Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks

Seoyoung Ahn
Department of Psychology, Stony Brook University,
Stony Brook, NY, USA

Gregory J. Zelinsky
Department of Psychology, Stony Brook University,
Stony Brook, NY, USA
Department of Computer Science, Stony Brook University,
Stony Brook, NY, USA

Gary Lupyan
Department of Psychology, University of Wisconsin-Madison,
Madison, WI, USA

Human visual recognition is outstandingly robust. People can recognize thousands of object classes in the blink of an eye (50–200 ms) even when the objects vary in position, scale, viewpoint, and illumination. What aspects of human category learning facilitate the extraction of invariant visual features for object recognition? Here, we explore the possibility that a contributing factor to learning such robust visual representations may be a taxonomic hierarchy communicated in part by common labels to which people are exposed as part of natural language. We did this by manipulating the taxonomic level of labels (e.g., superordinate-level [mammal, fruit, vehicle] and basic-level [dog, banana, van]), and the order in which these training labels were used during learning by a Convolutional Neural Network. We found that training the model with hierarchical labels yields visual representations that are more robust to image transformations (e.g., position/scale, illumination, noise, and blur), especially when images were first trained with superordinate labels and then fine-tuned with basic labels. We also found that Superordinate-label followed by Basic-label training best predicts functional magnetic resonance imaging responses in visual cortex and behavioral similarity judgments recorded while viewing naturalistic images. The benefits of training with superordinate labels in the earlier stages of category learning is discussed in the context of representational efficiency and generalization.

Introduction

Despite the remarkable achievements of recent computer vision models in image classification, the robustness of a model's performance given various transformations of object appearances lags far behind that of humans. The human visual system is unparalleled in its ability to extract invariant visual features of objects that enable good generalization across changes in object position and size (Ito, Tamura, Fujita, & Tanaka, 1995; Rust & DiCarlo, 2010), viewing direction (Biederman & Gerhardstein, 1993; Vuilleumier, Henson, Driver, & Dolan, 2002), illuminations (Vogels & Biederman, 2002), and contrast (Avidan, Harel, Hendler, Ben-Bashat, Zohary, & Malach, 2002; Rolls & Baylis, 1986), and humans are likely the only species who achieve such robustness for so many categories and at various levels of abstraction. In contrast, even state-of-the-art computer vision models such as those implemented by convolutional neural networks (CNNs) are challenged by this variability in visual objects, with recognition performance dropping dramatically by 40% to 50% when objects are presented in varied perspectives and backgrounds (Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, & Katz, 2019). CNNs are also known to be highly vulnerable to other image perturbations, such as adding noise or blur (Hendrycks & Dietterich, 2019), even for changes that are almost imperceptible to humans (Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, & Fergus, 2013) or that affect human perception only marginally (Dodge & Karam, 2017; Geirhos, Temme,

Citation: Ahn, S., Zelinsky, G. J., & Lupyan, G. (2021). Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, *21*(13):13, 1–19, https://doi.org/10.1167/jov.21.13.13.

https://doi.org/10.1167/jov.21.13.13 Received February 28, 2021; published December 30, 2021 ISSN 1534-7362 Copyright 2021 The Authors

Rauber, Schütt, Bethge, & Wichmann, 2018). Because tolerance to variability is what enables accurate and flexible visual object recognition, understanding how humans can successfully build such robust visual representations has been a core question in various disciplines ranging from cognitive science (Biederman, 1987; Tarr & Pinker, 1990) and neuroscience (Plaut & Farah, 1990; Rolls, 1994, p. 199) to computer vision (Marr, 1982; Ullman, 1989).

To understand what makes such robust object recognition possible, researchers have investigated architectural and functional characteristics of human and primate ventral visual stream where object recognition take place (DiCarlo, Zoccolan, & Rust, 2012) and identified biologically plausible algorithms that are helpful for robust recognition, such as pooling operations (Fukushima, 1980; Riesenhuber & Poggio, 2000) and recurrent connections (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Kubilius, Schrimpf, Nayebi, Bear, Yamins, & DiCarlo, 2018). In contrast, the computer vision community has taken a very different approach to improving model robustness, where one prominent approach is to expose CNNs to adversarial examples (e.g., images intentionally modified/perturbed to fool a recognition system) during training (see Akhtar & Mian, 2018, for a review). Here we focus on yet another important aspect of visual recognition that may influence the system's robustness but has been rarely considered: the structure of the category labels used during training. Image classification models that are most frequently used in Computer Vision (e.g., ResNet; He, Zhang, Ren, & Sun, 2016) are trained on 1000 categories from ImageNet (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009). These categories are mostly composed of subordinate-level labels, including, for example, 120 different dog breeds. This may be the appropriate input for creating an expert dog-breed classifier, but it results in a semantic structure that is highly atypical compared to that of an average person.

In contrast, category labels used by people have a roughly hierarchical organization structure, all chihuahuas are dogs, all dogs are mammals, and all mammals are animals. Rosch and her colleagues in 1976 found that adults are generally faster and more accurate to identify objects at the basic level of categorization (e.g., "dog") than at the superordinate level (e.g., "mammal") or at the subordinate level (e.g., "chihuahua"). Developmental studies have similarly observed that children find basic-level categories easier to learn and process, acquiring basic labels earlier than others (Mervis & Crisafi, 1982; Murphy & Lassaline, 1997; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), although some studies suggest that more general concepts like superordinate-level categories are often acquired first by infants and young children (Mandler, Bauer, & McDonough, 1991; Quinn & Johnson, 2000; see Murphy (2016) for how these conflicting results might be resolved). Although the basic-level advantage is one of the best known and most replicated phenomena in the field of human categorization (Murphy & Lassaline, 1997; Tanaka & Taylor, 1991; Tversky & Hemenway, 1984), the question we ask here is different: What effects do different kinds of labels have on learning robust visual representations?

Previous behavioral studies found that having semantic associations between categories facilitated extraction of invariant visual features for object recognition (Collins & Curby, 2013; Curby, Hayward, W. G., & Gauthier, 2004; Gauthier, James, Curby, & Tarr, 2003). In these studies, participants learned semantic associations between novel objects through training with adjective labels (e.g., sticky, loud, nocturnal). They then performed a perceptual-matching task where they compared two sequentially presented objects and judged whether they were from the same or different category. In the perceptual-matching task, the first object of the pair was always presented at a canonical orientation (0°), whereas the second object could be presented at one of four depth-orientations (0°, 30°, 60°, or 120°). Curby et al. (2004) found that learning semantic relationships between categories reduced viewpoint dependency in human object recognition: both the accuracy and response time needed to discriminate two objects being less impacted by changes in an object's orientation. Collins & Curby (2013) later extended this research by demonstrating that labels which were devoid of semantic association (e.g., numbers) did not create the same tolerance in visual recognition to depth-rotated objects; learning semantically meaningful associations between categories was key. Here, we build on this work by asking whether a model exposed to hierarchical labels during training learns more robust visual representations compared to models supervised with labels from only a single hierarchical level, and models not supervised by labels at all (methods and results for unsupervised models are in the Supplementary Materials SM7).

Leveraging semantic information from learned categories has been previously shown to improve the classification performance of CNNs (Annadani & Biswas, 2018; Frome, Corrado, Shlens, Bengio, Dean, Ranzato, & Mikolov, 2013; Lei Ba, Swersky, & Fidler, 2015, p. 20; Peterson, Soulos, Nematzadeh, & Griffiths, 2018). For example, Frome et al. (2013) re-trained a CNN to predict word vectors learned by a word embedding model (Mikolov, Chen, Corrado, & Dean, 2013) and found that learning the semantic similarity between categories significantly boosted the model's zero-shot learning performance (i.e., ability to predict novel categories that are never seen during training). More related to our work, Peterson et al. (2018) explored how labels at different levels in a category hierarchy, and the order that they were used in training, affected the visual representations learned by CNN models. In this work, CNNs were trained using either just one level of taxonomy labels (e.g., basic or subordinate) or using multi-level labels (e.g., training first with basic labels and then fine tuning with subordinate labels, or vice versa). They found that training on basic-level labels, either alone or following subordinate-level training, induced a more semantically structured representation, which better predicted human similarity judgments and category generalization patterns (e.g., generalizing to a new basic-level category after observing only a few subordinate exemplars). These results suggest that learning categories with a hierarchical structure of labels is beneficial for forming more human-like visual representations in CNNs; however a direct link between hierarchical label training and increased robustness in visual recognition has never been tested.

Here we investigated how the hierarchical structure of labels provided during category learning changes visual representations of objects that are learned by the networks. We aim to understand which regiment of label training achieves the most robust and human-like visual representations. To investigate this problem, we took a synergistic computational and behavioral approach. We simulated object-category learning using CNNs and thousands of naturalistic images (Deng et al., 2009), and manipulated the training of these models using labels from two different levels of a category hierarchy: superordinate and basic. We chose these two levels because they constitute the most basic and inherent structure of human semantic knowledge, even creating a debate over their advantages in category learning and acquisition (see Murphy, 2016, for a review). To compare the visual representations learned from these differently trained models, we conducted three computational experiments (Experiments 1–3) each focused on a specific type of representational robustness: (1) tolerance to various visual transformations in object recognition, (2) categorical separability (making representations of different categories more dissimilar or separable), and (3) shape bias (preserving shape information more than other visual features, e.g., texture, in object representation). We also compared models to see which produced the more human-like categorical representations. We did this by comparing their predictive performance to functional magnetic resonance imaging (fMRI) signals recorded by Chang, Pyles, Marcus, Gupta, Tarr, and Aminoff (2019) during the viewing of naturalistic images (Experiment 4) and by comparing behavioral responses that we collected using a triplet similarity judgement task (Experiment 5). Given the previous work, described above, relating semantic knowledge to visual representation, we hypothesize that the exposure to hierarchical relationships between labels during training will produce more robust and human-like visual representations compared to models trained with a single-level of category labels, although it is not clear from the literature which order of training (superordinate-then-basic or basic-then-superordinate) will result in the most robust visual representations.

Modeling method

In this study, we trained CNNs using identical architectures and training sets while manipulating the labels used to supervise the training. Specifically, we trained models with the following: basic-level labels only, superordinate-level labels only, first with basic labels and then fine-tuned with superordinate labels, or first with superordinate labels and then fine-tuned with basic labels. For analysis of the visual representations, we extract the bottleneck features from the encoder of each model (i.e., the 1568-dimensional output of the last convolutional layer; see Figure 1). Below we provide a detailed description of the model training and the methods used for statistical comparison.

Model architecture

Our CNN models consist of five blocks of two convolutional layers, each followed by max pooling and batch normalization layers (Figure 1). For all convolutional and max pooling operations, zero padding was used to produce output feature maps having the same size as the input. Rectified linear units (ReLU) were used to obtain an activation function after each convolution. The flattened output of the final Convolutional layer—the "bottleneck" features (dim = 1568)—were then extracted as a model's visual representation and fed into one fully connected dense layer. Softmax was used to obtain output activation functions for the supervised models.

Model training

We manipulated the types of labels used during the supervised training of these CNNs. The basic only model was trained with basic-level labels only, and the superordinate only model was trained only on superordinate-level labels. The basic-then-superordinate and the superordinate-then-basic models were trained with both types of labels, but in different order. For training and validation, we used 30 basic categories from the ImageNet 2012 dataset (Deng et

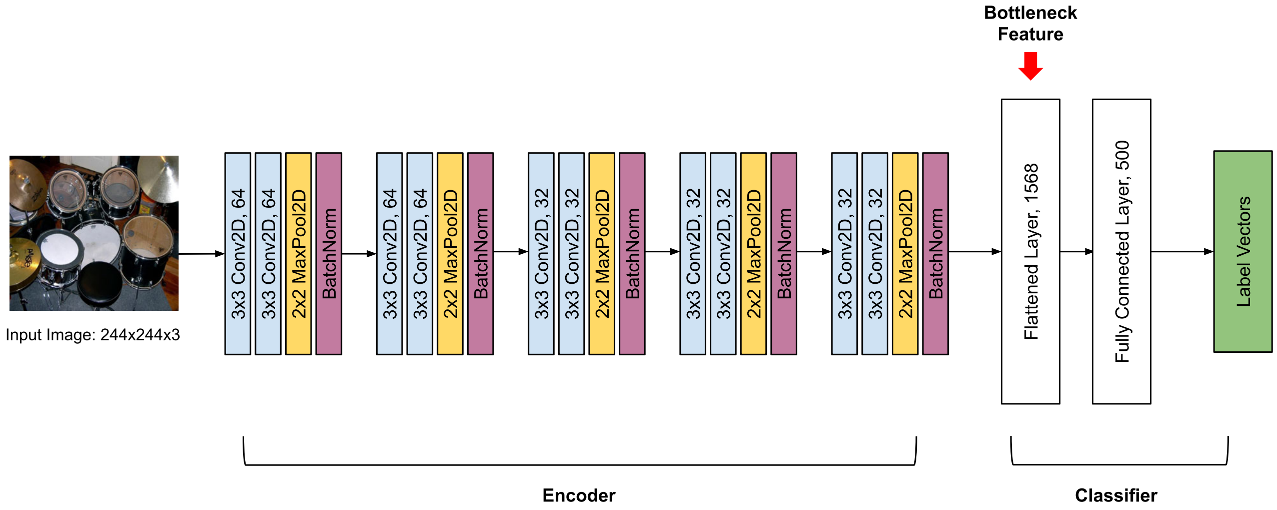

Figure 1. The Pipeline for the CNNs used in our study. The bottleneck features (the flattened output of the final Convolutional layer) are extracted and used as a model's visual representation (dim = 1568). The final predicted output, "Label Vector," is a one-hot embedding of labels according to the model's labeling scheme (e.g., basic-only, superordinate-only, basic-then-superordinate, superordinate-then-basic).

		Classification accuracy on the testing dataset					
		Using original softmax layer		Using linear classifier		Using prototypical representation	
Model	Number of categories learned	Superordinate	Basic	Superordinate	Basic	Superordinate	Basic
Basic only	30	N/A	0.9	0.90	0.87	0.85	0.80
Superordinate only	10	0.95	N/A	0.93	0.80	0.89	0.66
Basic-then-superordinate	40	0.95	N/A	0.94	0.85	0.91	0.78
Superordinate-then-basic	40	N/A	0.88	0.93	0.86	0.90	0.81

Table 1. Classification accuracy on the testing dataset. To evaluate the learned representation on both category levels, we report two accuracy measures in addition to the classification scores from the originally trained final Softmax layer: (1) using a linear classifier trained on top of the frozen base encoder (which outputs the "bottleneck" features), and (2) using the representation of the category prototype (used for analyzing robustness to image transformation, see Experiment 1 Method for details). Average precision and average recall scores are reported in Supplementary Material, SM2.

al., 2009), which can be grouped into 10 higher-level, superordinate categories: "mammal," "bird," "insect," "fruit," "vegetable," "vehicle," "container," "kitchen appliance," "musical instrument," and "tool'. This dataset includes both inanimate and animate categories and both natural and human-made categories, thereby making the taxonomic structure more representative of human conceptual knowledge. See Supplementary Material (SM1) for a full list of categories used, with their corresponding labels. For testing, we used the same 30 categories from the THINGS dataset (Hebart, Dickter, Kidder, Kwok, Corriveau, Van Wicklin, & Baker, 2019). We used different training and testing datasets so as to exploit the behavioral similarity ratings that we collected on the smaller THINGS dataset. All images were zero-centered with respect to the ImageNet images' distribution (i.e., RGB color values for each image were subtracted from the mean of the ImageNet training dataset), and categorical cross entropy loss was used for all models trained with labels. All parameter updates are done using Adam optimization (Kingma & Ba, 2014), with a mini-batch size of 64. Model training terminated based on the results from the validation dataset, and the model with the lowest validation loss was used for all subsequent analyses performed on a given model's visual representations.

As shown in Table 1, the models trained with both categories (basic-then-superordinate and superordinate-then-basic) performed similarly at both basic and superordinate-level classification. The models trained with a single type of label (basic or superordinate) show relatively low performance at the level the model was not trained on (e.g., basic only had 90% of the superordinate-level prediction accuracy when using a linear classifier, which is lower than the accuracy of superordinate only, 93%).

Statistical analyses

Statistical comparison between models was made using general linear mixed models (GLMMs) with binomial error distributions from the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). Each dependent variable, including stimuli items (e.g., images), participants, or both, were modeled as random effects (Baayen, Davidson, & Bates, 2008) and other independent variables of interest as fixed effects. The detailed modeling and coding schemes for the fixed effects are provided in each Results section. The significance of the interactions between fixed effects was checked by comparing the likelihoods of the models with and without interaction terms, using the likelihood ratio test. Random effects were set up to affect only the intercepts, not the slopes, to allow a nonsingular fit (Barr et al. 2013).

Experiment 1

In this experiment, we evaluated each model's robustness by measuring its recognition performance of images subjected to various transformations. We explored four types of image transformations: (1) affine transformation (vertical and horizontal shift, rotation, scale, and shear), (2) brightness, (3) salt-and-pepper noise, and (4) Gaussian blur. Examples of each type and level of transformation are shown in the Supplementary Material (SM3).

Method

Because the original classification output from the last Softmax layer is restricted to the categories used for training (e.g., the basic only model can only predict basic categories), we used the learned category representations to estimate recognition accuracy on both levels of categorical hierarchy by first extracting prototypical representations for every class from each model and then comparing these to the visual representation extracted from a transformed image. In the current analysis, a prototypical representation is computed by taking a mathematical average of all bottleneck features extracted from the images in the training dataset that belong to the category (Posner, 1970), a method widely used in vision models (Snell, Swersky, & Zemel, 2017). These prototypical representations were also used to predict human similarity judgments in Experiment 5. We considered recognition to be accurate when the highest cosine similarity is achieved between a visual representation extracted from a transformed image and the corresponding prototypical categorical representation generated from the ground truth class. For example, if a model's visual representation for a test image has the highest cosine similarity with the prototypical representation of its learned "banana" category, and if the ground truth label is also "banana," then the model would be scored as recognizing the object correctly. Differing levels of image transformations were applied to our test dataset, with higher levels denoting more extreme transformations.

Results and discussion

Visualizations of the models' average recognition accuracy are shown in Figure 2. Among the trained models, superordinate-then-basic model (the purple line) was consistently more accurate in recognizing transformed images than any other models, where the effect was most prominent in noise and blur conditions. We observed that the superiority of superordinate-then-basic model appears in both natural (e.g., "mammal," "bird," "insect," "fruit," "vegetable") and human-made categories (e.g., "vehicle," "container," "kitchen appliance," "musical instrument," and "tool"), whereas the effects were more pronounced in natural categories that are generally known to have higher visual consistency than artifact categories (Supplementary Material, SM5). For statistical analysis, the recognition accuracy on each testing image was modeled as a function of training type and transformation level using GLMMs with binomial family (see Statistical Analyses in Method section for details). For the fixed effects, the training types were coded using treatment contrasts with basic only as the baseline, and transformation level (0–6) was treated as a continuous variable.

A significant interaction between training type and transformation level was observed in all image transformations and taxonomic levels, where adding the interaction term significantly improved the statistical models (in superordinate-level recognition, $\chi^2(3) = 23.12$, p < 0.001 for affine transformation, $\chi^2(3) = 62.95$, $p < 0.001$ for brightness, $\chi^2(3) = 132.12$, $p < 0.001$ for noise; $\chi^2(3) = 47.52$, $p < 0.001$ for blur; in basic-level recognition, $\chi^2(3) = 11.83$, $p < 0.01$, for affine transformation, $\chi^2(3) = 9.43$, $p < 0.05$ for brightness, $\chi^2(3) = 43.51$, $p < 0.001$ for noise, $\chi^2(3) = 21.96$, $p < 0.001$ for blur). This indicates that there are significant differences in the robustness of our trained models, as measured by recognition performance over

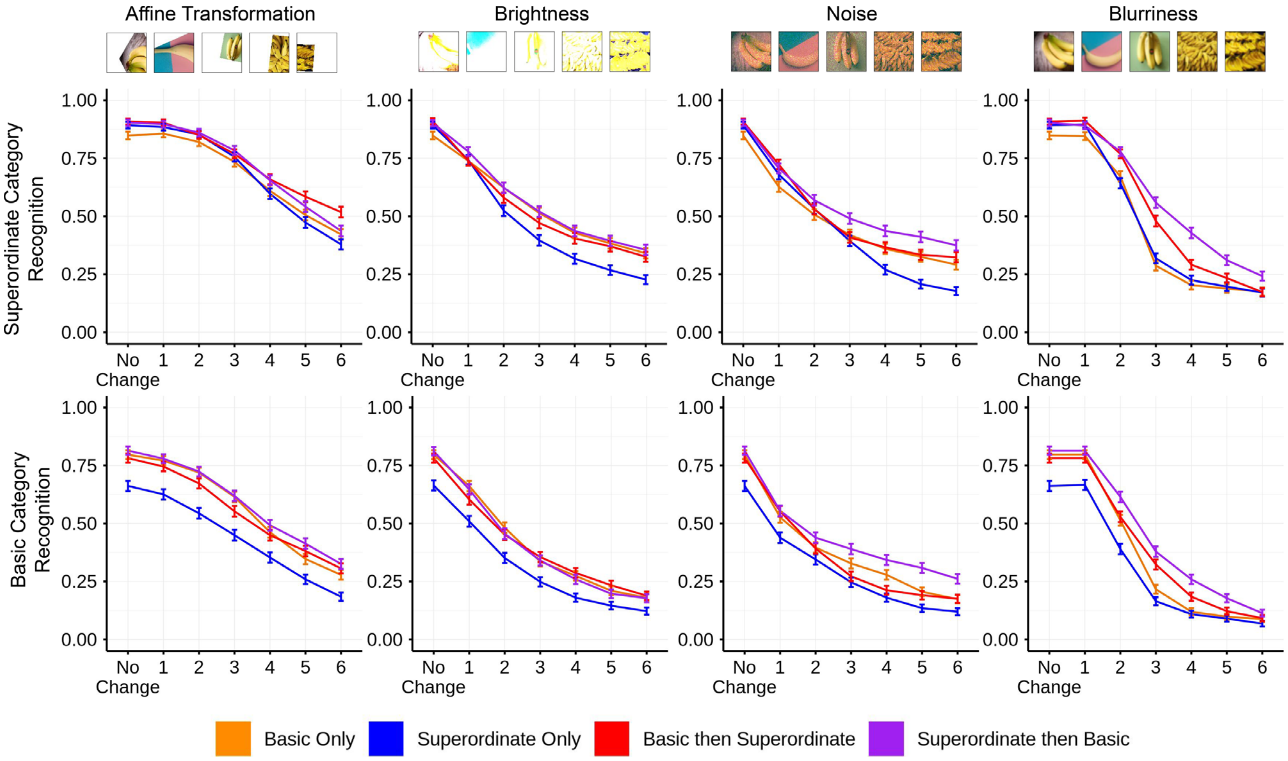

Figure 2. Measuring the robustness of superordinate-level and basic-level recognition (rows) to four image transformations (columns), with higher values along the x-axes indicating greater transformation applied to the image. Category recognition accuracy (y-axes) based on cosine similarity between each model's visual representation of a test image and its learned category prototype. Appearing at the top of each column are examples of images with the highest-level transformation applied. Complete examples are provided in the Supplementary Material. *Error bars* are SEM calculated over 467 testing images.

differing levels of distorted images. As shown in Table 2, recognition performance for all trained models declined sharply with increasing level of transformation, indicated by negative coefficients of transformation level, exposing the brittleness of CNN representations. However, the degree of performance degradation caused by image transformation depended on how the models were trained. Superordinate-then-basic was observed to be comparatively more robust for most image transformation types, exhibiting both higher intercept (recognition accuracy of images not subjected to any transformation) and smaller decline in accuracy as a function of transformation level (shown as positive interaction terms), compared to other models.

Post-hoc pairwise comparisons with Bonferroni corrections confirmed that when effects were collapsed over differing levels of transformation, the recognition performance of superordinate-then-basic model was significantly superior to all other models for most types of image transformations, especially for noise and blur transformations (all comparisons were significant with adjusted $p < 0.05$), and this was true for both recognition at both the basic and superordinate levels. For affine image transformation and brightness change, the performance from superordinate-then-basic was often on par with basic only or basic-then-superordinate models (adjusted $p > 0.05$). See Supplementary Material, SM4 for full comparison results.

Experiment 2

We often perceive exemplars from one category of object as more similar to each other and less similar to exemplars from other categories (Harnad, 1987). Given the variations that exist in object appearance, such categorical perception helps to efficiently process the visual features that are relevant to recognizing the visual object. In this experiment, we examined the extent to which visual representations learned by each model are categorical (their "categorical separability"), by measuring a visual similarity distance for objects belonging to different categories compared to the same category (Goldstone & Hendrickson, 2010). We also compare t-distributed stochastic neighbor embedding (t-SNE) visualizations of each model's visual representations.

Method

We estimated each model's categorical separability using a variation of the Davies–Bouldin index (Davies

	Affine transformation			Brightness			Noise			Blur		
	b	SE	Z	b	SE	Z	b	SE	Z	b	SE	Z
Superordinate-level recognition												
Intercept (basic only)	3.32	0.15	**22.35**	2.91	0.17	**17.26**	2.19	0.16	**14.11**	3.84	0.19	**20.26**
Superordinate only	0.64	0.16	**4.02**	0.19	0.14	1.38	0.73	0.13	**5.46**	0.43	0.16	**2.70**
Basic-then-superordinate	0.54	0.16	**3.42**	0.16	0.14	1.17	0.51	0.13	**3.90**	1.18	0.17	**6.91**
Superordinate-then-basic	0.75	0.16	**4.64**	0.36	0.14	**2.57**	0.47	0.13	**3.60**	0.80	0.17	**4.82**
Transformation level	−0.63	0.03	**−22.64**	−0.74	0.03	**−25.61**	−0.73	0.03	**−25.60**	−1.40	0.04	**−34.09**
Superordinate only × transformation level	−0.17	0.04	**−4.20**	−0.31	0.04	**−7.53**	−0.37	0.04	**−9.21**	−0.08	0.05	−1.67
Basic-then-superordinate × Transformation Level	−0.01	0.04	−0.34	−0.09	0.04	**−2.24**	−0.08	0.04	**−2.19**	−0.09	0.05	−1.69
Superordinate-then-basic × Transformation Level	−0.11	0.04	**−2.76**	−0.06	0.04	−1.55	0.06	0.04	1.51	0.22	0.05	**4.40**
Basic-level Recognition												
Intercept (basic only)	2.52	0.14	**18.27**	1.88	0.14	**13.55**	1.51	0.14	**10.45**	2.44	0.14	**17.31**
Superordinate only	−1.31	0.13	**−10.07**	−0.98	0.12	**−7.86**	−0.71	0.13	**−5.64**	−1.12	0.14	**−8.28**
Basic-then-superordinate	−0.41	0.13	**−3.06**	−0.29	0.12	**−2.33**	−0.02	0.13	−0.19	−0.10	0.14	−0.69
Superordinate-then-basic	0.02	0.14	0.11	−0.01	0.13	−0.08	0.03	0.12	0.24	0.24	0.14	1.70
Transformation level	−0.69	0.03	**−25.68**	−0.83	0.03	**−28.48**	−0.81	0.03	**−27.24**	−1.23	0.04	**−32.77**
Superordinate only × Transformation level	0.11	0.04	**3.22**	0.03	0.04	0.78	−0.03	0.04	−0.74	0.19	0.05	**4.07**
Basic-then-superordinate × Transformation level	0.09	0.04	**2.60**	0.09	0.04	**2.39**	−0.06	0.04	−1.42	0.14	0.05	**2.99**
Superordinate-then-basic × Transformation Level	0.05	0.04	1.49	−0.02	0.04	−0.50	0.17	0.04	**4.58**	0.19	0.05	**4.06**

Table 2. Results from GLMMs predicting recognition accuracy. Training type and transformation level, and their interactions were modeled as fixed effects, and testing images (n = 467) were treated as random effects (see Method for details). Significant z-scores are in bold (adjusted $p < 0.05$).

& Bouldin, 1979). This index takes into account the ratio of two components: (1) the distance between visual representations of objects from different categories (between-class dispersion), and (2) how closely the visual representations of objects within the same category are located in the representational space (within-class dispersion). To measure the visual similarity between exemplars, the cosine similarity was calculated on the "bottleneck" features extracted from each exemplar image in the testing dataset.

To visually compare how tightly clustered are the visual representations within each category, we projected high-dimensional visual feature vectors into hundred-dimensional space using principal component analysis (PCA), and further projected them into two-dimensional space using t-SNE with the perplexity of 50 with random initialization, with maximum iteration of 1000 (Maaten & Hinton, 2008).

Results and discussion

According to one-way analysis of variance, category separability did not significantly differ between different training schemes in either the basic and superordinate-levels , $F(3,116) = 1.07$, $p > 0.05$, and $F(3,36) = 0.79$, $p > 0.05$, respectively. This finding is particularly interesting given the coarse training of the superordinate only model. Indeed, superordinate only showed categorical separability at the basic level as well, despite never being trained on basic-level categories, and its score was even comparable to that of the basic only model. The model trained with basic labels before training with superordinate labels had numerically higher categorical separability for visual representations at both the basic and superordinate hierarchical levels (see Table 3), but the differences were not significant (adjusted $p > 0.05$).

Similar patterns can be observed from an inspection of the t-SNE visualizations of each model's visual representations. In Figure 3, each data point indicates a visual representation of a testing image projected in two-dimensional space. Note that colors in the figure are coding the ground truth labels for the testing dataset and are not available to the t-SNE visualization algorithm. The trained models produced basic-level and superordinate-level category structure in their representations, which are signified as clusters having

	Superordinate-level separability	Basic-level separability
Basic only	1.10 (0.06)	1.73 (0.07)
Superordinate only	1.23 (0.10)	1.64 (0.07)
Basic then superordinate	1.29 (0.12)	1.82 (0.09)
Superordinate then basic	1.20 (0.07)	1.79 (0.07)

Table 3. Category separability (Inverse of Davies-Bouldin Index) of visual representations for each model, with higher values indicating greater categorical separability, for example, greater dissimilarity of visual representation between different categories compared to similarity of visual representations within the same category. Standard errors are calculated over all categories at each taxonomic level and reported in the parentheses.

the same color in the t-SNE plots. There is even evidence for some hierarchical structure as well, for example, the three small groups of light blue dots corresponding to “lion,” “gazelle,” and “orangutan” exemplars within the global “mammal” cluster in the basic-then-superordinate and superordinate-then-basic models. Again remarkably, training only with superordinate categories allows the model to learn considerable structure at the *basic* level.

Experiment 3

Previous research suggests that a CNN’s classification is heavily influenced by differences in texture whereas human object recognition is much more reliant on overall shape information (Baker, Lu, Erlikhman, & Kellma, 2018; Geirhos, Rubisch, Michaelis, Bethge, Wichmann, & Brendel, 2019). This experiment examined how biased our trained models became in using shape versus texture when recognizing images with swapped (transferred) texture information (see Method for details).

Method

We created texture-transferred testing images where the original object’s shape is preserved, but the texture is replaced with that of other categories using AdaIN style transfer (Huang & Belongie, 2017; Geirhos et al., 2019; See Figure 4A for an example image) and calculated each model’s shape bias index. Here we define shape bias as the relative proportion of images that are classified according to its shape with respect to all images either classified by their shape or texture. If shape bias is above 0.50, the model prefers to use shape information rather than texture during recognition. Each model’s recognition was generated using the same method used in testing recognition robustness, by finding the category whose prototypical categorical representation has the highest cosine similarity with a visual representation extracted from a texture-transferred image.

Results and discussion

Figure 4C suggests that our trained models have different levels of shape bias for object recognition. Overall, the models supervised with hierarchical categorical labels produced higher shape bias than models trained with a single level of hierarchy, such as superordinate only or basic only. In particular, when recognition was performed at the basic-level (second row in Figure 4), the superordinate-then-basic model was shown to be the most shape-biased, whose shape bias score reflected an actual shape bias rather than a bias to use texture. This pattern was confirmed by statistical analysis, where GLMMs with binomial family were used to model the recognition accuracy on each testing image as a function of training type, which were coded using treatment contrasts with basic only as the baseline (see Statistical Analyses in Modeling Method section for details).

A significant main effect of training type on shape bias was observed in both levels of recognition ($\chi^2(3) = 111.55$, $p < 0.001$ for superordinate-level recognition; $\chi^2(3) = 129.07$, $p < 0.001$ for basic-level recognition). Although none of trained networks were biased to use shape rather than texture information in superordinate-level recognition models, when recognition was performed at the basic-level, superordinate-then-basic (0.59) and basic-and-superordinate (0.55) achieved a shape bias score higher than 0.50, which indicates a preference for these two models to use shape than texture information when recognizing visual objects, as humans do. Post-hoc comparison with Bonferroni adjustments also confirmed that superordinate-then-basic achieved the highest shape bias compared to any other trained model in basic-level recognition, and all comparisons were statistically significant ($Z_{ratio} = 9.32$, $p < 0.001$ from basic only; $Z_{ratio} = 9.29$, $p < 0.001$ from superordinate only; $Z_{ratio} = 3.42$, $p < 0.001$ from basic-then-superordinate). Shape bias scores for individual basic-level categories were also reported in the Supplementary Material SM12.

Experiment 4

In this experiment, we evaluated the model’s ability to predict patterns of neural activity, as measured by fMRI in humans, elicited by the same images shown to the model during training. This was made possible by a publicly-available human fMRI dataset, BOLD5000

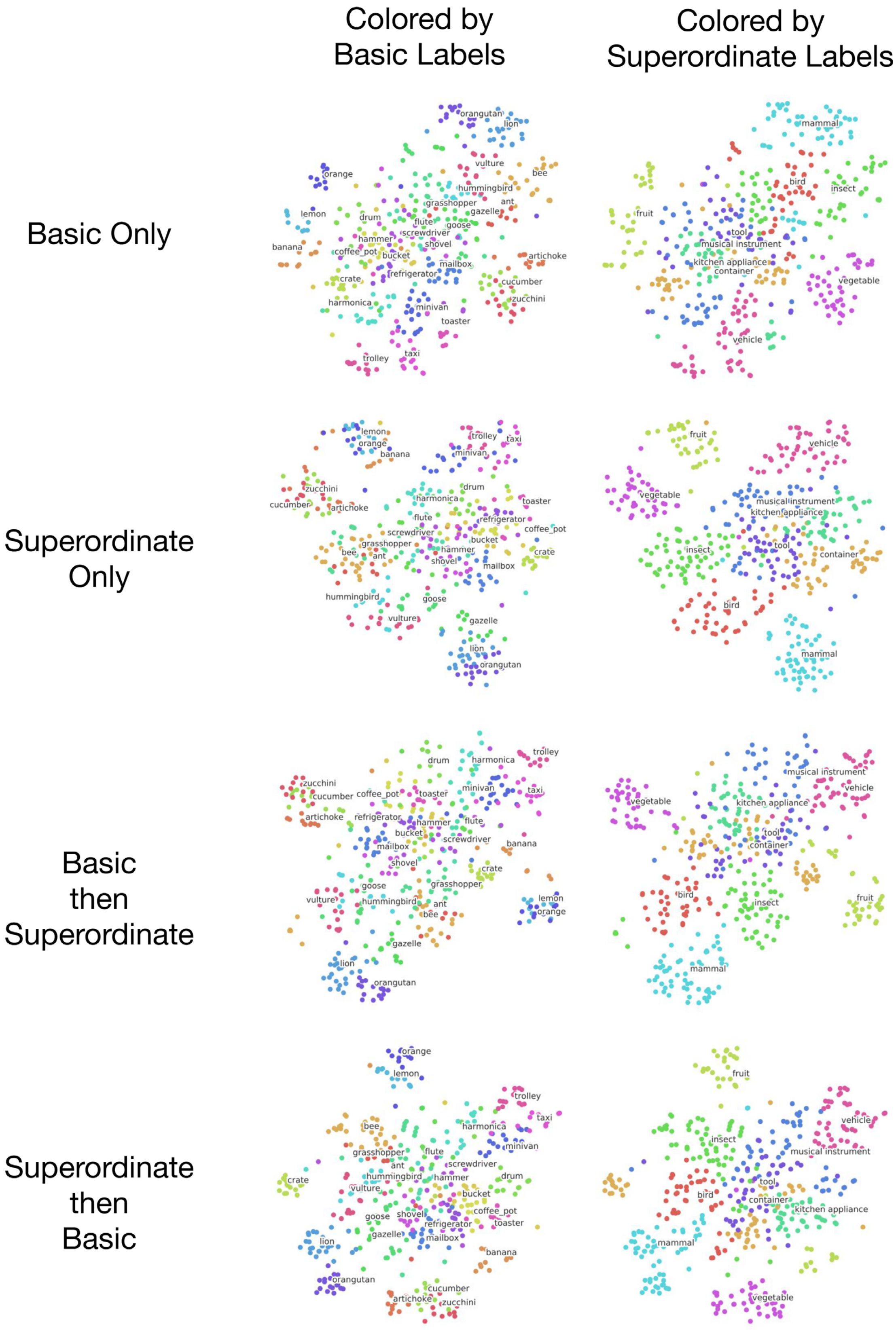

Figure 3. The t-SNE Visualizations of image representations extracted from the trained models. The same feature distribution for each model is color-coded differently by superordinate-level labels (first column) and basic-level labels (second column). Best viewed in PDF form with magnification.

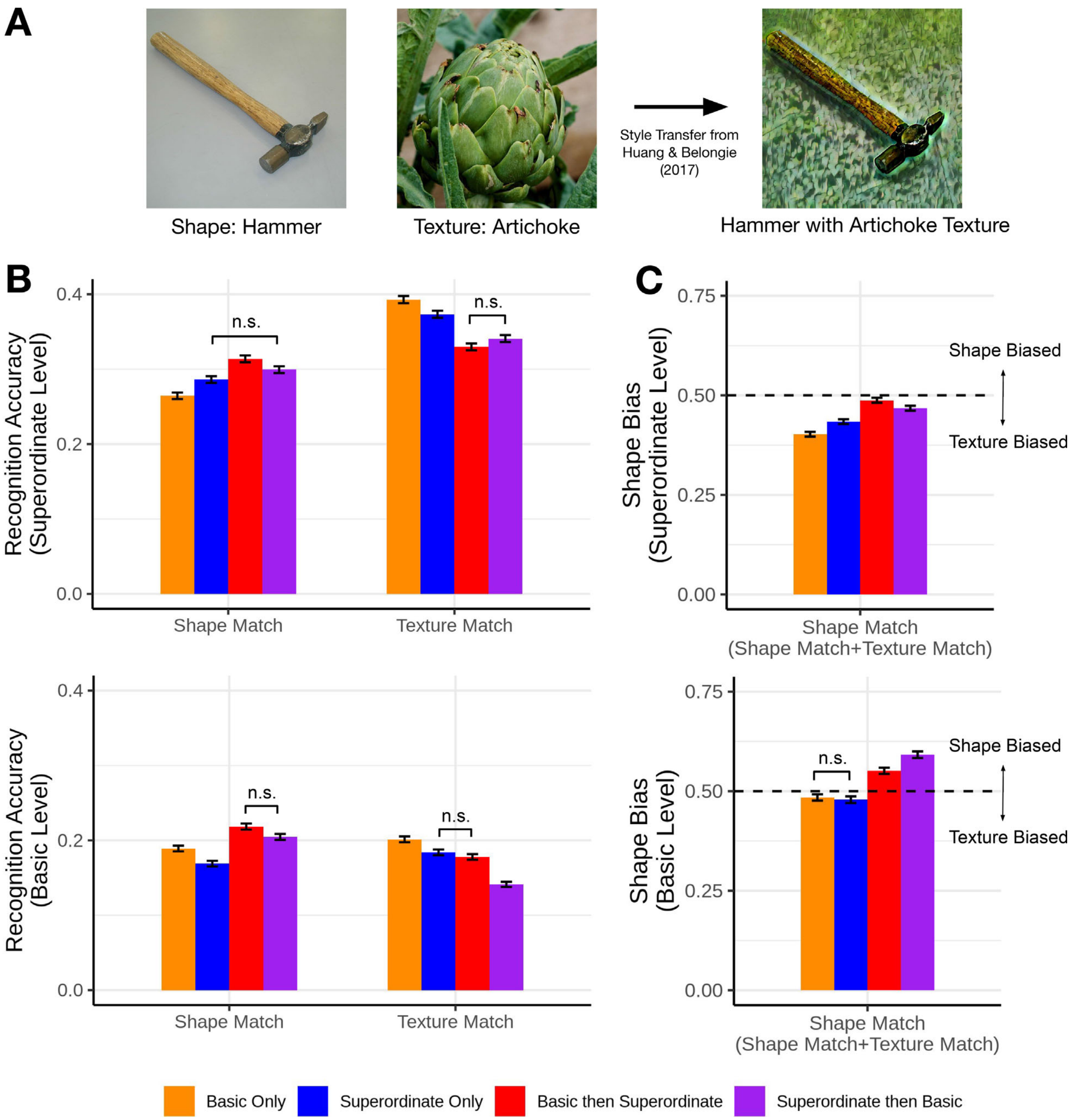

Figure 4. Shape versus texture biases. (A) An example of texture transfer using AdaIN style transfer (Huang & Belongie, 2017). (B) Average recognition accuracy on the texture transferred images for basic and superordinate category classifications. Shape Match: Recognition accuracy when ground-truth categories are based on an object's shape (original image's identity); Texture Match: Recognition accuracy when ground-truth categories are based on an object's texture (texture image's identity). (C) Shape Bias: Relative proportion of images that are correctly classified according to its shape with respect to all images classified either by its shape or texture. To the extent that the Shape Bias score is above 0.50, the model is shape-biased; to the extent that a model is below 0.50, it is texture-biased. All pairwise comparison with Bonferroni corrections were statistically significant at $p < 0.05$ unless specified as n.s.

(Chang et al., 2019), which consisted of neural data collected on more than 5000 images from datasets commonly used for training computer vision models (e.g., ImageNet, COCO, SUN datasets).

The fMRI data

BOLD5000 is a large-scale publicly-available dataset (Chang et al., 2019) consisting of slow event-related

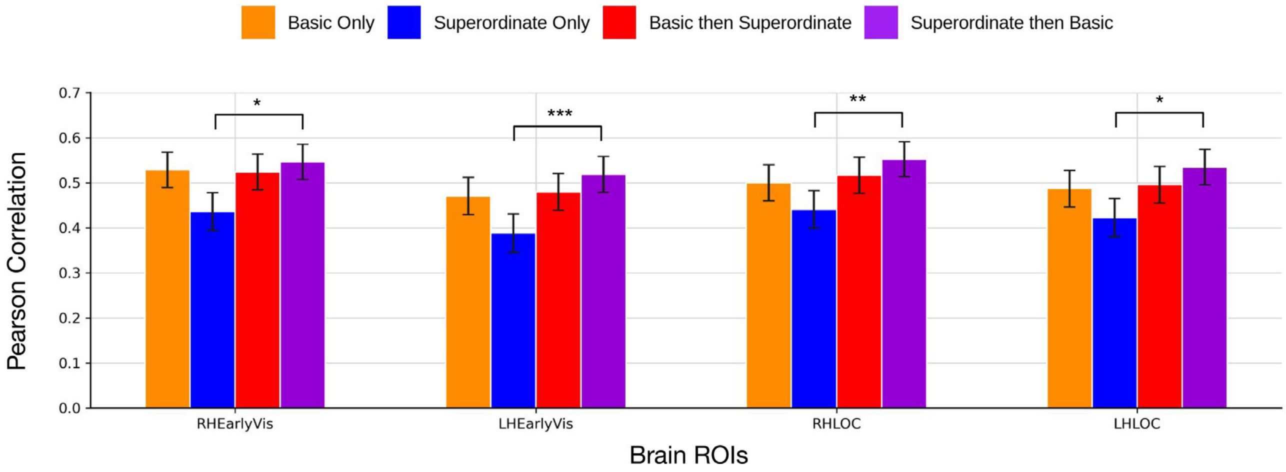

Figure 5. Pearson correlation between RSMs computed for human fMRI activation and each model. Human RSM was calculated by taking an average of RSMs over all participants (n = 3). *Error bars* denote the standard error of the Pearson correlation coefficient. Four different brain ROIs were included in the analysis: earlier visual cortex, including V1 or V2 in the right and left hemisphere (RHEarlyVis and LHEarlyVis, respectively), and LOC in the right and left hemisphere (RHLOC and LHLOC, respectively). All correlation coefficients were significant at 0.05 alpha level. Significance of pairwise differences from post-hoc analyses were signified as *arrows* and *asterisks* above the corresponding bars (* adjusted $p < 0.05$, ** adjusted $p < 0.01$, *** adjusted $p < 0.001$).

fMRI data collected while people viewed 5000 naturalistic images. For the purpose of our study, we only analyzed the neural data recorded for the ImageNet images corresponding to the 30 basic-level categories used to train our models. There were two exemplar images for each category on which blood oxygenation level–dependent (BOLD) activation levels were recorded, and a simple average of these two BOLD signals was taken to represent each individual's categorical visual representation. These categorical visual representations extracted from fMRI data were used to calculate a representational similarity matrix (RSM) comparing the human brain and CNN activation. The full list of image names and their URLs are shared in the Supplementary Material (SM8). Recordings were made from four participants in the original dataset, but we excluded the data from one participant because that person did not complete the entire experimental sessions, and consequently did not view all 30 of our categories of interest. Each image was displayed for one second, followed by a nine-second fixation cross (sampling rate of 0.5 Hz, i.e., TR = 2000 ms), and BOLD signals were recorded during this time (Chang et al., 2019). Brain activity was found to peak from four to eight seconds, so we averaged BOLD signals over TR3 and TR4 and used these for comparison to the models. Whole brains were scanned in the original dataset, but we selected two regions of interest (ROIs) for our analyses. These were early visual cortex and the lateral occipital complex (LOC), brain regions at different levels in the ventral visual stream that are known to play an important role in the recognition of visual objects (Grill-Spector, Kourtzi, & Kanwisher, 2001). Complete descriptions of the experimental design, fMRI recording, and preprocessing pipeline can be found in the original dataset study (Chang et al., 2019).

Method

Architectural and dimensional differences between a CNN and visual cortical responses (human vision) prevent direct comparison of the prototypical representations from the models to the fMRI responses, so to make this comparison we used RSM. RSM is a method initially developed to compare patterns generated from different agents, for example, comparing activation from a human to the neural activity patterns from a monkey (Kriegeskorte, Mur, & Bandettini, 2008). We created a similarity matrix for each model and each brain ROI, where each entry in the RSM was calculated by taking a Pearson correlation between the visual representations of the 30 basic-level categories. We then directly compared how similar their representational patterns were by calculating a Pearson correlation between the two RSMs, one from human fMRI responses and the other from a model's visual prototype representation (See Experiment 1 Method for details).

Results and discussion

We computed an RSM (see Method section for details) for each model and fMRI participant, and computed the Pearson correlation between the two matrices. Doing this for each brain ROI for all three participants and then averaging RSMs over participants gives a value that we will refer to as *human RSM*. Visual comparison between RSMs from each model and humans are provided in the Supplementary Material (SM9). Overall, as shown in Figure 5, superordinate-then-basic again showed the highest similarity, this time in brain activation across different ROIs (average $r = 0.538$, all $p < 0.05$). This was followed by basic-then-superordinate (average $r = 0.503$, all $p < 0.05$) and basic only (average $r = 0.496$, all $p < 0.05$). Superordinate only showed the lowest similarity with humans among the models trained with categorical labels (average $r = 0.422$, all $p < 0.05$). To test whether these correlations are statistically different from one another, we conducted post-hoc pairwise comparisons on the Fisher Z-transformed correlation coefficients, with Bonferroni corrections. Across all brain ROIs in both hemispheres, superordinate-then-basic was the only model whose similarity with human visual representations was significantly higher than superordinate only ($Z_{\text{diff}} = 3.15$, adjusted $p < 0.05$ in RHEarlyVis; $Z_{\text{diff}} = 3.54$, adjusted $P < 0.001$ in LHEarlyVis; $Z_{\text{diff}} = 3.19$, adjusted $p < 0.01$ in RHLOC; $Z_{\text{diff}} = 3.13$, adjusted $p < 0.05$ in LHLOC). Other pairwise comparisons made between the models trained with categorical labels did not show any statistically significant differences (adjusted $p > 0.05$). The higher neural correspondence of superordinate-then-basic model was observed in both natural and human-made categories (see Supplementary Material SM10).

Experiment 5

This experiment evaluated how well models produce human-like visual judgments in a triplet similarity task. In this task, the participant had to select one image that they judged to be the most different among three images of objects shown simultaneously on each trial. We reasoned that this task might engage the human visual representations used in perceptual decision making, thereby providing a salient representational signature in behavior as shown in Hebart, Zheng, Pereira, and Baker (2020). We also wanted a behavioral performance measure to provide contrast to the neural performance measure, in the hope of finding converging evidence for one model or another.

Behavioral data

We collected human similarity judgments in an odd-one-out task, where participants were shown three object images of and asked to choose which was most different from the other two. This paradigm (Roberson, Davidoff, & Braisby, 1999; Zheng et al., 2019) is especially well suited for our goal because we can select the objects in this task to differ in their level in the semantic hierarchy. We did this by varying the number of unique superordinate categories that appeared in a triplet. For example, when all three images in a triplet come from the same superordinate category (e.g., “lemon,” ”orange,” “banana”), the perceived similarity will be compared at the basic level. However, when only two images in a triplet belong to the same superordinate category (e.g., “lemon,” “orange,” “minivan”), the semantic oddity of the other image will focus the similarity comparison at the superordinate level. Each triplet consisted of three exemplar objects from the 30 categories used for our model training. All exemplar images came from Zheng et al. (2019), except for “crate,” “hammer,” “harmonica,” and “screwdriver,” which were replaced with new exemplars to improve image quality and category representativeness. There are 4060 possible triplets that can be generated from all 30 categories, but constraints on behavioral data collection required that we sample only a subset of these. This subset included (1) the ten triplets having objects coming from the same superordinate category (e.g., mammals: “orangutan,” “lion,” “gazelle”), (2) all 435 triplets where two objects came from the same superordinate category (e.g., “orangutan,” “lion,” “minivan”), and (3) 1375 triplets where all objects came from different categories (e.g., “orangutan,” “minivan,” “lemon”), yielding 1820 unique triplets in total. Participants were 51 Amazon Mechanical Turk workers, each making responses on ∼200 triplets (5% , 42%, and 52% of these triplets belong to the subsets 1, 2, and 3, respectively). After removing responses having reaction times below 500 ms, we obtained 9697 similarity judgments where each unique triplet was viewed by 5.6 workers on average.

Method

We simulated responses from each model to each image triplet by calculating cosine similarities between the prototypical category representation for each image and selecting the one most dissimilar from the other two. We computed a lower bound of accuracy (Null accuracy), achieved by predicting every sample with the people’s most frequent choice for the entire dataset, and the effective upper bound (Bayes accuracy) achieved by predicting people’s most frequent choice for each unique triplet. We also report predictions from the SPoSE model (Zheng et al., 2018), where human categorical representations (49 dimensional vectors) are parameterized and estimated from human similarity judgments collected on the entire 1854 categories from the THINGS dataset (Hebart et al., 2019).

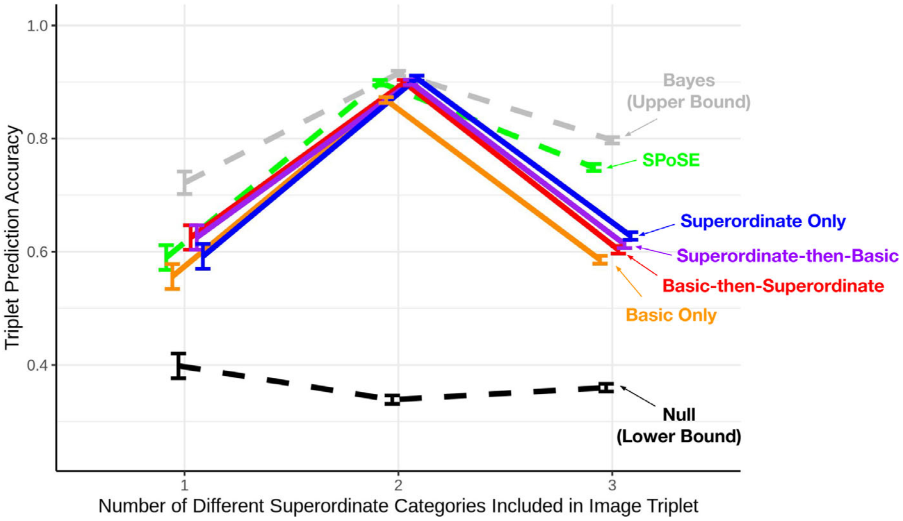

Figure 6. Average prediction accuracy (y-axis) for human similarity judgments for each triplet type according to the number of unique superordinate categories in the triplet (1, 2, or 3, x-axis). For the middle condition, a correct response can be made on the basis of which item is from a different superordinate category. *Error bars* represent standard errors. Null and Bayes accuracy constitutes the lower and upper bounds of performance, respectively. The SPoSE model is the representational embedding model trained on behavioral data collected from Zheng et al. (2018).

Results and discussion

We used GLMMs with binomial family (see Statistical Analyses in Modeling Method section for details), predicting accuracy from triplet-type and model-type. Seven different models (3 baselines + 4 trained models) were coded using treatment contrasts with the Null model (lower limit of performance) as the reference group. Triplet types were also coded using treatment contrasts and the triplet where only two images belong to the same superordinate category (e.g., “lemon,” “orange,” “minivan”) was treated as the reference condition.

Both interaction and main effects for model and triplet type were significant, evidenced by the addition of the corresponding term significantly increasing the likelihood of the statistical model ($\chi^2(12) = 1880.05$, $p < 0.001$ for the interaction effect, $\chi^2(2) = 355.61$, $p < 0.001$ for the triplet-type main effect, $\chi^2(6) = 9367.70$, $p < 0.001$ for the model-type main effect; full coefficient estimates and their significance from the model with an interaction term are provided in the Supplementary Material (SM13). Comparison of marginal means of triplet prediction accuracy estimated from the mixed model in Figure 6 suggests that our trained models predicted human similarity judgments very well, considerably better than Null accuracy. Performances from some of these models were observed to be comparable to the mathematical upper limit (Bayes model) and even the model directly trained with behavioral data (SPoSE model), especially in the condition where there were two unique superordinate categories in a triplet (second column in the figure).

Post-hoc analysis with Bonferroni corrections confirmed these observations. When a triplet consisted of two unique superordinate categories, resulting in the creation of a semantic oddity at a superordinate-level (e.g., “lemon,” “orange,” “minivan”), the models trained with superordinate labels (superordinate only, superordinate-then-basic, and basic-then-superordinate) achieved about 90% prediction accuracy, which was not statistically different from Bayes (92%) and SPoSE model (90%; adjusted $p > 0.05$). Basic only had an accuracy of 87%, which was significantly lower than both the Bayes and SPoSE models ($Z_{\text{ratio}} = -16.31$ adjusted $p < 0.001$; $Z_{\text{ratio}} = -9.35$, adjusted $p < 0.001$; respectively).

Similar trends were observed in a condition where all items in a triplet came from different superordinate categories (e.g., “orangutan,” “minivan,” “lemon”). Although performances of these trained models were generally poorer than the upper baselines, Bayes (80%) and SPoSE (75%), superordinate only achieved the best accuracy (63%) among the trained models, and this accuracy was significantly higher than basic only (59%; $Z_{\text{ratio}} = 5.12$, adjusted $p < 0.001$). There were no statistically meaningful differences between the superordinate only, superordinate-then-basic, and basic-then-superordinate models (adjusted $p > 0.05$). The models whose training included superordinate labels benefited specifically for triplets in which the correct (typical human) response was made on the basis of which image was from a different superordinate category. When all three images in a triplet came from the same superordinate category (e.g., "lemon," "orange," "banana"), we expected that perceived similarity would be compared at the basic level and the models primarily trained with basic-level labels might show an advantage. Surprisingly, experience with superordinate labels was helpful here too, numerically increasing accuracy from 56% obtained by the basic only condition to 62% obtained by both the basic-then-superordinate model and the superordinate-then-basic model. Training with just the Superordinate labels achieved a comparable 59% (adjusted *p* for comparisons of label-training conditions > 0.05).

Summary and general discussion

Despite the remarkable progress in using deep CNNs for classifying images (Rawat & Wang, 2017), their performance still pales in comparison to the robustness of human recognition, which has an outstanding tolerance to changes in object appearances, such as changes in position and size (Ito et al., 1995; Rust & DiCarlo, 2010), viewing direction (Biederman & Gerhardstein, 1993; Vuilleumier et al., 2002), illuminations (Vogels & Biederman, 2002), and contrast (Avidan et al., 2002; Rolls & Baylis, 1986). The aim of the current study was to understand the possible role of category labels in learning more robust visual representations. We hypothesized that the semantic structure of category labels, which is often provided by the labeling hierarchy of our language, contributes to learning flexible and robust categorical representations in humans. To test this idea, we trained multiple CNNs with the same architecture on different types of category labels and conducted an extensive analysis to test the robustness of the visual representations learned by each model. We also identified the models generating the most human-like fMRI and behavioral data during visual tasks.

We found that the models trained with labels at basic and superordinate levels of abstraction learned more robust category-diagnostic visual features compared to models trained at only a single level of abstraction (i.e., only basic or only superordinate). The robustness of the superordinate-then-basic model was particularly impressive, which consistently achieved a significantly higher recognition accuracy than other models over various visual transformations, including changes of position/rotation, illumination, noise, and blurriness. Consistent with this finding, superordinate-then-basic achieved the highest shape bias among our trained models, signifying the model's relatively high tolerance to textural changes. We found that categorical separability for basic-then-superordinate was higher than superordinate-then-basic and other models trained at a single taxonomic level, but these differences were not statistically significant. Previous studies demonstrated that learning non-visual semantic relationships between novel objects, such as one object being "sticky, loud, and nocturnal" and another object being "strong, soft, and friendly" made visual recognition of those objects less viewpoint dependent (Collins & Curby, 2013; Curby et al., 2004). Here we extend these previous findings by showing that learning a taxonomic structure, a basic human semantic association, can also drive significant improvement in visual object recognition.

We also determined the types of labels used for model training that resulted in the learning of visual representations that were most comparable to those of people. We did this by analyzing the representational similarities (Kriegeskorte et al., 2008) between our trained models and visually-evoked responses obtained from the human visual cortex and by comparing the models with respect to their prediction of similarity judgments obtained by people performing an odd-one-out triplet task. Consistent with the robustness analysis above, we found that the superordinate-then-basic model learned representations that were most similar to the visual representations formed across different ROIs in visual cortex, including early visual cortex and LOC that are known to contribute to the identity-preserving visual representation of objects (DiCarlo et al., 2012; Grill-Spector et al., 2001).

We also used our models to predict which image out of three is the most visually different according to people's judgments (Odd-one-out triplet task). The models trained with superordinate labels (superordinate-only, superordinate-then-basic, and basic-then-superordinate) achieved the highest prediction accuracy (~ 90%) for triplets consisting of two unique superordinate categories (e.g., "drum," "harmonica," "refrigerator"). For these trials, both people and the models tended to choose the image with the different superordinate category. Because superordinate-level knowledge is all that is required to answer these trials "correctly," it is not surprising that models trained with just superordinate labels were successful. What *is* surprising is that the superordinate-only training also led to successful performance in other conditions in which each image was from a different superordinate-level category (e.g., "ant," "hammer," "lemon") or the same superordinate-level category (e.g., "coffee-pot," "refrigerator," "toaster'). The model trained only on superordinate class labels was not only more sensitive to global class distinctions compared to models trained on just basic-level labels, but it learned enough within-class structure to predict the performance of people making odd-one-out judgments at the basic-level (when the triplets come from the same superordinate category). Indeed, the model trained only with superordinate-level labels showed considerable basic structure in their learned visual representations according to the t-SNE visualizations, and also performed quite well in classifying basic-level categories (accuracy of 80% when using a linear classifier; Table 1).

Previous studies in computer vision have reported that training with basic-level labels (e.g., dog) helps classifying the finer-grained, subordinate-level objects (e.g., breeds of dog) and attributed the observed benefits to the inherent characteristics of basic-level labels (e.g., informativeness; Peterson et al., 2018; Wang & Cottrell, 2016). Our study suggests that this benefit is not specific to learning basic-level categories but rather reflects a more general advantage of training with coarser-grained labels on fine-grained classification tasks. This idea is further supported by our additional observation of training benefits from superordinate-level labels even when the superordinate categories do not represent the actual semantics of the language. We first trained the model with randomly mixed and thus semantically meaningless superordinate category labels (e.g., “flute,” “lemon,” “lion” would make a new superordinate category A) and then trained with the original 30 basic labels in a second stage. We found that this “superordinate-mixed-then-basic” model performed very well, at times even being comparable to the original superordinate-then-basic model in terms of representational robustness (Supplementary Material, SM6) and correspondence with human brain representations (Supplemental Material, SM11). These findings suggest that training with coarse-level structure is generally helpful in learning more robust and more human-like visual representations of finer-level categories.

Since the seminal work by Rosch et al. (1976), the basic-level advantage has been one of the most-cited and well-known concepts in cognitive psychology. Basic labels are assumed to be easier to learn and access, and for these reasons basic-level object representations dominate our everyday interactions with objects. Mervis and Crisafi (1982) studied how children acquire categories across different hierarchical levels and found higher classification accuracy for basic-level categories compared to superordinate-levels. The authors explained their result by appealing to the relatively low within-category similarity between superordinate-level categories making it difficult to learn the visual regularities needed to group objects at that level compared to at the basic level. However, when we put this idea to the test, the computational experiments from our study suggest the greater visual heterogeneity at the superordinate level may serve a purpose: the supervised learning of superordinate-level categories before basic-categories may improve the robustness of visual recognition compared to training with basic labels alone. In future work we hope to validate the existence of this coarse-to-fine advantage in human visual category learning by testing whether training novel categories with superordinate labels before basic labels increases the visual recognition performance on both unseen and distorted test images.

Why does training with labels spanning levels of abstraction—particularly beginning with superordinate labels and proceeding to basic-level ones—lead to more robust visual representations? One possibility is that, because shared features within superordinate categories are not as salient as in basic-level categories, training with superordinate-level categories first may promote finding features (i.e., regions of the representational space) that are diagnostic of to-be-learned category distinctions (Damiano & Walther, 2019). On the other hand, because the extraction of diagnostic features of basic-level categories is relatively easy (Mervis & Crisafi, 1982; Rosch et al., 1976), training on basic-level categories may discourage exploration by focusing one’s attention more narrowly on the selection of categorical features. Previous studies found that, whereas selective attention enables efficient encoding of stimuli by ignoring category-irrelevant information, a negative consequence of this “learned inattention” is a muted exploration of newly relevant visual features in subsequent learning (Blanco & Sloutsky, 2019; Hoffman & Rehder, 2010), which might have implications for the learning of robust object representations.

One major limitation of our study is that the models we tested are still far from achieving human-like robustness. Although we found that the superordinate-then-basic network was the generally best-performing model, it was still highly vulnerable to image transformations, with recognition accuracy dropping by ∼50% despite levels of changes that are visually insignificant to us (see Supplementary Material, SM3 for examples of images with transformations applied). Note also that the CNNs in our study all had a simple feedforward architecture and used far fewer trainable parameters (∼1 million) than what is typical in the Computer Vision literature (e.g., ResNet50, ∼23 million). We decided for this simplified architecture to more effectively model and analyze the effects of various labelling schemes on the learned visual representations. However, future studies should confirm whether our results generalize to more complex models trained with larger datasets and test whether the use of superordinate-level training in state-of-the-art vision models further reduces the robustness gap between humans and CNNs. One obstacle slowing progress toward this future direction is the unavailability of a dataset having sufficient superordinate labels for model training. Although superordinate-level concepts have long been considered one of the most basic and inherent structures of human semantic knowledge, together with basic-level categories (Murphy, 2016), there does not yet exist a dataset structured as such. For example, ImageNet 2012 (Deng et al., 2009) is one of the most commonly used datasets for training computer vision models, but most of the 1000 categories in ImageNet come from subordinate-level labels, including 120 different dog breeds. We hope that our findings pointing to a superordinate-level training benefit will fuel additional effort into creating a dataset having a more carefully designed semantic structure.

We also investigated how the various label-training conditions compared to conditions lacking supervision from labels altogether, that is, learning categories only from the visual structure. Previous behavioral findings showed that the categorical distinctions made by linguistic labels facilitate the extraction of categorical diagnostic features, as well as the abstraction over irrelevant perceptual information (Althaus & Mareschal, 2014; Edmiston & Lupyan, 2015; Levin & Beale, 2000; Lupyan, Rakison, & McClelland, 2007; Macpherson, 2012; Meteyard, Bahrami, & Vigliocco, 2007; Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009). The success of recent unsupervised methods in computer vision (e.g., contrastive learning; Chen, Kornblith, Norouzi, & Hinton, 2020) suggests that exposure to category labels may not be necessary to learn effective visual representations for human-like classification (Konkle & Alvarez; 2020). Indeed, when we put "SimCLR-Resnet50" unsupervised learning model to the test (Supplementary Material SM7), its recognition was generally more robust to image transformation than our supervised models (except for salt-and-pepper noise, where our supervised models performed better). However, other differences between these models make a direct comparison currently impossible, making it difficult to ascertain the limits of unsupervised training for learning visual structure from visual input alone. Notably, SimCLR–the currently state of the art in unsupervised image category learning–was trained using a more complex convolutional architecture than our supervised models, and was also trained on a much larger dataset (1000 Imagenet categories). When these factors are controlled (e.g., "Conv. AutoEncoder" or "SimCLR-Matched" in the Supplementary Material SM7), the unsupervised models performed far worse than the supervised model. Recent work directly comparing visual representations learned using SimCLR to those that emerge when unsupervised structure is fine-tuned with category labels shows that the latter experience leads to more human-like categorical structure (Luo, Sexton, & Love, 2021). More systematic and controlled experiments will be required to better understand the unique role of category labels on learning robust visual representations.

In this study we explored the effects of linguistic taxonomy on the visual representations learned by CNNs and found that training across a hierarchy, especially in the superordinate-then-basic order, resulted in the learning of robust and human-like visual representations. Consistent with previous findings showing a benefit of learning semantic associations between categories, our results suggest that learning object associations across a simple taxonomic hierarchy can similarly mitigate the challenges imposed on visual object recognition by the various visual transformations imposed by nature on the appearance of objects. Understanding the superior efficiency and flexibility of the human visual system relative to existing artificial systems will likely require extending beyond traditional behavioral science, into domains such as computer vision and robotics. Our work suggests that the semantic structure of labels and datasets should be carefully constructed if the goal is to build vision models that learn visual feature representations having a human-like tolerance to variability. Future research should test whether the benefits of superordinate-then-basic training that we observed here in computational models translates into improved category learning in people.

Keywords: visual object recognition, robustness, CNNs, superordinate labels

Acknowledgments

Supported by NSF-PAC grant 1734260 to G.L. and G.Z.

Commercial relationships: none.
Corresponding author: Seoyoung Ahn.
Email: seoyoung.ahn@stonybrook.edu.
Address: Department of Psychology, Stony Brook University, Stony Brook, NY, USA.

References

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access, 6*, 14410–14430.

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PloS One, 9*(7), e99670.

Annadani, Y., & Biswas, S. (2018). Preserving semantic relations for zero-shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7603–7612.

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., & Malach, R. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *Journal of Neurophysiology, 87*(6), 3102–3116.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*(12), e1006613.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., . . . Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems*, 9448–9458.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *Lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014*.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological review*, *94*(2), 115.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162.

Blanco, N. J., & Sloutsky, V. M. (2019). Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Developmental Psychology*, *55*(10), 2060.

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data, 6*(1), 1–18.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *In International conference on machine learning* (pp. 1597–1607). PMLR.

Collins, J. A., & Curby, K. M. (2013). Conceptual knowledge attenuates viewpoint dependency in visual object recognition. *Visual Cognition*, *21*(8), 945–960.

Curby, K. M., Hayward, W. G., & Gauthier, I. (2004). Laterality effects in the recognition of depth-rotated novel objects. *Cognitive, Affective, & Behavioral Neuroscience, 4*(1), 100–111.

Damiano, C., & Walther, D. B. (2019). Distinct roles of eye movements during memory encoding and retrieval. *Cognition, 184*, 119–129.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227, https://doi.org/10.1109/TPAMI.1979.4766909.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

Dodge, S., & Karam, L. (2017). A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. *ArXiv:1705.02498 [Cs]*, http://arxiv.org/abs/1705.02498.

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100, https://doi.org/10.1016/j.cognition.2015.06.008..

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., . . . Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 2121–2129.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202, https://doi.org/10.1007/BF00344251.

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology, 20*(3–6), 507–523.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems* (2018).

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, https://openreview.net/forum?id=Bygh9j09KX.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(1), 69–78.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41*(10–11), 1409–1422.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision* *and Pattern Recognition (CVPR)*, 770–778, https://doi.org/10.1109/CVPR.2016.90.

Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., . . . Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS One, 14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour, 4*(11), 1173–1185.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *ArXiv:1903.12261 [Cs, Stat]*, http://arxiv.org/abs/1903.12261.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139*(2), 319.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1501–1510).

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology, 73*(1), 218–226.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience, 22*(6), 974.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Konkle, T., & Alvarez, G. A. (2020). Instance-level contrastive learning yields human brain-like representation without category-supervision. *BioRxiv*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, https://doi.org/10.3389/neuro.06.004.2008.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

Lei Ba, J., Swersky, K., & Fidler, S. (2015). Predicting deep zero-shot convolutional neural networks using textual descriptions. *Proceedings of the IEEE International Conference on Computer Vision*, 4247–4255.

Levin, D. T., & Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception & Psychophysics, 62*(2), 386–401.

Luo, X., Sexton, N. J., & Love, B. C. (2021). A deep learning account of how language affects thought. Language, Cognition and Neuroscience, https://doi.org/10.1080/23273798.2021.2001023.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science, 18*(12), 1077–1083.

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), 2579–2605.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 24–62.

Mandler, J. M., Bauer, P. J., & McDonough, L. (1991). Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology, 23*(2), 263–298, https://doi.org/10.1016/0010-0285(91)90011-C.

Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, 258–266.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs: Language affects low-level visual perception. *Psychological Science, 18*(11), 1007–1013.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Murphy, G. (2016). Explaining the Basic-Level Concept Advantage in Infants … or Is It the Superordinate-Level Advantage? In *Psychology of Learning and Motivation* (Vol. *64*, pp. 57–92). Philadelphia: Elsevier.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. *Knowledge, Concepts, and Categories*, 93–131.

Peterson, J. C., Soulos, P., Nematzadeh, A., & Griffiths, T. L. (2018). Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *ArXiv Preprint ArXiv:1805.07647*.

Plaut, D. C., & Farah, M. J. (1990). Visual object representation: Interpreting neurophysiological data within a computational framework. *Journal of Cognitive Neuroscience, 2*(4), 320–343.

Posner, M. I. (1970). Abstraction and the process of recognition. In *Psychology of learning and motivation* (Vol. *3*, pp. 43–100). Philadelphia: Elsevier.

Quinn, P. C., & Johnson, M. H. (2000). Global-before-basic object categorization in connectionist networks and 2-month-old infants. *Infancy, 1*(1), 31–46.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation, 29*(9), 2352–2449.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*(11), 1199–1204.

Roberson, D., Davidoff, J., & Braisby, N. (1999). Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition, 71*(1), 1–42.

Rolls, E., & Baylis, G. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research, 65*(1), 38–48.

Rolls, E. T. (1994). Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes, 33*(1–2), 113–138.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience, 30*(39), 12978–12995.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 4077–4087.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., . . . Fergus, R. (2013). Intriguing properties of neural networks. *ArXiv Preprint ArXiv:1312.6199*.

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science, 1*(4), 253–256.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder?. *Cognitive Psychology*, *23*(3), 457–482.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences, 106*(11), 4567–4570.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113*(2), 169.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition, 32*(3), 193–254.

Vogels, R., & Biederman, I. (2002). Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. *Cerebral Cortex, 12*(7), 756–766.

Vuilleumier, P., Henson, R., Driver, J., & Dolan, R. J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience, 5*(5), 491–499.

Wang, P., & Cottrell, G. W. (2016). Basic level categorization facilitates visual object recognition. *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico. Workshop Track Proceedings*.

Zheng, C. Y., Pereira, F., Baker, C. I., & Hebart, M. N. (2019). Revealing interpretable object representations from human behavior. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA. Conference Track Proceedings*, https://openreview.net/forum?id=ryxSrhC9KX.