

Note that the number of train objects that are 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 the objects in the train and test triplets are 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 $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1\}$.

A.2 TEMPERATURE SCALING

It is widely known that classifiers trained to minimize cross-entropy tend to be overconfident in their predictions (Szegedy et al., 2016; Guo et al., 2017; Roelofs et al., 2022), which is in stark contrast to the high-entropy predictions of VICE. For this purpose, we performed temperature scaling (Guo et al., 2017) 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 Eq. 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,

- 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) (Guo et al., 2017)

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. (2020). For a given triplet $\{i,j,k\}$ and similarity matrix $\boldsymbol{S}$ we define confidence as

$$\operatorname{conf}(\{i,j,k\},\boldsymbol{S}) \coloneqq \max_{\{a,b\}\subset\{i,j,k\}} p(\{a,b\} \mid \{i,j,k\},\boldsymbol{S}).$$

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

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

For a similarity matrix, $\boldsymbol{S}$, and a set of triplets with responses, $\mathcal{D}' \subset \mathcal{D}$, we define $\operatorname{acc}(\mathcal{D}', \boldsymbol{S})$ 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 $\operatorname{conf}(\mathcal{D}')$ to be the average confidence over that set (triplet responses are simply ignored). The ECE is defined as

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

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 $\tau \in \{1\cdot10^0, 7.5\cdot10^{-1}, 5\cdot10^{-1}, 2.5\cdot10^{-1}, 1\cdot10^{-1}, 7.5\cdot10^{-2}, 5\cdot10^{-2}, 2.5\cdot10^{-2}, 1\cdot10^{-2}, 7.5\cdot10^{-3}, 5\cdot10^{-3}, 2.5\cdot10^{-3}, 1\cdot10^{-3}, 5\cdot10^{-4}, 1\cdot10^{-4}, 5\cdot10^{-5}, 1\cdot10^{-5}\}$.

A.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 Appendix A.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 Eq. 3 using `RidgeCV` from Pedregosa et al. (2011). We performed a grid search over α, where $\alpha \in \{0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 1000000\}$.

B MODELS

First, we evaluate supervised models trained on ImageNet (Russakovsky et al., 2015), such as AlexNet (Krizhevsky, 2014), various VGGs (Simonyan & Zisserman, 2015), ResNets (He et al., 2016), EfficientNets (Tan & Le, 2019), ResNext models (Xie et al., 2017), and Vision Transformers (ViTs) trained on ImageNet-1K (Dosovitskiy et al., 2021) or ImageNet-21K (Steiner et al., 2022) respectively. Second, we analyze recent state-of-the-art models trained on image/text data, CLIP-RN & CLIP-ViT (Radford et al., 2021), ALIGN (Jia et al., 2021) and BASIC-L (Pham et al., 2022). Third, we evaluate self-supervised models that were trained with a contrastive learning objective such as SimCLR (Chen et al., 2020) and MoCo-v2 (He et al., 2020), recent SSL models that were trained with a non-contrastive learning objective (no negative examples), BarlowTwins (Zbontar et al., 2021), SwAV (Caron et al., 2020), and VICReg (Bardes et al., 2022), as well as earlier SSL, non-Siamese models, Jigsaw (Noroozi & Favaro, 2016), and RotNet (Gidaris et al., 2018). All self-supervised models have a ResNet-50 architecture. For SimCLR, MoCo-v2, Jigsaw and RotNet, we leverage model weights from the VISSL library (Goyal et al., 2021). For the other models we use weights from their official GitHub repositories. Last, we evaluate the largest available ViT (Zhai et al., 2022), trained on the proprietary JFT-3B image classification dataset, which consists of approximately three billion images belonging to approximately 30,000 classes (Zhai et al., 2022). See Table B.1 for further details regarding the models evaluated.

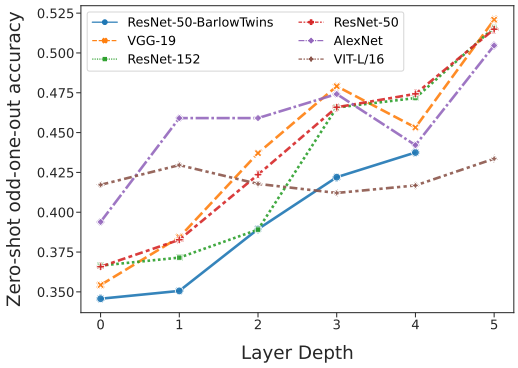


Figure B.1: Zero-shot odd-one-out accuracy for different layers for a subset of selected models.

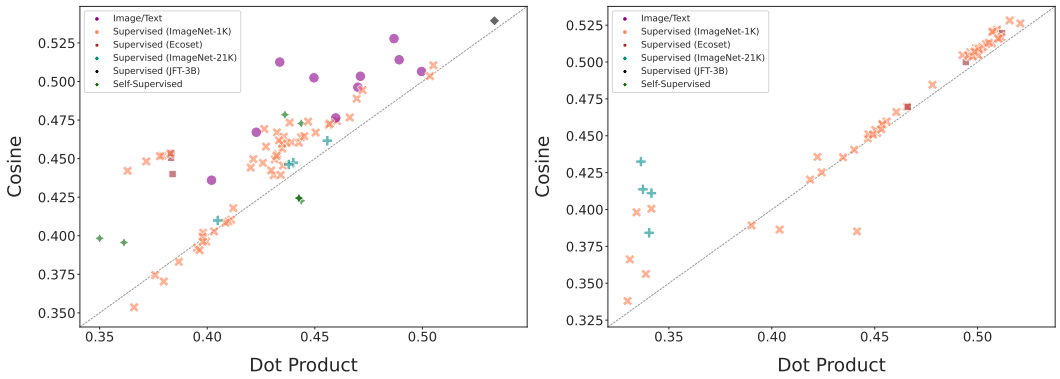


Figure B.2: Zero-shot odd-one-out accuracy using the cosine similarity nearly always is better than using the dot product as a similarity measure.

Model	Source	Architecture	Dataset	Objective	ImageNet	Zero-Shot	Probing
AlexNet	(Krizhevsky, 2014)	AlexNet	ImageNet-1K	Supervised (softmax)	56.52%	50.47%	53.84%
AlexNet	(Muttenthaler & Hebart, 2021)	AlexNet	Ecoset	Supervised (softmax)	-	50.00%	51.30%
ALIGN	(Jia et al., 2021)	EfficientNet	ALIGN dataset	Image/Text (contr.)	85.11%	43.60%	60.81%
Basic-L	(Pham et al., 2022)	CoAtNet	ALIGN + JFT-5B	Image/Text (contr.)	89.45%	50.65%	61.24%
CLIP RN101	(Radford et al., 2021)	ResNet	CLIP dataset	Image/Text (contr.)	75.70%	51.26%	60.22%
CLIP RN50	(Radford et al., 2021)	ResNet	CLIP dataset	Image/Text (contr.)	73.30%	52.78%	59.92%
CLIP RN50x16	(Radford et al., 2021)	ResNet	CLIP dataset	Image/Text (contr.)	81.50%	49.63%	60.86%
CLIP RN50x4	(Radford et al., 2021)	ResNet	CLIP dataset	Image/Text (contr.)	78.20%	50.24%	60.38%
CLIP RN50x64	(Radford et al., 2021)	ResNet	CLIP dataset	Image/Text (contr.)	83.60%	47.64%	61.07%
CLIP ViT-B/16	(Radford et al., 2021)	ViT	CLIP dataset	Image/Text (contr.)	80.20%	50.34%	60.72%
CLIP ViT-B/32	(Radford et al., 2021)	ViT	CLIP dataset	Image/Text (contr.)	76.10%	51.41%	60.54%
CLIP ViT-L/14	(Radford et al., 2021)	ViT	CLIP dataset	Image/Text (contr.)	83.90%	46.71%	60.64%
DenseNet-121	(Kornblith et al., 2019b)	DenseNet	ImageNet-1K	Supervised (softmax)	75.64%	50.37%	55.18%
DenseNet-169	(Kornblith et al., 2019b)	DenseNet	ImageNet-1K	Supervised (softmax)	76.73%	50.52%	55.36%
DenseNet-201	(Kornblith et al., 2019b)	DenseNet	ImageNet-1K	Supervised (softmax)	77.14%	50.45%	55.37%
EfficientNet B0	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	77.69%	45.42%	50.82%
EfficientNet B1	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	79.84%	45.08%	51.30%
EfficientNet B2	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	80.61%	43.23%	49.33%
EfficientNet B3	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	82.01%	39.94%	50.79%
EfficientNet B4	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	83.38%	38.52%	50.65%
EfficientNet B5	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	83.44%	45.10%	51.47%
EfficientNet B6	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	84.01%	45.97%	51.56%
EfficientNet B7	(Tan & Le, 2019)	EfficientNet	ImageNet-1K	Supervised (softmax)	84.12%	45.77%	52.41%
Inception-RN V2	(Kornblith et al., 2019b)	Inception	ImageNet-1K	Supervised (softmax)	80.26%	51.31%	55.78%
Inception-V1	(Kornblith et al., 2019b)	Inception	ImageNet-1K	Supervised (softmax)	73.63%	50.43%	54.97%
Inception-V2	(Kornblith et al., 2019b)	Inception	ImageNet-1K	Supervised (softmax)	75.34%	50.70%	54.97%
Inception-V3	(Kornblith et al., 2019b)	Inception	ImageNet-1K	Supervised (softmax)	78.64%	51.59%	55.84%
Inception-V4	(Kornblith et al., 2019b)	Inception	ImageNet-1K	Supervised (softmax)	79.92%	51.58%	55.47%
MobileNet-V1	(Kornblith et al., 2019b)	MobileNet	ImageNet-1K	Supervised (softmax)	72.39%	50.70%	54.98%
MobileNet-V2	(Kornblith et al., 2019b)	MobileNet	ImageNet-1K	Supervised (softmax)	71.67%	50.45%	55.17%
MobileNet-V2 (1.4)	(Kornblith et al., 2019b)	MobileNet	ImageNet-1K	Supervised (softmax)	74.66%	50.67%	55.11%
NASNet-L	(Kornblith et al., 2019b)	NASNet	ImageNet-1K	Supervised (softmax)	80.77%	51.68%	55.78%
NASNet-Mobile	(Kornblith et al., 2019b)	NASNet	ImageNet-1K	Supervised (softmax)	73.57%	51.23%	55.48%
RN-50-BarlowTwins	(Zbontar et al., 2021)	ResNet	ImageNet-1K	Self-sup. (non-contr.)	71.80%	42.44%	50.50%
RN-50-Jigsaw	(Goyal et al., 2021)	ResNet	ImageNet-1K	Self-sup. (non-Siamese)	48.57%	39.56%	50.11%
RN-50-MoCo-v2	(Goyal et al., 2021)	ResNet	ImageNet-1K	Self-sup. (contr.)	66.40%	47.85%	55.94%
RN-50-RotNet	(Goyal et al., 2021)	ResNet	ImageNet-1K	Self-sup. (non-Siamese)	54.93%	39.83%	50.82%
RN-50-SimCLR	(Goyal et al., 2021)	ResNet	ImageNet-1K	Self-sup. (contr.)	69.68%	47.28%	56.37%
RN-50-SWAV	(Caron et al., 2020)	ResNet	ImageNet-1K	Self-sup. (non-contr.)	74.92%	42.27%	51.79%
RN-50-VICReg	(Bardes et al., 2022)	ResNet	ImageNet-1K	Self-sup. (non-contr.)	73.20%	42.44%	50.50%
RN-18	(He et al., 2016)	ResNet	ImageNet-1K	Supervised (softmax)	69.76%	51.05%	54.97%
RN-34	(He et al., 2016)	ResNet	ImageNet-1K	Supervised (softmax)	73.31%	50.93%	55.30%
RN-50	(He et al., 2016)	ResNet	ImageNet-1K	Supervised (softmax)	80.86%	49.44%	53.72%
RN-101	(He et al., 2016)	ResNet	ImageNet-1K	Supervised (softmax)	81.89%	47.40%	52.06%
RN-152	(He et al., 2016)	ResNet	ImageNet-1K	Supervised (softmax)	82.28%	46.61%	50.74%
RN-101	(Kornblith et al., 2019b)	ResNet	ImageNet-1K	Supervised (softmax)	78.56%	50.86%	55.95%
RN-152	(Kornblith et al., 2019b)	ResNet	ImageNet-1K	Supervised (softmax)	79.29%	50.95%	56.04%
RN-50	(Kornblith et al., 2019b)	ResNet	ImageNet-1K	Supervised (softmax)	76.93%	51.02%	56.05%
RN-50 (dropout)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.42%	51.26%	55.40%
RN-50 (extra WD)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.82%	52.62%	56.16%
RN-50 (label smoothing)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.63%	45.78%	55.52%
RN-50 (logit penality)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.67%	47.67%	54.21%
RN-50 (mixup)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.92%	46.70%	56.29%
RN-50 (AutoAugment)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax)	77.64%	50.67%	56.10%
RN-50 (logit norm)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.83%	51.05%	55.63%
RN-50 (cosine softmax)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax+)	77.86%	52.82%	56.73%
RN-50 (sigmoid)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (sigmoid)	78.18%	44.72%	55.34%
RN-50 (softmax)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (softmax)	76.94%	50.89%	55.97%
RN-50 (squared error)	(Kornblith et al., 2021)	ResNet	ImageNet-1K	Supervised (sq. error)	77.13%	41.04%	49.87%
RN-50	(Muttenthaler & Hebart, 2021)	ResNet	Ecoset	Supervised (softmax)	-	46.96%	50.63%
ResNeXt-50 32x4d	(Xie et al., 2017)	ResNeXt	ImageNet-1K	Supervised (softmax)	81.20%	46.97%	51.44%
ResNeXt-101 32x8d	(Xie et al., 2017)	ResNeXt	ImageNet-1K	Supervised (softmax)	81.89%	48.46%	50.82%
VGG-11	(Simonyan & Zisserman, 2015)	VGG	ImageNet-1K	Supervised (softmax)	69.02%	52.04%	55.91%
VGG-13	(Simonyan & Zisserman, 2015)	VGG	ImageNet-1K	Supervised (softmax)	69.93%	52.24%	55.84%
VGG-16	(Simonyan & Zisserman, 2015)	VGG	ImageNet-1K	Supervised (softmax)	71.59%	52.09%	55.86%
VGG-19	(Simonyan & Zisserman, 2015)	VGG	ImageNet-1K	Supervised (softmax)	72.38%	52.09%	55.86%
VGG-16	(Muttenthaler & Hebart, 2021)	VGG	Ecoset	Supervised (softmax)	-	51.96%	53.58%
ViT-B/16 I1K	(Steiner et al., 2022)	ViT	ImageNet-1K	Supervised (sigmoid)	77.66%	38.31%	50.48%
ViT-B/16 I21K	(Steiner et al., 2022)	ViT	ImageNet-21K	Supervised (sigmoid)	83.77%	44.62%	56.49%
ViT-B/32 I1K	(Steiner et al., 2022)	ViT	ImageNet-1K	Supervised (sigmoid)	72.08%	37.45%	46.99%
ViT-B/32 I21K	(Steiner et al., 2022)	ViT	ImageNet-21K	Supervised (sigmoid)	79.16%	44.74%	56.78%
ViT-L/16 I1K	(Steiner et al., 2022)	ViT	ImageNet-1K	Supervised (sigmoid)	75.11%	37.03%	51.42%
ViT-L/16 I21K	(Steiner et al., 2022)	ViT	ImageNet-21K	Supervised (sigmoid)	83.13%	40.99%	51.27%
ViT-S/32 I1K	(Steiner et al., 2022)	ViT	ImageNet-1K	Supervised (sigmoid)	72.18%	40.28%	48.31%
ViT-S/32 I21K	(Steiner et al., 2022)	ViT	ImageNet-21K	Supervised (sigmoid)	72.93%	46.16%	56.60%
ViT-G/14 JFT	(Zhai et al., 2022)	ViT	JFT-3B	Supervised (sigmoid)	89.01%	53.94%	61.18%
ViT-B-16	(Dosovitskiy et al., 2021)	ViT	ImageNet-1K	Supervised (softmax)	81.07%	42.02%	47.89%
ViT-B-32	(Dosovitskiy et al., 2021)	ViT	ImageNet-1K	Supervised (softmax)	75.91%	42.52%	48.69%

Table B.1: Pretrained neural networks that we considered in our analyses. "RN" = ResNet, "Self-sup." = self-supervised, "contr." = contrastive. "RN-50 (extra WD)" is a ResNet-50 with higher weight decay on the final network layer. The "ImageNet" column contains the accuracy on the ImageNet dataset. The "Zero-Shot" column contains the THINGS zero-shot odd-one-out accuracy of the better of the embedding and logits layer. The "Probing" column contains the THINGS probing odd-one-out accuracy of the embedding layer.

Figure B.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 B.1. Figure B.2 compares the odd-one-out accuracy of using dot product versus cosine similarity and shows that cosine similarity generally yields better alignment.

C CIFAR-100 TRIPLET TASK

Figure C.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 rightmost image is the odd-one-out.

In a similar vein to the THINGS triplet task, we constructed a reference triplet task from the CIFAR-100 dataset (Krizhevsky & Hinton, 2009). 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 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 C.1 shows an example triplet for this task.

We find that ImageNet accuracy is highly correlated with odd-one-out accuracy for the CIFAR-100 coarse task (see Figure C.2; $r = 0.70$), which is in stark contrast to its correlation with accuracy on human odd-one-out judgments, which is significantly weaker (see Figure 2).

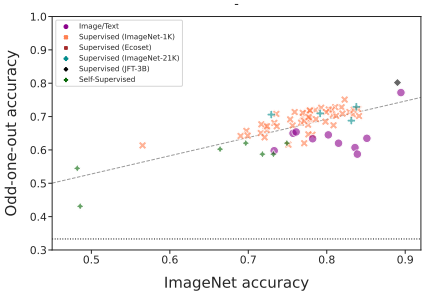


Figure C.2: Zero-shot odd-one-out accuracy on a triplet task based on CIFAR-100 coarse exhibits a strong correlation with ImageNet accuracy. Diagonal line indicates a least-squares fit.

The main reason for constructing this task was to examine whether or not any findings from comparing human to neural network responses for the THINGS triplet odd-one-out task can be attributed to the nature of the triplet task. Instead of using the CIFAR-100 class labels, we specifically used the coarse super-classes that are possibly comparable to higher-level concepts that are relevant to human similarity judgments on the THINGS odd-one-out task. Hebart et al. (2020) and Muttenthaler et al. (2022) have shown that humans only use a small set of concepts for making similarity judgments in the triplet odd-one-out task. These concept representations are sparse representations. That is, there are only k objects that are important for a concept, where $k \ll m$. Recall that m denotes the number of objects in the data (e.g., 1854 for THINGS). The importance of objects is defined in Hebart et al. (2020) and Muttenthaler et al. (2022). Similarly, the coarse super-classes in CIFAR-100 are sparse. Although the CIFAR-100 triplet task may be different, we believe that additionally testing models on this task is one reasonable way to figure out whether findings (e.g., the correlation of ImageNet accuracy with triplet odd-one-out accuracy) are attributable to the nature of the triplet task itself rather than to variables related to alignment.

D TRANSFERABILITY OF PENULTIMATE LAYER VS. LOGITS

In Figure E.2, we show that the logits typically outperform the penultimate layer in terms of zero-shot odd-one-out accuracy. In this section, we perform a similar comparison of the performance of the penultimate layer and logits in the context of transfer learning. We find that, contrary to odd-one-out accuracy, transfer learning accuracy is consistently highest in the penultimate layer.

Following Kornblith et al. (2019b), we report the accuracy of ℓ_2-regularized multinomial logistic regression classifiers on 12 datasets: Food-101 dataset (Bossard et al., 2014), CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009), Birdsnap (Berg et al., 2014), the SUN397 scene dataset (Xiao et al., 2010), Stanford Cars (Krause et al., 2013), FGVC Aircraft (Maji et al., 2013), the PASCAL VOC 2007 classification task (Everingham et al., 2010), the Describable Textures Dataset (DTD) (Cimpoi et al., 2014), Oxford-IIIT Pets (Parkhi et al., 2012), Caltech-101 (Fei-Fei et al., 2004), and Oxford 102 Flowers (Nilsback & Zisserman, 2008). We use representations of the 16 models previously studied by Kornblith et al. (2019b) (see Table B.1), and follow the same procedure for training and evaluating the classifiers.

Results are shown in Figure D.1. For nearly all models and transfer datasets, the penultimate layer provides better representations for linear transfer than the logits layer.

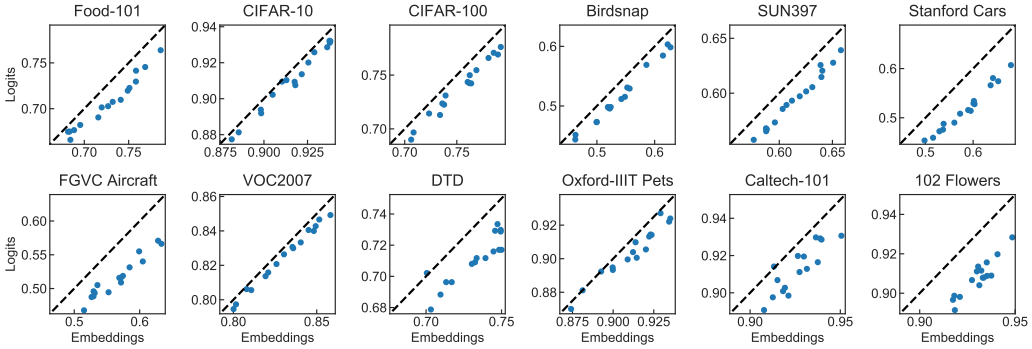


Figure D.1: Penultimate layer embeddings consistently offer higher transfer accuracy than the logits layer. Points reflect the accuracy of a multinomial logistic regression classifier trained on the penultimate layer embeddings (x-axis) or logits (y-axis) of the 16 models from Kornblith et al. (2019b), which were all trained with vanilla softmax cross-entropy. Dashed lines reflect $y = x$.

E LINEAR PROBING

In Figure E.1 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 and zero-shot odd-one-out accuracies.

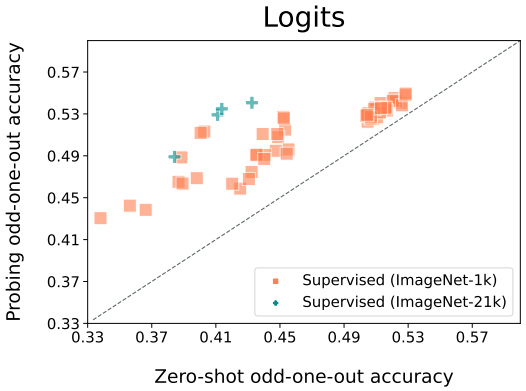


Figure E.1: Probing odd-one-out accuracy as a function of zero-shot odd-one-out accuracy for the logits layer of all ImageNet models in Table B.1.

In the top left panel of Figure E.3, we show probing odd-one-out accuracy as a function of ImageNet accuracy for all models in Table B.1. Similarly to the findings depicted in the top left panel of Figure 2, we observe a low Pearson correlation coefficient ($r = 0.213$). The remaining panels of Figure E.3 visualize probing odd-one-out accuracy as a function of ImageNet accuracy for the same subsets of models as shown in Figure 2. Again, the relationships are similar to those observed for zero-shot odd-one-out accuracy in Figure 2.

Probing accuracies show less variability than zero-shot accuracies among the networks trained with the different loss functions from Kornblith et al. (2021) (Figure 2 top center), although cosine softmax, which performed best for the zero-shot setting, is also among the best-performing models here. Moreover, probing reduced the differences between different Siamese self-supervised learning models (Figure 2 bottom left), although Siamese models still performed substantially better than the non-Siamese models. Yet, as in our analysis of zero-shot accuracy (§4.1), architecture (Figure 2 top right) or model size (Figure 2 bottom right) did not affect odd-one-out accuracy. These findings hold across every dataset we have considered in our analyses (see Figure E.4).

Whereas the logits often achieve a higher degree of alignment with human similarity judgments than the penultimate layer for zero-shot performance, alignment is nearly always highest for the penultimate layer after applying $\boldsymbol{W}$ — the transformation matrix learned during linear probing — to a model's raw representation space and then performing the odd-one-out task or RSA (see Figure E.2).

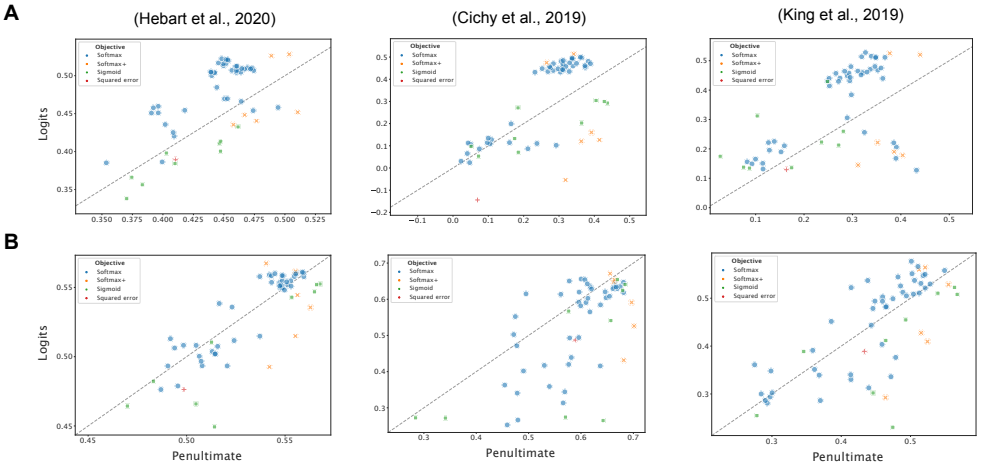


Figure E.2: Performance of the logits vs. penultimate layer for all models across all three datasets without (top row) and with (bottom row) applying the transformation matrix obtained from linear probing to a model's raw representation space. x-axis and y-axis represent odd-one-out accuracy for Hebart et al. (2020) and Spearman's ρ for Cichy et al. (2019) and King et al. (2019). Networks are colored by their loss function. "softmax+" indicates softmax cross-entropy with additional regularization or normalization. Dashed lines indicate $y = x$.

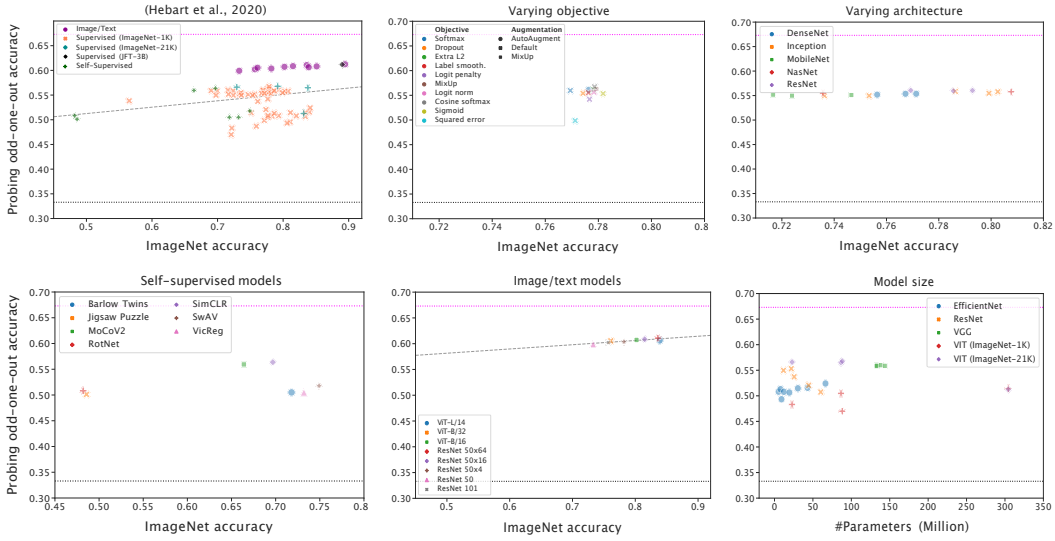


Figure E.3: Probing odd-one-out accuracy as a function of ImageNet accuracy or number of model parameters. **Top left**: Probing odd-one-out accuracies for the embedding layer of all models considered in our analysis. **Top center**: Models have the same architecture (ResNet-50) but were trained with a different objective function (Kornblith et al., 2021). **Top right**: Models were trained with the same objective function but vary in architecture (Kornblith et al., 2019b). **Bottom left**: Performance for different SSL models. **Bottom center**: Different image/text models with their ImageNet accuracies. **Bottom right** A subset of ImageNet models including their number of parameters. Dashed diagonal lines indicate a least-squares fit. Dashed horizontal lines reflect chance-level or ceiling accuracy respectively.

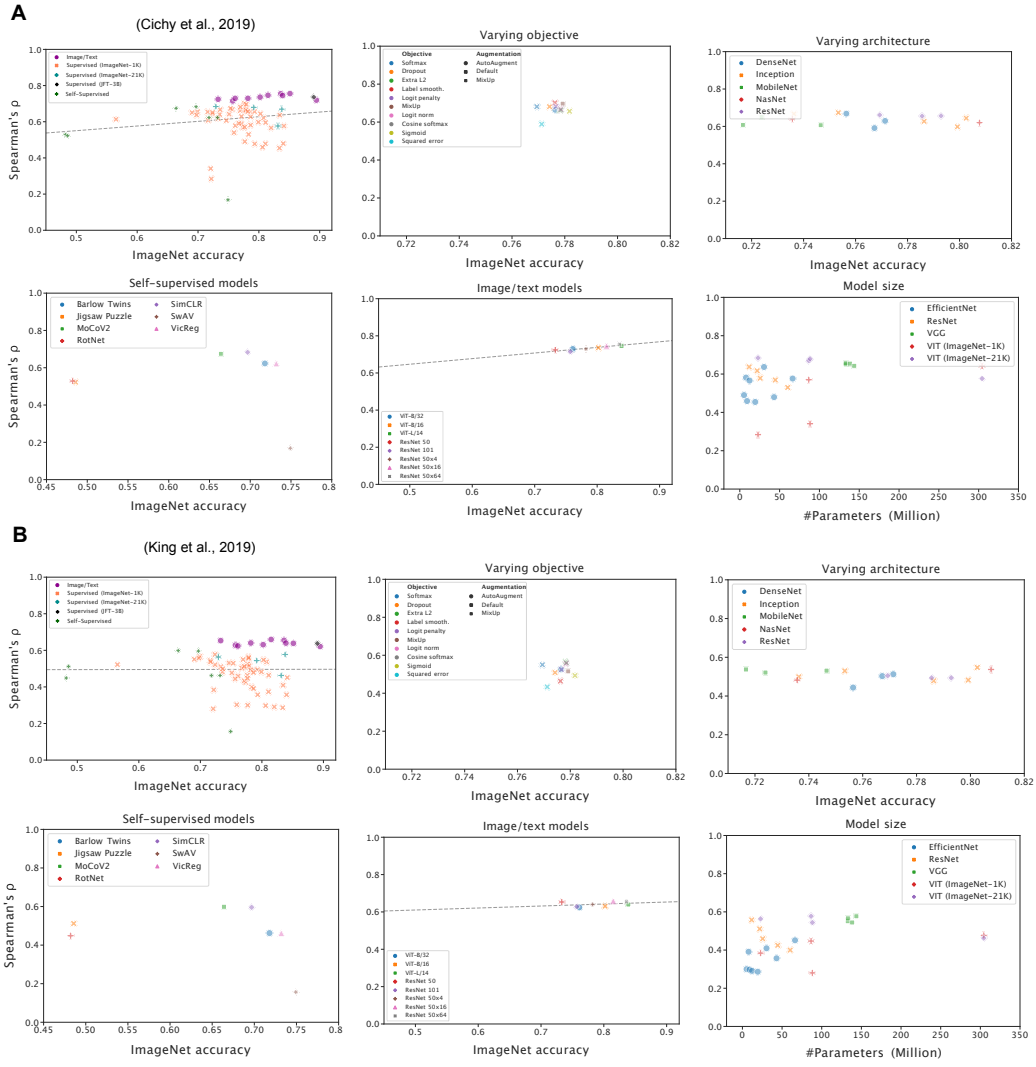


Figure E.4: Spearman rank correlation after applying $\boldsymbol{W}$ to a neural net's representation space is weakly correlated with ImageNet accuracy and varies with training objective but not with model architecture or number of parameters for both human similarity judgment datasets from Cichy et al. (2019) and King et al. (2019) respectively. Diagonal lines indicate a least-squares fit.

F LINEAR REGRESSION

In this section, we elaborate upon the results that we presented in §4.4.2 in more detail. For each of the 45 representation dimensions j from VICE, we minimize the following least-squares objective

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

where $Y_{i,j}$ is the value of the j^{th} VICE dimension for image i, $\boldsymbol{x}_i$ is the neural net representation of image i, and $\alpha_j > 0$ is a regularization hyperparameter. We optimize each dimension separately, selecting α_j via cross validation (see Appendix A.3), and assess the fit in two ways. First, we directly measure the R^2 for held-out images. Second, we evaluate odd-one-out accuracy of the transformed neural net representations using a similarity matrix $\boldsymbol{S}$ with $S_{ij} \coloneqq (\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b})^\top (\boldsymbol{A}\boldsymbol{x}_j + \boldsymbol{b})$, with $\boldsymbol{A}$ and $\boldsymbol{b}$ obtained from Eq. 3 (i.e., *regression odd-one-out accuracy*).

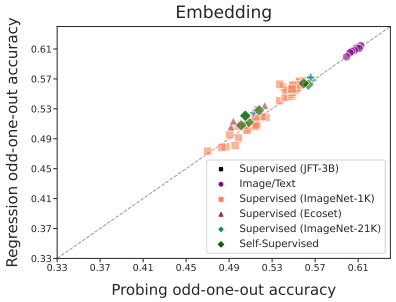


Figure F.1: Regression as a function of probing odd-one-out accuracies for all models in Table B.1.

In Figure F.1, we compare odd-one-out accuracies after linear probing and regression respectively. The two performance measures are highly correlated for the embedding ($r = 0.982$) and logit ($r = 0.972$; see Figure F.3) layers.[2] We provide R^2 values for individual concepts in Appendix F. We observe that the leading VICE dimensions, which are the most important for explaining human triplet responses, could be fitted with an R^2 score of > 0.7 for most of the models – the quality of the regression fit generally declines with the importance of a dimension (see Figure F.4).

We compared zero-shot with regression odd-one-out accuracies (as defined in §4.4.2) for the embedding layer of all models in Table B.1 and observe a strong positive relationship ($r = 0.795$; Figure F.2). In addition, we contrasted regression odd-one-out accuracy between logit and penultimate layers for ImageNet models. The results are consistent with the findings obtained from the linear probing experiments, shown in Figure 4. Moreover, we observe that probing and regression odd-one-out are highly correlated, thus applying similar transformations to a neural net's representation space. This is demonstrated in Figure F.1 for the embedding layer of all models in Table B.1 and in Figure F.3 for the logit layers of the ImageNet models. Figure F.4 shows R^2 scores of the linear regression fit from embedding-layer representations of all models in Table B.1 to the same subset of VICE dimensions that we leveraged for the linear probing experiments in Figure 5.

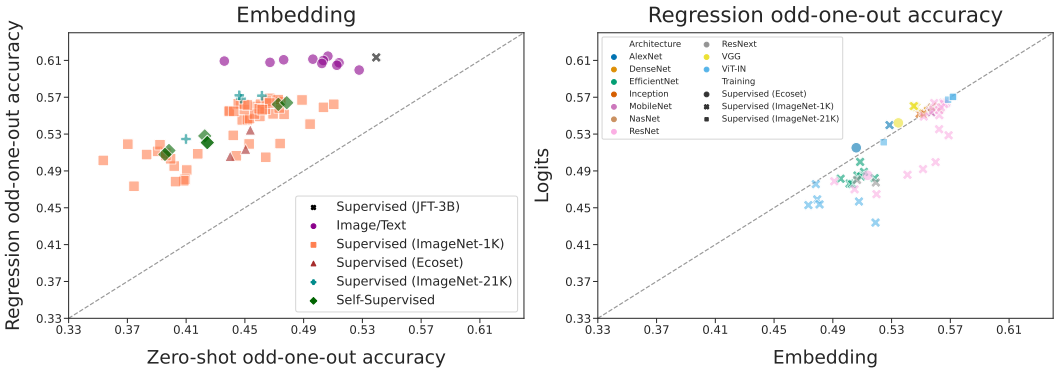


Figure F.2: **Left**: Zero-shot and regression odd-one-out accuracies for the embedding layer of all neural nets. **Right**: Regression odd-one-out accuracy for the embedding and logits layer for all supervised models trained on ImageNet-1K or ImageNet-21K.

[2]Note that odd-one-out accuracies are slightly higher for linear regression (Figure F.1). We hypothesize that this is because VICE is trained on all images, and thus the transformation matrix learned in linear regression has indirect access to the images it is evaluated on, whereas the linear probe has no access to these images (see Appendix A.1).

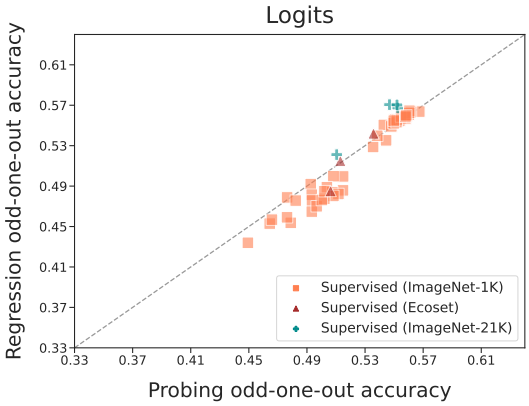


Figure F.3: Regression as a function of probing odd-one-out accuracies for the logits layers of all ImageNet models in Table B.1.

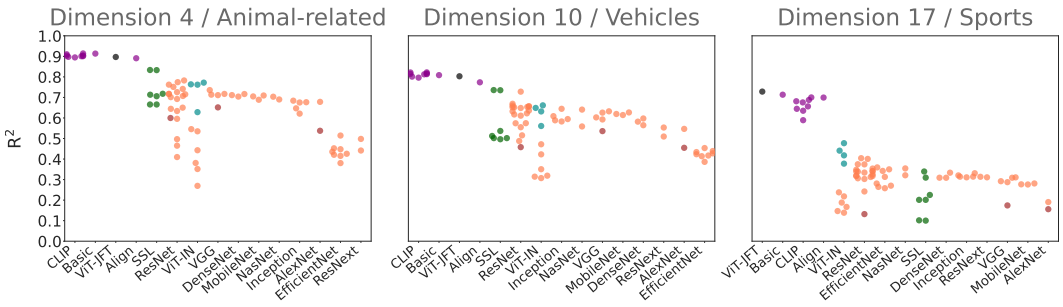


Figure F.4: R^2 scores for all models in Table B.1 after fitting an ℓ_2-regularized linear regression to predict individual VICE dimensions from the embedding-layer representation of the images in THINGS. Color-coding was determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Brown: Supervised (Ecoset). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).

G ENTROPY

Let $\mathbb{A} \coloneqq \{\{i,j)\},\{i,k\},\{j,k\}\}$ be the set of all combinations of pairs in a triplet. The entropy of a triplet can then be written as

$$H(\{i,j,k\}) = -\sum_{\{a,b\}\in\mathbb{A}} \hat{p}(\{a,b\}|\{i,j,k\}) \log \hat{p}(\{x,y\}|\{i,j,k\}),$$

where $\hat{p}(\{a,b\}|\{i,j,k\})$ is derived from the VICE model and is defined precisely in Equation 6 in Muttenthaler et al. (2022).

To understand how aleatoric uncertainty of odd-one-out predictions varies across different models, we calculated the entropy of each triplet in THINGS for every model in Table B.1 and subsequently computed the Pearson correlation coefficient of these per-triplet entropies for every pair of models.

Although models with the same architecture often correlated strongly with respect to their aleatoric uncertainty across triplets, not a single model achieves a strong positive correlation to VICE (see Figure G.1 and Figure G.2 respectively). Interestingly, the choice of the objective function appeared to play a crucial role for the entropy-alignment of a neural net with other neural nets or with VICE.

G.1 HUMAN UNCERTAINTY DETERMINES ODD-ONE-OUT ERRORS

Since VICE provides a probabilistic model of humans' responses on the odd-one-out task, we can use the entropy of a given triplet's probability distribution to infer humans' uncertainty regarding the odd-one-out. In this section, we evaluate a model's odd-one-out classification error as a function of a triplet's entropy. For this analysis, we partitioned the original triplet dataset $\mathcal{D}$ into 11 sets $\mathcal{D}_1,\ldots,\mathcal{D}_{11}$, corresponding to 11 bins. A triplet belongs to a subset $\mathcal{D}_i$ if the entropy of the VICE

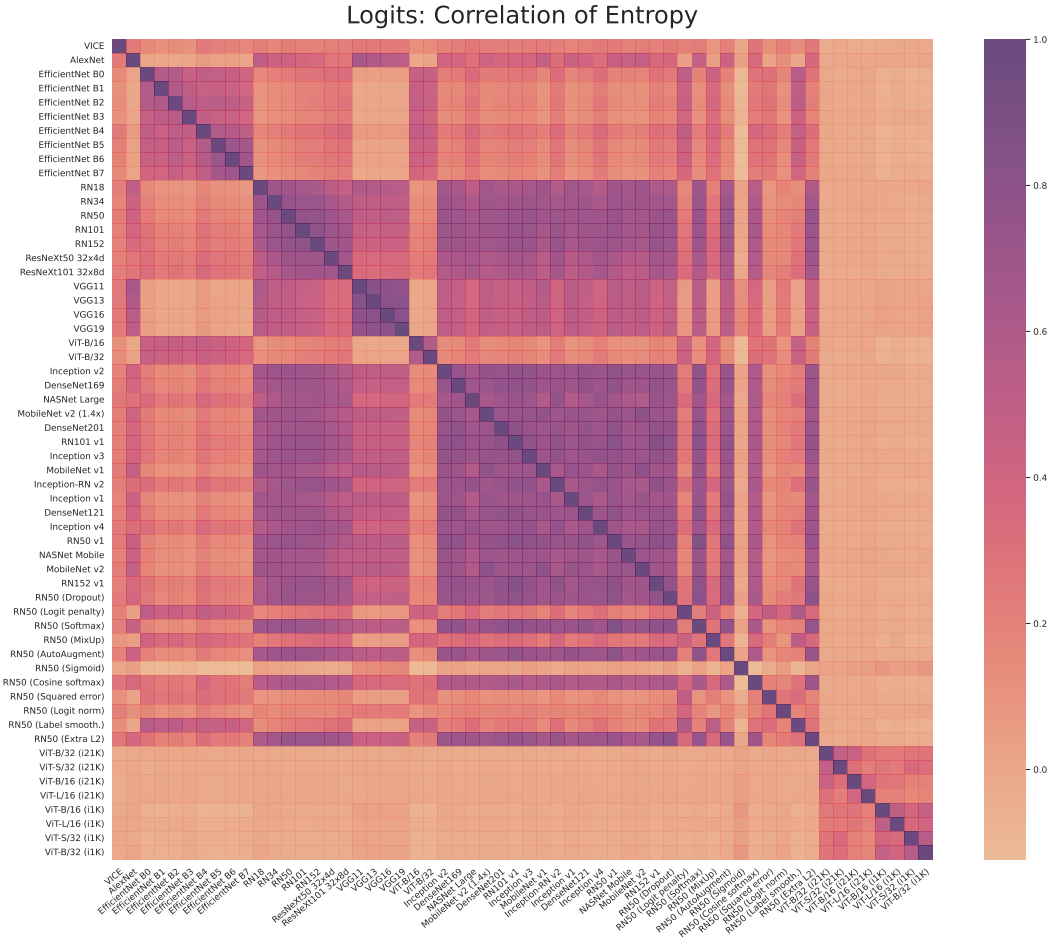


Figure G.1: The correlation coefficient of entropy of the output probabilities for each triplet in THINGS.

output distribution for that triplet falls in between the $(i-1)^{\text{th}}$ and i^{th} bins' boundaries. We define a triplet's entropy as the entropy over the three possible odd-one-out responses, estimated using 50 Monte Carlo samples from VICE (see details in Appendix G). Note that the entropy for a discrete probability distribution with 3 outcomes lies in $[0, \log(3)]$.

Unsurprisingly, all models in Table B.1 yield a high zero-shot odd-one-out classification error for triplets that have high entropy, and all models' error rates increase monotonically as human triplet entropies increase. However, most models make a substantial number of errors even for triplets where entropy is low and thus humans are very certain. We find that VGGs, ResNets, EfficientNets, and ViTs trained on ImageNet-1K or ImageNet-21K and SSL models show a similarly high zero-shot odd-one-out error, between 0.1 and 0.3, for triplets with low entropy, whereas ALIGN and in particular CLIP-RN, CLIP-ViT, BASIC-L and ViT-G/14 JFT achieve a near-zero zero-shot odd-one-out error for the same set of triplets (see Figure G.3).

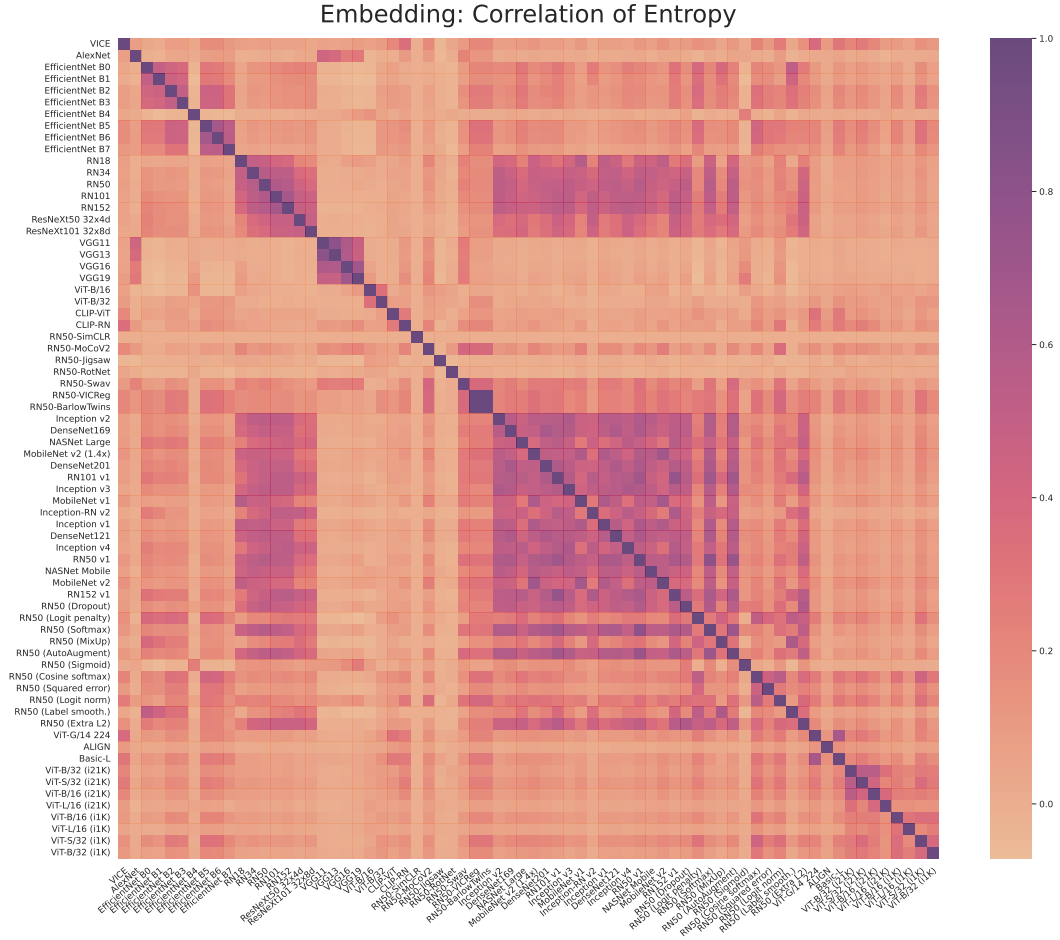


Figure G.2: The correlation coefficient of entropy of the output probabilities for each triplet in THINGS.

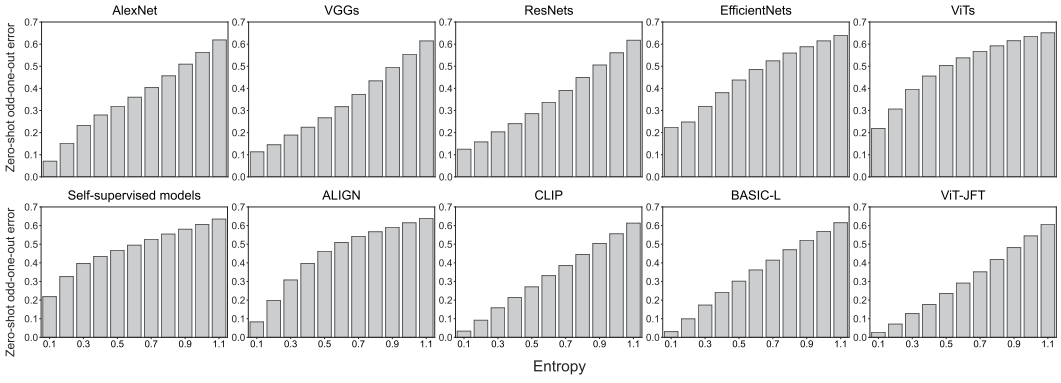
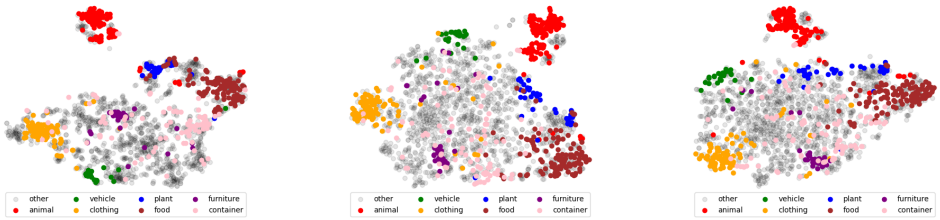


Figure G.3: Zero-shot odd-one-out prediction errors as a function of a triplet's entropy differ across model classes. **Top**: Logits layer of ImageNet supervised models. **Bottom**: Embedding layer of SSL, Image/Text models and ViT-G/14 JFT. Since models with the same architecture, trained on the same data (e.g., ImageNet-1K) with the same objective function, perform very similarly in their odd-one-out choices, we aggregated their predictions and report the average. To isolate architecture and training data from any other potentially confounding variables, we excluded models from Kornblith et al. (2021) when aggregating the ResNet predictions.



(a) VICE (b) CLIP-RN (c) CLIP-ViT

(d) AlexNet (e) VGG-16 (f) ResNet-18

(g) ResNet-152 (h) VIT-B/32 (i) EfficientNet B3

(j) ResNet-50 (SimCLR) (k) ResNet-50 (BarlowTwins) (l) ResNet-50 (Jigsaw)

Figure H.1: t-SNE visualizations for the embedding layer of a subset of the models in Table B.1. Data points are labeled according to higher-level categories provided in the THINGS database (Hebart et al., 2019).

H DIFFERENT CONCEPTS ARE DIFFERENTLY DISENTANGLED IN DIFFERENT REPRESENTATION SPACES

In this section, we show t-SNE (Van der Maaten & Hinton, 2008) visualizations of the embedding layers of a subset of the models in Table B.1. To demonstrate the difference in disentanglement of higher-level concepts - which are provided in the THINGS database (Hebart et al., 2019) - in different representation spaces, we have chosen a representative model for a subset of the architectures and training objectives. Figure H.1 shows that the animal concept is substantially more disentangled from other high-level categories in the representation space of image/text models such as CLIP-RN/CLIP-ViT (Radford et al., 2021) than it is for ImageNet models. The bottom row of Figure H.1 shows that higher-level concepts appear to more distributed and poorly disentangled in Jigsaw, a non-Siamese self-supervised model, compared to SimCLR and BarlowTwins, contrastive and non-contrastive self-supervised models respectively.

I ODD-ONE-OUT AGREEMENTS REFLECT REPRESENTATIONAL SIMILARITY

To understand whether odd-one-out choice agreements between different models reflect similarity of their representation spaces, we compared the agreement in odd-one-out choices between pairs of models with their linear Centered Kernel Alignment (CKA), a widely adopted similarity metric for neural network representations (Kornblith et al., 2019a; Raghu et al., 2021). Let $\boldsymbol{X} \in \mathbb{R}^{m \times p_1}$ and $\boldsymbol{Y} \in \mathbb{R}^{m \times p_2}$ be representations of the same m examples obtained from two neural networks with p_1 and p_2 neurons respectively. Assuming that the column (neuron) means have been subtracted from each representation (i.e., centered representations), linear CKA is defined as

$$\mathrm{CKA}_{\mathrm{linear}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathrm{vec}(\boldsymbol{X}\boldsymbol{X}^\top) \cdot \mathrm{vec}(\boldsymbol{Y}\boldsymbol{Y}^\top)}{\|\boldsymbol{X}\boldsymbol{X}^\top\|_{\mathrm{F}} \|\boldsymbol{Y}\boldsymbol{Y}^\top\|_{\mathrm{F}}}.$$

Intuitively, the representational similarity (Gram) matrices $\boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{Y}\boldsymbol{Y}^\top$ measure the similarity between representations of different examples according to the representations contained in $\boldsymbol{X}$ and $\boldsymbol{Y}$. Linear CKA measures the cosine similarity between these representational similarity matrices after flattening them to vectors.

In Figure I.1 we show heatmaps for both zero-shot odd-one-out agreements and CKA for the same pairs of models. The regression plot shows zero-shot odd-one-out choice agreement between all pairs of models in Table B.1 as a function of their CKA. We observe a high correlation between odd-one-out choice agreements and CKA for almost every model pair. That is, odd-one-out choice agreements appear to reflect similarities of neural network representations.

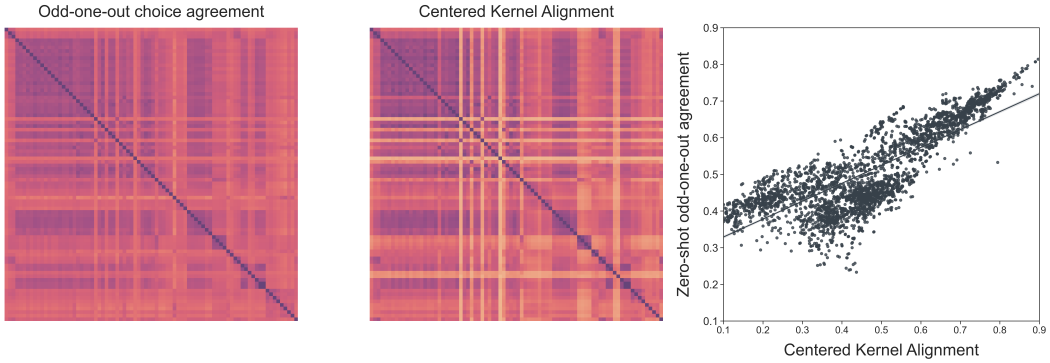


Figure I.1: **Left**: Heatmaps for zero-shot odd-one-out choice agreements and CKA for the same pairs of models. For better readability, we omitted labeling the names of the individual models. **Right**: Zero-shot odd-one-out choice agreements between all pairs of models as a function of CKA plus a linear regression fit.

J HUMAN ALIGNMENT IS CONCEPT SPECIFIC

In Figure J.1 and Figure J.2 we compare zero-shot odd-one-out accuracy and R^2 scores respectively between the best-aligned ImageNet model and the best-aligned overall model. Although ViT-G/14 JFT achieves better alignment with human similarity judgments for most VICE dimensions, the best ImageNet model outperformed ViT-G/14 JFT for a small subset of the concepts, e.g., `tools`, `technology`, `paper`, `liquids`. That is, there appear to be some human concepts which ImageNet models can represent fairly well without an additionally learned linear transformation matrix, even better than a model trained on a larger, more diverse dataset. However, the R^2 scores show a different pattern. In linear regression, representations of ViT-G/14 JFT could clearly be fitted better to the VICE dimensions for every concept compared to the best ImageNet model.

In Figure J.3 we show per-concept zero-shot and probing odd-one-out accuracies for all models in Table B.1 for all 45 VICE dimensions. Whereas zero-shot odd-one-out performances did not show a consistent pattern across the individual dimensions, probing odd-one-out performances clearly demonstrated that image/text models and ViT-G/14 JFT are better aligned with human similarity judgments for almost every concept. However, there are some concepts (e.g., outdoors-related objects/dimension 8; powdery/dimension 22) for which these models are worse aligned than ImageNet models. The difference in human alignment between image/text models plus ViT-G/14 JFT and ImageNet models is largest for royal and sports-related objects - i.e, dimension 7 and 17 respectively

-, where image/text models and ViT-G/14 JFT outperformed ImageNet models by a large margin. The bottom panel of the figure shows R^2 scores of the regression fit for each VICE dimension using the embedding layer representations of all models in Table B.1. We observe that more important concepts are generally easier to recover. Recall that dimensions are numbered according to their importance (Muttenthaler et al., 2022).

In addition, Figure J.4 - Figure J.7 show both zero-shot and probing odd-one-out accuracies for all models in Table B.1 for a larger set of human concepts as we do in the main text, including the 6 most important THINGS images/categories for the dimensions themselves.

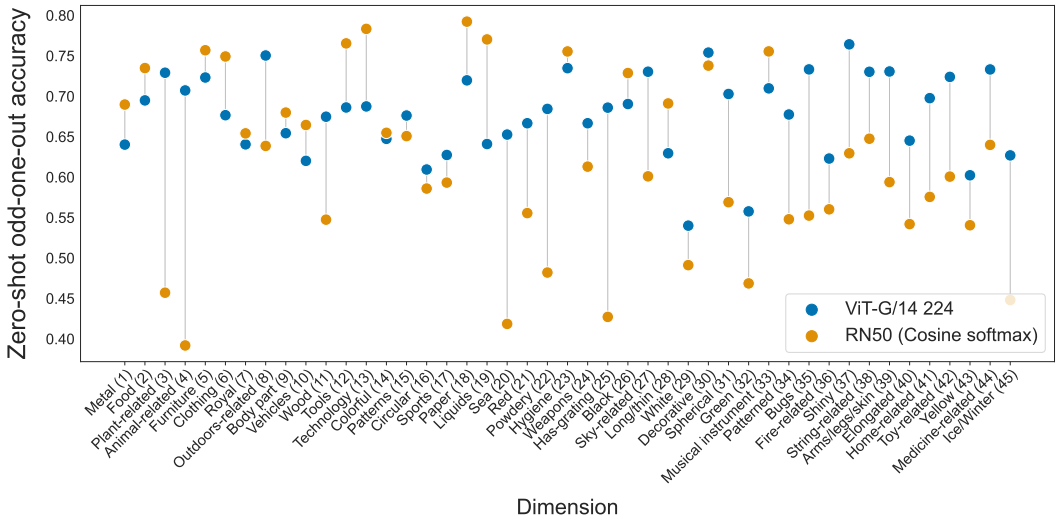


Figure J.1: Comparison of zero-shot odd-one-out accuracy between the best ImageNet and the best overall model for all 45 VICE dimensions.

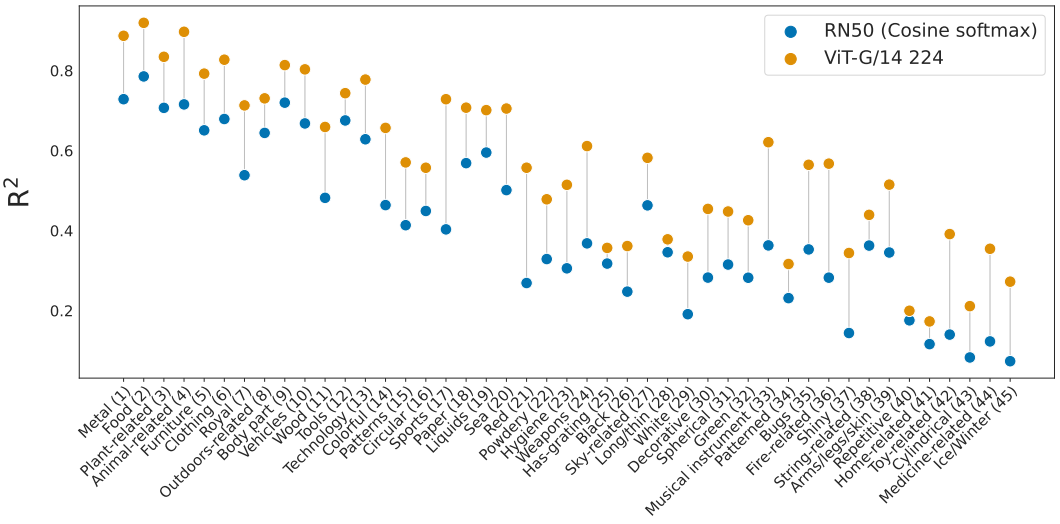


Figure J.2: Comparison of the regression R^2-scores between the best ImageNet and the best overall model for all 45 VICE dimensions.

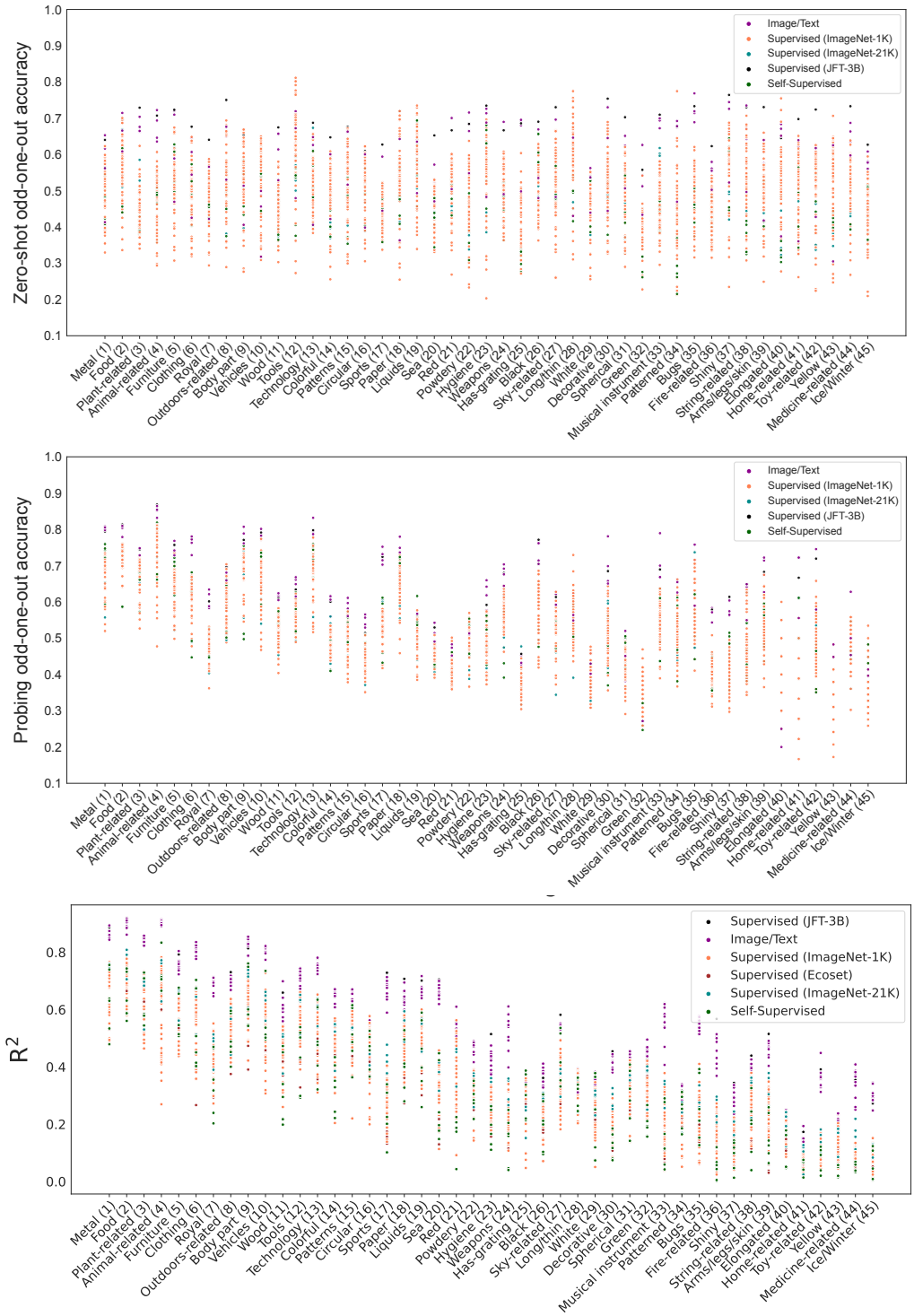


Figure J.3: **Top**: Zero-shot odd-one-out accuracies of all models in Table B.1 for all 45 VICE dimensions. **Middle**: Probing odd-one-out accuracies of all models in Table B.1 for all 45 VICE dimensions. **Bottom**: R^2 scores for all models in Table B.1 after fitting an ℓ_2-regularized linear regression to predict individual VICE dimensions/human concepts.

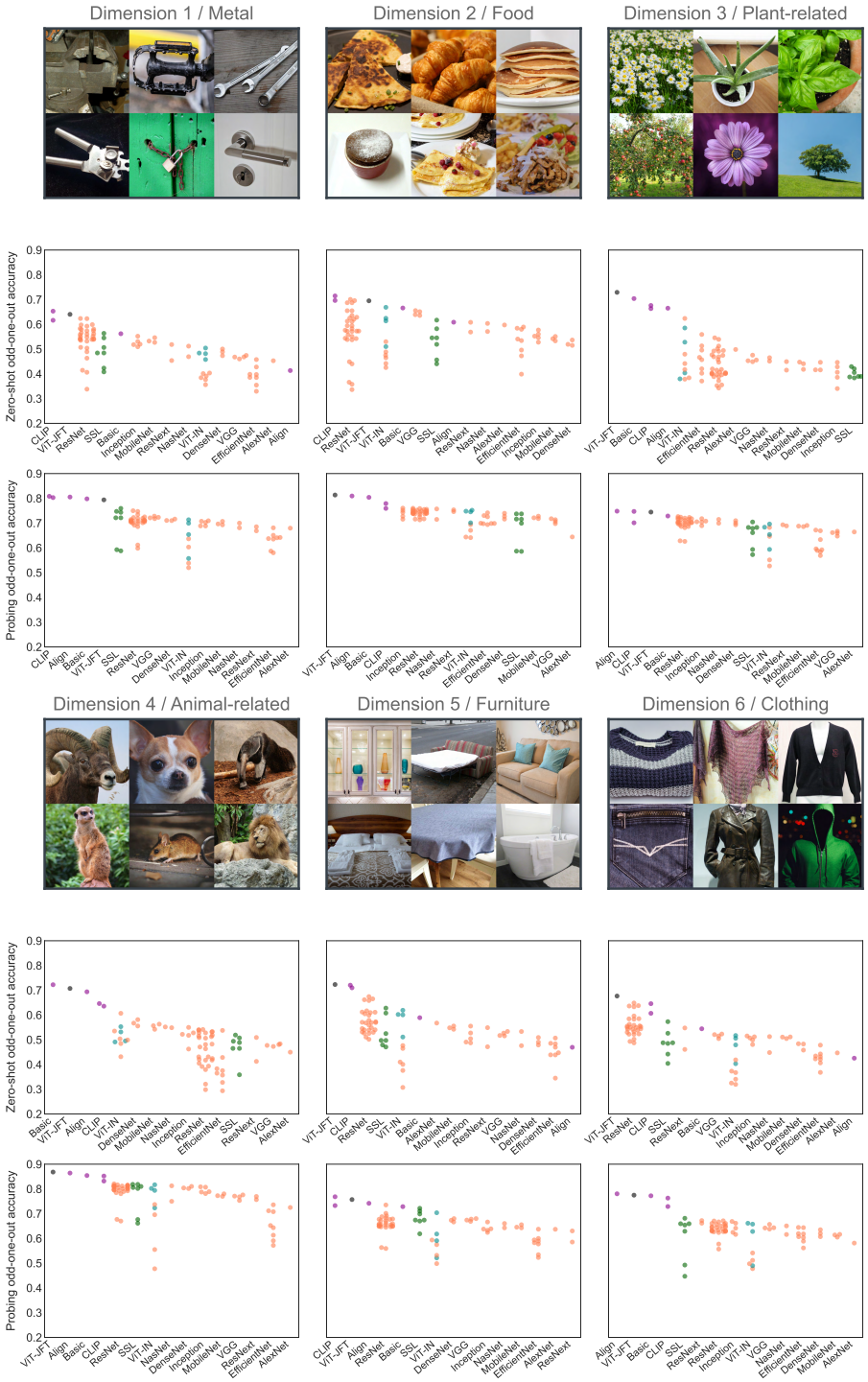


Figure J.4: Zero-shot and probing accuracy for triplets discriminated by VICE dimensions 1-6, following the approach described in §4.4. Color-coding is determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).

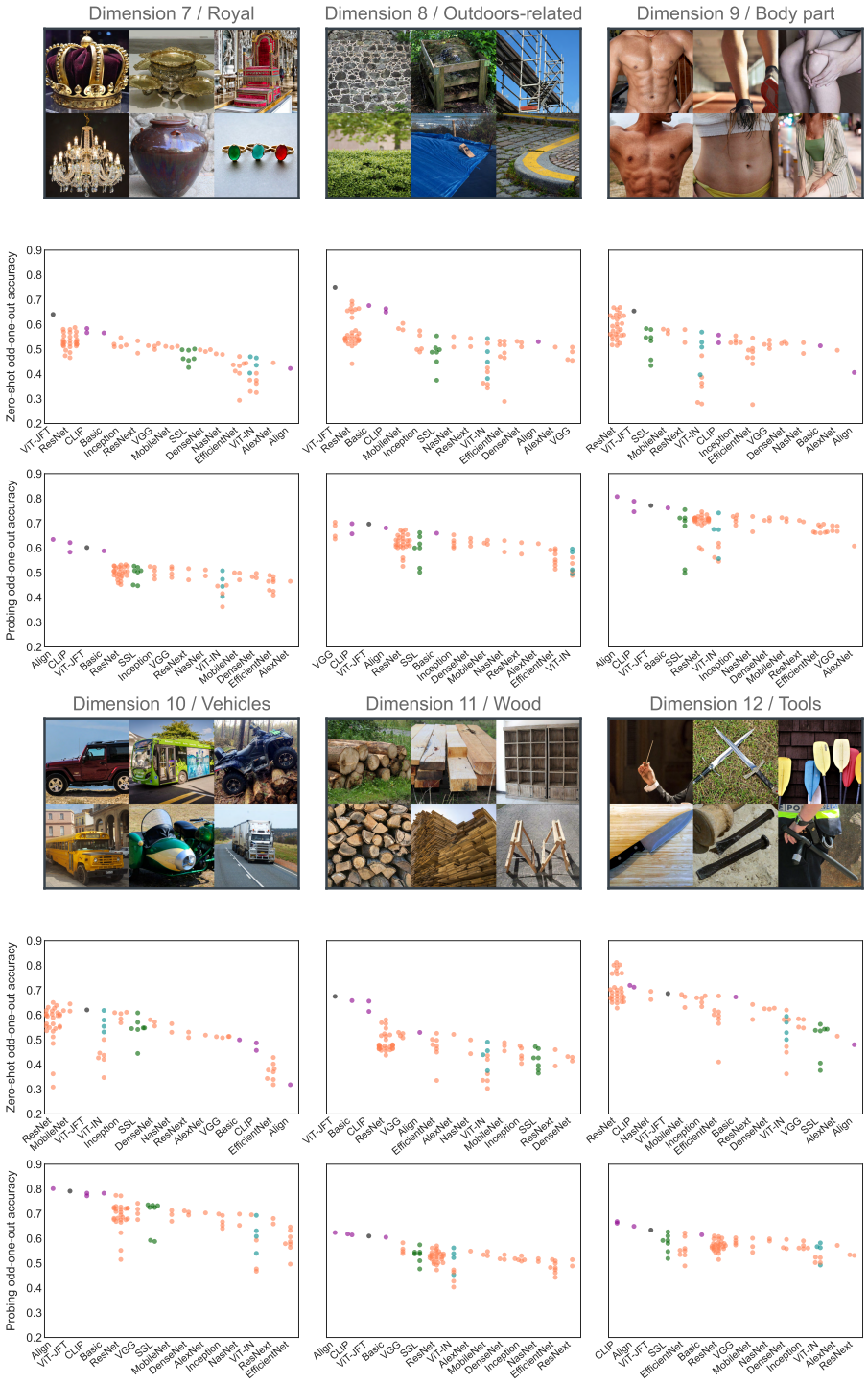


Figure J.5: Zero-shot and probing accuracy for triplets discriminated by VICE dimensions 7-12, following the approach described in §4.4. Color-coding is determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).

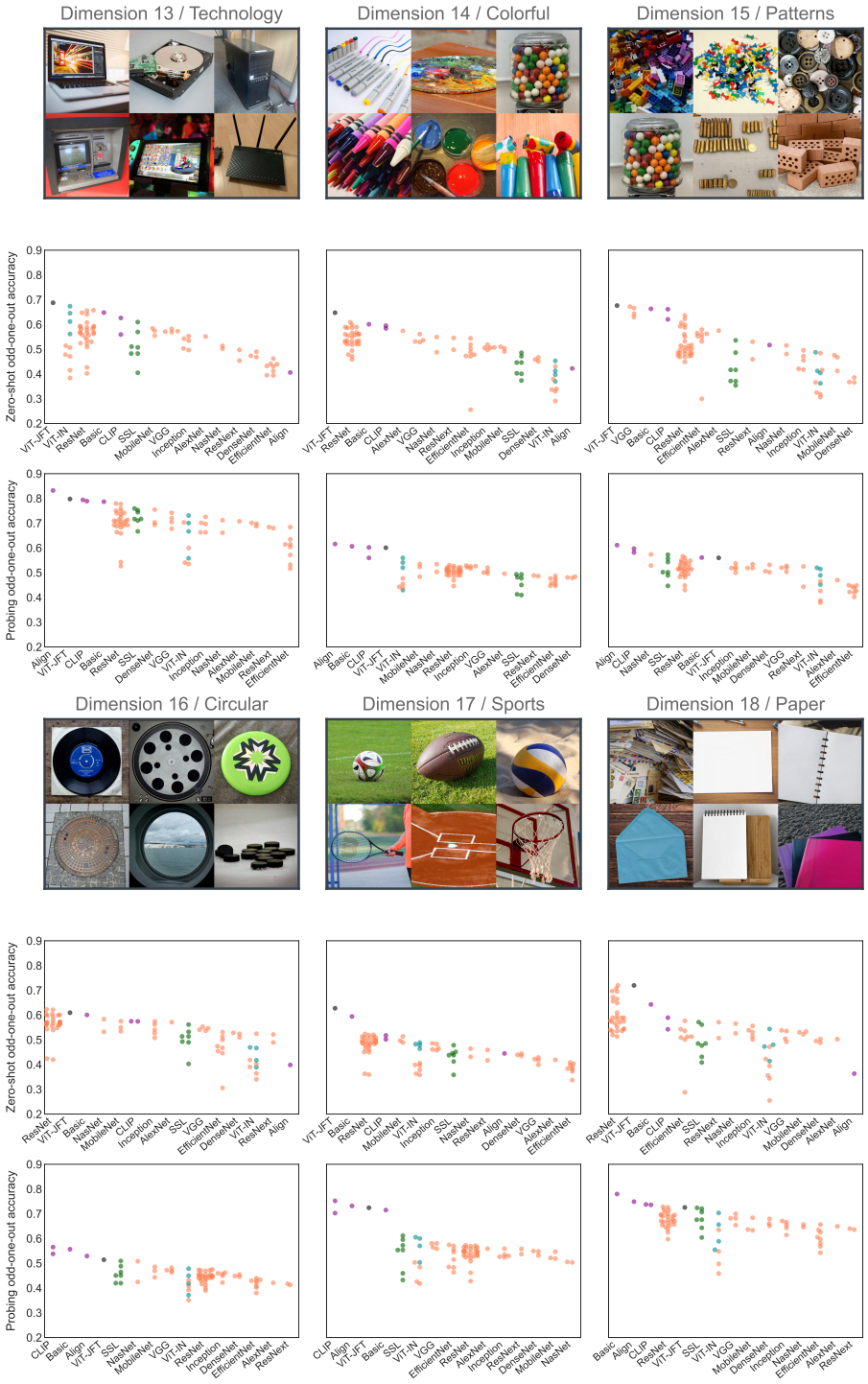


Figure J.6: Zero-shot and probing accuracy for triplets discriminated by VICE dimensions 13-18, following the approach described in §4.4.Color-coding is determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).

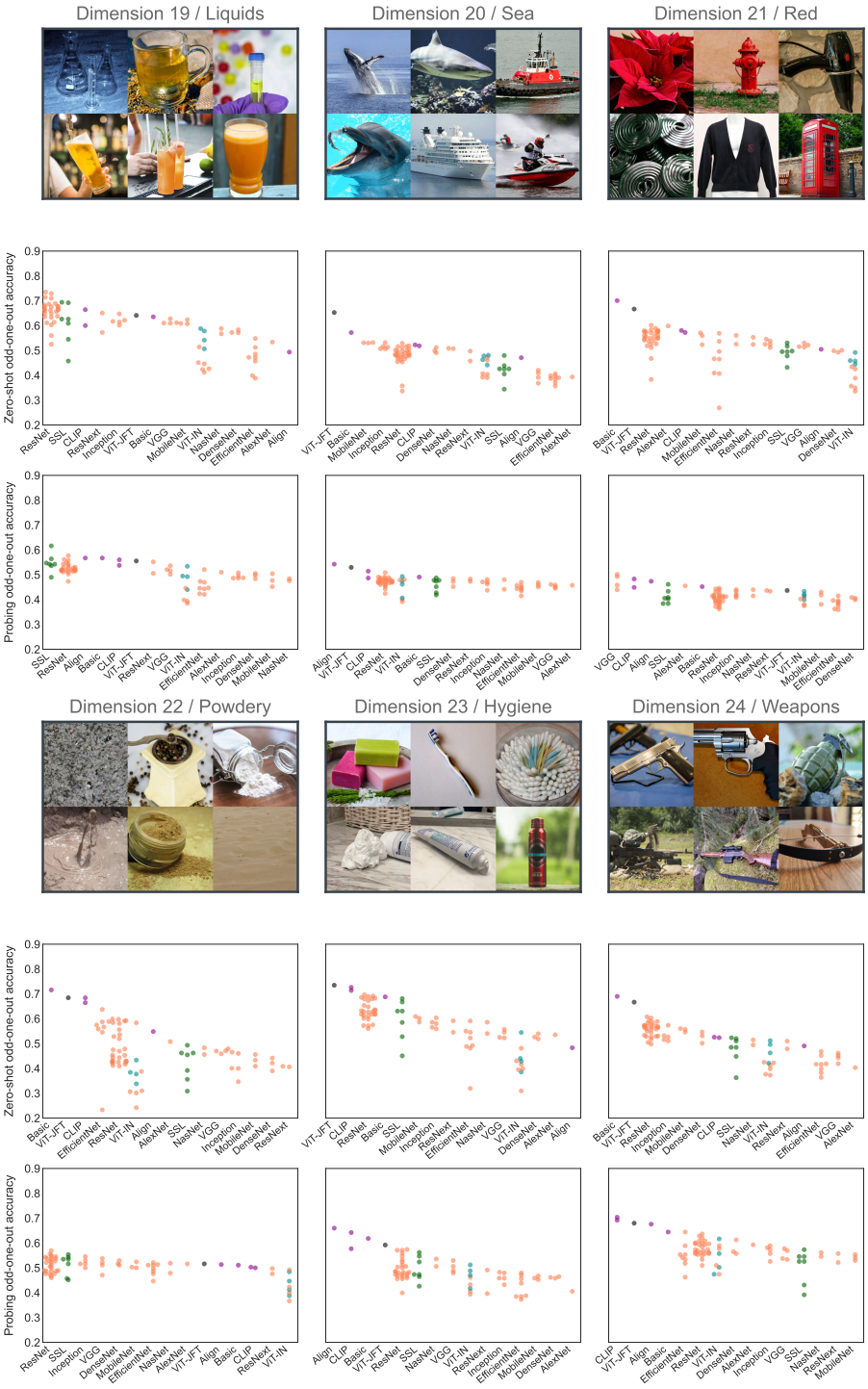


Figure J.7: Zero-shot and probing accuracy for triplets discriminated by VICE dimensions 19-24, following the approach described in §4.4.Color-coding is determined by training data/objective. Violet: Image/Text. Green: Self-supervised. Orange: Supervised (ImageNet-1K). Cyan: Supervised (ImageNet-21K). **Black**: Supervised (JFT-3B).