Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Scene context is predictive of unconstrained object similarity judgments

Caterina Magri, Eric Elmoznino, Michael F. Bonner [*]

Department of Cognitive Science, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, United States of America

ARTICLE INFO

Keywords:
Contextual associations
Objects
Scenes
Similarity
Convolutional neural networks
Natural image statistics

ABSTRACT

What makes objects alike in the human mind? Computational approaches for characterizing object similarity have largely focused on the visual forms of objects or their linguistic associations. However, intuitive notions of object similarity may depend heavily on contextual reasoning—that is, objects may be grouped together in the mind if they occur in the context of similar scenes or events. Using large-scale analyses of natural scene statistics and human behavior, we found that a computational model of the associations between objects and their scene contexts is strongly predictive of how humans spontaneously group objects by similarity. Specifically, we learned contextual prototypes for a diverse set of object categories by taking the average response of a convolutional neural network (CNN) to the scene contexts in which the objects typically occurred. In behavioral experiments, we found that contextual prototypes were strongly predictive of human similarity judgments for a large set of objects and rivaled the performance of models based on CNN representations of the objects themselves or word embeddings for their names. Together, our findings reveal the remarkable degree to which the natural statistics of context predict commonsense notions of object similarity.

1. Introduction

Our understanding of the similarities between objects allows us to group them in meaningful ways and underlies behaviors ranging from freeform associative thinking to the creation of scientific taxonomic systems (Goldstone & Son, 2012). Because objects can be compared in myriad ways, there is no single definition of similarity—rather, it is highly task-dependent. Nonetheless, when humans reason about objects, the similarities that spontaneously come to mind are largely in agreement across individuals, demonstrating that there is systematicity in our intuitions about what makes objects alike (Hebart et al., 2019; Peterson et al., 2018).

Similarity judgments are of broad interest in the cognitive, neural, and computational sciences because they provide a window into the mental representation of concepts. They have been used to test theories of concept learning and generalization (Shepard, 1987; Shepard & Arabie, 1979), to characterize the representational dimensions of the mind (Greene et al., 2016; Hebart et al., 2019, 2020; Jozwik et al., 2017; Mur et al., 2013; Peterson et al., 2018; Tarhan et al., 2021), to identify the neural bases of concepts (Dima et al., 2022; King et al., 2019; Mur et al., 2013; Tarhan et al., 2021), and to assess the quality of object and word representations in computational systems (Baroni et al., 2014; Jozwik et al., 2017; Roads & Love, 2020). Across all these disciplines, a major focus has been on the similarity responses that observers spontaneously provide when task instructions do not specify the dimensions along which stimuli should be compared. These spontaneous similarity judgments are of primary theoretical interest because they reflect the core representational dimensions that underlie our intuitive understanding of the relationships among objects. It is these intuitions about object similarity that we set out to model in the current work.

In seeking to understand how similarities are represented in the mind, a key objective is the development of computational models that can account for human behavior (Hebart et al., 2019, 2020; Jozwik et al., 2017; Peterson et al., 2018). Computational models ground psychological theories in the language of mathematics, they provide insight into the statistical and geometric principles of mental representations, and they are well-suited for characterizing the complex, multifaceted properties of natural stimuli. To model the similarities between objects, previous computational approaches have largely focused on convolutional neural network (CNN) representations of object images, which characterize their high-level visual features, or word embeddings for object names, which characterize their co-occurrence with other words in written language (Hebart et al., 2019, 2020; Jozwik et al., 2017; Peterson et al., 2018). However, these approaches have overlooked a major component of object representations in the mind: contextual associations.

Objects are often encountered in a highly regular set of scene contexts (e.g., refrigerators are often found in kitchens). There is both

[*] Corresponding author.
E-mail address: mfbonner@jhu.edu (M.F. Bonner).

https://doi.org/10.1016/j.cognition.2023.105535
Received 20 September 2022; Received in revised form 21 June 2023; Accepted 23 June 2023
Available online 24 July 2023

psychological and neural evidence that humans use these contextual regularities to facilitate object recognition and to generate predictions during perception (Aminoff & Tarr, 2015; Bar & Aminoff, 2003; Biederman et al., 1982; Bonner & Epstein, 2021; Davenport & Potter, 2004; Lauer et al., 2021; Oliva & Torralba, 2007; Palmer, 1975). Furthermore, recent work has shown that CNNs trained on object classification also learn this contextual information, and they do so in a manner that is highly consistent with explicit human judgments of contextual similarity (Aminoff et al., 2022; Bracci et al., 2022). Despite the importance of contextual regularities in object perception, we do not yet have answers to fundamental questions about the role of context in the kinds of intuitive similarity judgments that are often used to assess object representation (Greene et al., 2016; Hebart et al., 2019, 2020; Jozwik et al., 2017; Mur et al., 2013; Peterson et al., 2018; Tarhan et al., 2021). First, to what degree do these contextual regularities explain human behavior in intuitive similarity judgments? And second, how can these regularities be computationally modeled?

Here we set out to answer these questions by determining whether human behavioral judgments of object similarity could be explained by the natural statistics of object context. To accomplish this, we developed a novel computational approach for characterizing the rich contextual associations of natural objects, and we collected spontaneous judgments of object similarity for a diverse set of object categories. Specifically, our modeling approach leveraged a large dataset of annotated images to quantify the co-occurrences between objects and scene contexts. We used this dataset to construct contextual-prototype representations based on CNN responses to the scene contexts themselves. These contextual prototypes capture the rich visual and semantic information that can be learned from the scenes in which an object is typically encountered. For comparison, we also examined object-based representational models constructed from CNN responses to the objects themselves or from word embeddings for their names in language.

To anticipate our results, we found that the statistical regularities captured by our contextual model were highly predictive of spontaneous behavioral judgments of object similarity and rivaled object-based models that used images of the objects or their associated word embeddings. Furthermore, we observed a dissociation under different task conditions: while the contextual model performed well at explaining spontaneous similarity judgments, it was outperformed by the object-based models at explaining shape similarity judgments. Thus, our findings show that contextual prototypes specifically emphasize the core representational dimensions that come to mind when participants intuitively group objects based on their similarities. These findings reveal the remarkable degree to which the natural statistics of context can predict human intuitions about object similarity.

2. Methods

2.1. Behavioral experiments

2.1.1. Overview

We sought to explain how humans spontaneously group objects according to their similarities. To do so, we collected behavioral data in the form of a multi-arrangement task, in which participants were asked to arrange images of objects on a 2D display, such that nearby objects are more similar (Fig. 1). We were primarily interested in how participants judged similarity in an unguided task without being instructed to focus on any specific object attributes. For our first experiments, we therefore did not instruct participants on how to evaluate similarity, other than explaining the basic mechanics of the task. We refer to this as an intuitive similarity task because it relies on the participants' intuitions about the core factors that define similarity. The object images came from a previously published study and included 81 categories of objects encountered in everyday real-world scenes (Bonner & Epstein, 2021). The object categories in this stimulus set were specifically chosen because they appear in the image annotations of the ADE20K scene database, which we used to model contextual associations (Zhou et al., 2019). We also conducted follow-up experiments with new participants and stimuli to determine if our findings could generalize to objects sampled from a more object-centric database—specifically, the COCO database (Lin et al., 2015). Thus, throughout the manuscript, we report findings from two sets of experiments using object categories sampled from ADE20K and COCO. The categories from both stimulus sets are listed in Supplementary Table 1.

While our primary hypothesis was that intuitive judgments of similarity could be explained by context, we also predicted that these intuitive judgments would differ from another salient form of similarity: the similarity of object shapes. Shape is a fundamental property of objects that is strongly represented throughout the ventral stream of visual cortex (Bracci & Op de Beeck, 2016; Kourtzi & Kanwisher, 2000; Proklova et al., 2016). Local shape features are also encoded in CNN representations (Kubilius et al., 2016; Zeman et al., 2020), and shape has been a major focus in the computational modeling of object similarity (Hebart et al., 2019; Jozwik et al., 2017; Peterson et al., 2018). We, therefore, conducted a second multi-arrangement experiment in which participants were asked to arrange images of objects on a 2D display based specifically on the similarity of their shapes (Fig. 1). The stimuli and methods for this experiment were identical to the intuitive-similarity task, except for the instructions and practice trials, which asked participants to specifically evaluate shape similarity.

2.1.2. Participants

80 participants were recruited through the online platform Prolific (https://www.prolific.co/) and redirected to perform a multi-arrangement task on the Meadows Research platform (https://meadows-research.com/). Four separate experiments were performed: two experiments with object categories from ADE20K and two experiments with object categories from COCO. Twenty participants were assigned to each experiment. Participants gave informed consent in compliance with procedures approved by the Institutional Review Board at Johns Hopkins University. One participant from the shape-guided COCO experiment was excluded before data analysis due to technical issues that prevented the correct retention of their data.

2.1.3. Stimuli

ADE20K stimulus set. The stimuli consisted of 81 inanimate object categories, with one item per category (all categories are listed in Supplementary Table 1). All stimuli depicted isolated objects on a white background. The images were taken from a previously developed stimulus set (Bonner & Epstein, 2021). The object categories in this experiment were chosen because they can also be found in the ADE20K database of scene images, which have been densely segmented and annotated for their constituent objects (Zhou et al., 2014) (https://groups.csail.mit.edu/vision/datasets/ADE20K/).

COCO stimulus set. The stimuli consisted of 77 object categories (including both inanimate and animate categories), with one item per category (all categories are listed in Supplementary Table 1). All stimuli depicted isolated objects on a white background. The object categories in this experiment were chosen because they can also be found in COCO (https://cocodataset.org/#home), a large-scale object image database (Lin et al., 2014). Images for these object categories were hand-selected based on a search of freely available images.

2.1.4. Multi-arrangement task

Participants completed a multi-arrangement task adapted from Kriegeskorte and Mur (2012). Objects from the stimulus set were presented to the right and left of a circular arena. Participants were required to drag the objects inside the arena according to their similarity, so that more similar objects were closer together and more dissimilar objects were farther apart. We ran two separate versions of this experiment for each of the two object datasets, for a total of

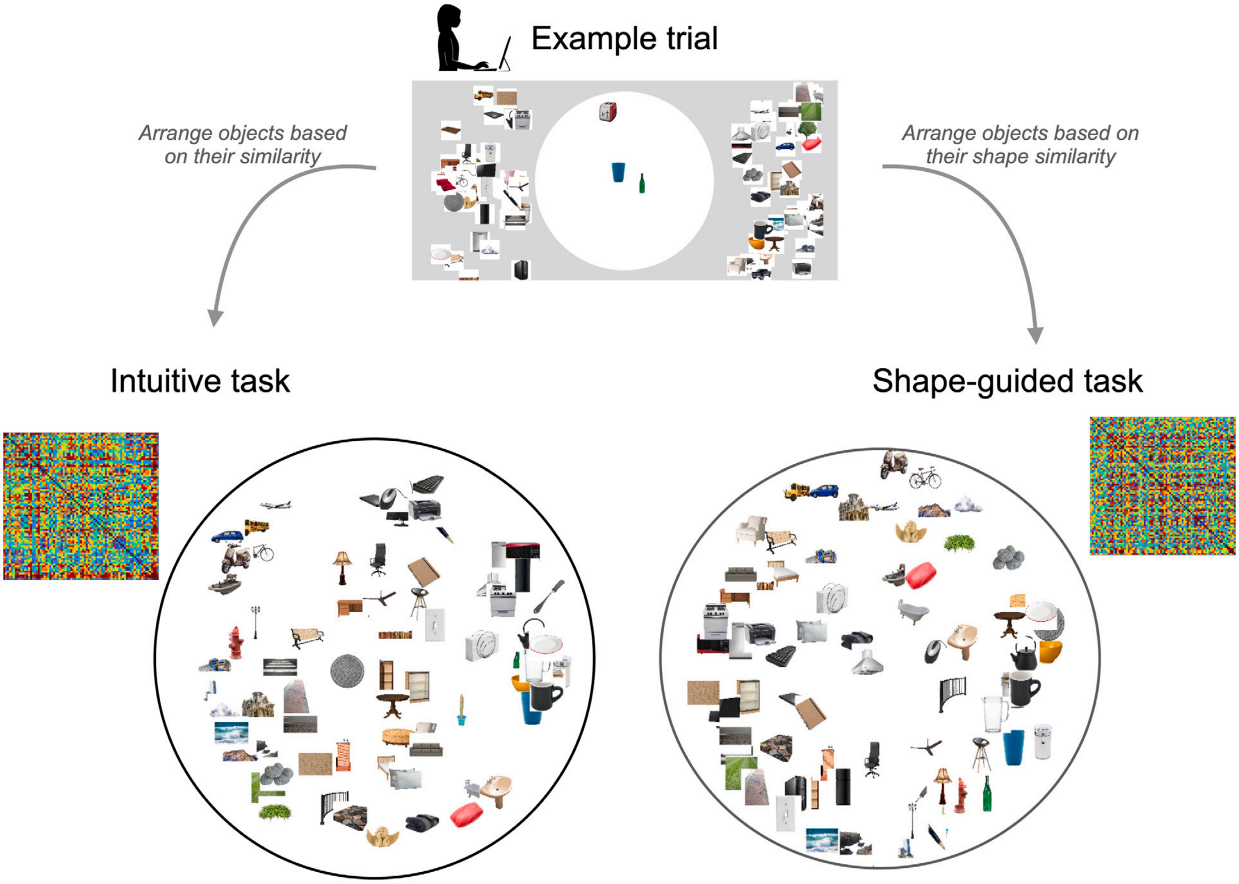
Example trial
Arrange objects based on their similarity
Arrange objects based on their shape similarity
Intuitive task
Shape-guided task

Fig. 1. Multi-arrangement tasks for intuitive similarity and shape similarity. Top: Example trial of the multi-arrangement task for the ADE20K experiments. Lower left: Behavioral RDM based on data from the intuitive similarity task, along with a 2D visualization of this similarity space (using multidimensional scaling). Lower right: Behavioral RDM based on data from the shape similarity task, along with a 2D visualization.

four separate experiments with independent groups of participants. The two versions of this experiment differed on how the participants were instructed to judge similarity:

- *Intuitive sorting task*. Participants were instructed to arrange objects by similarity. We did not specify which factors to consider when evaluating similarity.
- *Shape sorting task*. Participants were instructed to arrange objects by their shape similarity, and they were asked to ignore all other aspects of similarity.

In a pilot study of the shape sorting experiment, we observed that participants would often dismiss shape and resort to intuitive similarity if not thoroughly instructed. Thus, we provided an additional practice trial before the start of the experiment in which participants had to arrange four simple geometrical shapes (a blue square, a yellow hexagon, a red parallelogram, and a red circle) based on shape similarity. After the practice trial, feedback was provided by displaying one of the possible meaningful arrangements that captured shape similarity. Adding this additional practice trial at the beginning ensured that participants clearly understood the task before starting the experiment.

All experiments took place over several trials. Participants could advance to the next trial only when all objects were inside the arena. On the first trial, all objects were presented, while on the subsequent trials the number of objects presented at once could vary based on the accumulated weighted evidence estimated for each object pair (as described in Kriegeskorte & Mur, 2012). At the end of each trial, the object pairs for which the weighted evidence was the least were selected to be presented on the next trial. This procedure was repeated for each trial until the object pair with the smallest evidence weight was above the evidence weight threshold of 0.5 (selected based on prior literature), at which point the experiment terminated. If such a threshold was not reached, the experiment would terminate after 60 min, excluding time for breaks. For the ADE20K dataset, the 0.5 threshold was reached by all participants for the shape task, and by 11 of the 20 participants for the intuitive task. For the COCO dataset, the 0.5 threshold was reached by 12 of the 19 participants for the shape task, and by 12 of the 20 participants for the intuitive task. The output data for each experiment consisted of the lower triangle of a matrix reflecting the distance between each pair of the object categories inferred from the distances between images in the experimental trials (inverse multidimensional scaling algorithm, see Kriegeskorte & Mur, 2012 for details). We refer to these distance matrices as representational dissimilarity matrices (RDMs). For our main analyses, we averaged the RDMs across participants to produce a group-level RDM for each experiment. We additionally report participant-level analyses using the individual participant RDMs in the Supplementary Material.

2.2. Computational models

2.2.1. Overview of image-based models

We set out to test the hypothesis that intuitive judgments of object similarity could be explained by the contextual regularities of the environment, such that objects that tend to occur in the same contexts would be judged as similar. To accomplish this, we developed a modeling approach for characterizing the high-level contextual associations of objects based on the scene contexts in which they are typically encountered. We took advantage of the ADE20K database (Zhou et al., 2019), which contains 22,210 natural scene images with corresponding

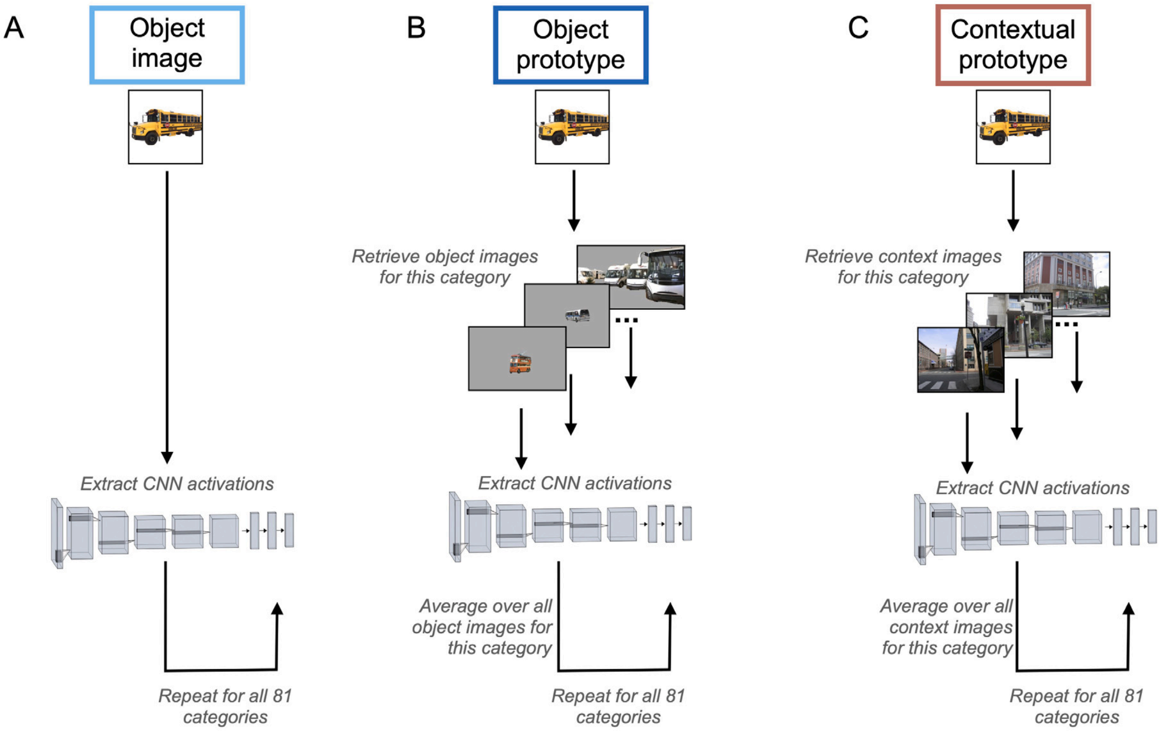
A
Object image
Extract CNN activations
Repeat for all 81 categories
B
Object prototype
Retrieve object images for this category
Extract CNN activations
Average over all object images for this category
Repeat for all 81 categories
C
Contextual prototype
Retrieve context images for this category
Extract CNN activations
Average over all context images for this category
Repeat for all 81 categories

Fig. 2. Image-based representational models. This schematic depicts the extraction of CNN features for the object-image model (A), the object-prototype model (B), and the contextual-prototype model (C). The stimuli shown here come from the ADE20K experiments. For all models, we extracted feature activations from an AlexNet CNN pre-trained on ImageNet. Features were extracted at each layer of AlexNet and concatenated across all layers. RDMs were created by computing the squared Euclidean distances of the feature vectors for all pairs of object categories. For the object-image model, we extracted CNN activations for the exact images that were presented to participants in the behavioral experiments. For the object-prototype model, we sought to obtain an average CNN representation of object-related features based on many instances of each object category. To accomplish this, we retrieved images of each object category from the ADE20K or COCO databases, masked the image backgrounds to leave only the segmented object intact, and extracted the average CNN activation vectors across all images for each category. For the contextual-prototype model, we sought to obtain an average CNN representation of context-related features for each object category. To accomplish this, we used the ADE20K database to retrieve sample images of the scene contexts associated with each object. Importantly, these context images never contained the target object. We then extracted the average CNN activation vectors across all context images for each object.

human annotations for their constituent objects (e.g., notebook, desk) and overall scene categories (e.g., office). For each object category, we obtained CNN responses to sample images from its associated scene contexts, and we computed the average CNN representation across these context images, weighted by the frequency of association between the object and scene categories. Importantly, the context images never contained the object being modeled. Instead, we selected other images from the same scene categories that happened to not contain the object of interest. For example, when obtaining context images for the object *notebook*, we only used images of offices and other related contexts that happen to not contain notebooks. Thus, our modeling procedure emphasizes information associated with the scenes in which the objects are found rather than the features of the objects themselves.

Our context representations are based on CNN responses to the context images for each object category. CNNs are the leading computational models of the human visual system, and they are well-suited for extracting high-level visual representations from natural images (Agrawal et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). We used an AlexNet CNN pre-trained on ImageNet, though, as we report in the Results, similar findings were obtained with a different pre-training dataset. The average CNN representations generated from our procedure are statistical summaries of the scene contexts in which objects are encountered, and we refer to these representations as contextual prototypes (Fig. 2C).

For comparison, we also created two object-based CNN models (Fig. 2A and 2B). The first was based on CNN representations of the actual object images used in the behavioral experiments, which we refer to as the object-image model. This model closely resembles how others have used CNNs in previous work to characterize object similarity (Hebart et al., 2019; Jozwik et al., 2017; Peterson et al., 2018). The second was an object-prototype model, which was a complement to our contextual-prototype model. For each object, we used the ADE20K or COCO image database to retrieve all images containing instances of the object category. We then used segmentation data to mask all pixels except for those containing the object (see examples in Fig. 2). We passed these masked images to the CNN and computed the average across all instances to generate an object prototype representation. Much like the contextual prototypes, these object prototypes reflect a statistical summary across a large sample of images. However, they emphasize the features of the objects themselves rather than the contexts in which they are encountered.

2.2.2. Image-based model details

For all three image-based models, we extracted feature activations from ImageNet-pre-trained AlexNet (Pytorch implementation) at each layer. For the convolutional layers, we applied global max-pooling to focus on feature properties rather than spatial properties. For our main analyses, these feature activations were concatenated across layers (all 5 convolutional layers and 3 fully connected layers) and used to build an RDM reflecting feature information across the entire network hierarchy. We also report layerwise versions of these models in the supplement in which we constructed separate RDMs for each layer. All model RDMs were constructed by computing pairwise squared Euclidean distances of feature-activation vectors. The feature activations were first z-scored by items and by features before building the RDM.

For the object image model (Fig. 2A), all images from the ADE20K stimulus set and the COCO stimulus set were submitted to AlexNet to obtain feature activations and to compute RDMs.

For the object prototype model (Fig. 2B), we took advantage of the way in which the object categories in our stimulus set were selected:

the ADE20K object categories were obtained from the ADE20K image database, and the COCO object categories were obtained from the COCO image database. ADE20K comprises 22,210 images containing labels for 3148 unique object labels and 871 unique scene labels. Each image is densely annotated by a human observer, providing segmentations and labels for objects in every scene. For each object category in this stimulus set, we used the set of labels from ADE20K that were specified in Bonner and Epstein (2021), which included variations of the object names (e.g., bath, bath tub, bathtub) as well as plural forms. COCO comprises 330 K images from 80 object categories and 91 stuff categories, with segmentation masks for each object class. To create object-prototype representations, we identified all images containing a given object and used the corresponding object segmentation data to mask all pixels not containing the object with a uniform gray background. When multiple instances from the same object category were present in an image, all instances were left intact and other pixels were masked. These masked images were submitted to AlexNet to obtain feature activations, which were then averaged across all images associated with each object category. An RDM was built based on the distances between the average category-level activations.

For the contextual prototype model (Fig. 2C), we sought to examine images of the scene categories in which the objects are typically encountered while avoiding images that contained instances of the objects themselves. To accomplish this, we took advantage of the fact that the images in the ADE20K database contain labels for both the objects within each scene and the category of the scene itself. Specifically, for each object category, we identified all images with an instance of that object and counted the frequency of these images in each scene category (e.g., the object *pen* occurred in 100 *office* scenes, 10 *kitchen* scenes, and so on). For each of these images, we obtained an alternative image with the same scene-category label but with the constraint that the alternative image could not contain an instance of the target object (e.g., we obtained 100 office scenes without a pen, 10 kitchen scenes without a pen, and so on). Beyond these constraints, the alternative images were chosen randomly. If ADE20K did not include enough alternative scene images without the target object, we randomly sampled with replacement from the available images within a scene category (e.g., if we needed 100 offices without pens but there were only 90 such office images, we randomly sampled 100 images with replacement from the 90 available images). This ensured that the frequency of each scene category was preserved. The resulting set of images served as context images for the target object. Contextual prototypes were created by passing these context images to AlexNet and performing the same pooling and averaging procedures as in the object-prototype model. The average feature activations from AlexNet were then used to compute an RDM.

To compute contextual-prototype representations in the same way for the COCO stimulus set, we needed to combine and remove some of the object categories. This became necessary because we changed our procedure for computing contextual prototypes after we had already collected behavioral data for the COCO experiments. Specifically, in our original procedure for computing contextual prototypes, we created a set of context images by masking the targets objects in their original images rather than by obtaining alternative images without the target objects. However, we wanted to rule out the possibility that our findings could be explained by object shape properties that might be preserved by the masking procedure, such as size, aspect ratio, and coarse shape features. We therefore developed a cleaner approach of obtaining alternative images without the target objects (as described for the stimuli in the ADE20K experiments), which allowed us to completely avoid using masks. For this updated contextual-prototype procedure, we needed to obtain alternative context images from ADE20K for each object category in the COCO stimulus set. Some of the COCO object labels were not present in ADE20K or were only present in a small number of images. To overcome these issues, we reduced the original 77 categories in our COCO stimuli to a new set of 48 categories, each of which occurred in at least 49 images in the ADE20K dataset (49 being the lowest number of images we had across all categories in the ADE20K experiments). Supplementary Figure S1 specifies how we created this reduced set of 48 object categories, which involved combining and removing some of the original 77 categories. To match the RDMs across all COCO analyses, we reduced all behavioral and model RDMs of COCO stimuli to the same set of 48 categories.

We directly compared these representational models by computing the correlations of their RDMs (see Supplementary Fig. S2 for a visualization of the RDMs). We found that the correlation of the contextual-prototype model with each of the object-based models was lower than the correlation between the two object-based models, suggesting that the contextual prototypes capture information that is partially distinct from representations of the objects themselves (contextual prototype and object image: $\rho = 0.20$; contextual prototype and object prototype: $\rho = 0.19$; object prototype and object image: $\rho = 0.33$). We observed a similar result for the COCO objects (contextual prototype and object image: $\rho = 0.18$; contextual prototype and object prototype: $\rho = 0.22$; object prototype and object image: $\rho = 0.52$).

2.2.3. Distributional models

We examined models of the distributional semantics of our categories based on word co-occurrence in language, using word2vec (Mikolov, Chen, et al., 2013) and object co-occurrence in images, using object2vec (Bonner & Epstein, 2021). For word2vec, we obtained a set of 300-dimensional embeddings trained on the Google News corpus (https://code.google.com/archive/p/word2vec/). For object2vec, we obtained a set of 8-dimensional embeddings trained on the image annotations from ADE20K, which contain information about the co-occurrence of object labels in densely annotated natural scenes (https://osf.io/ug5zd/). Details of the object2vec model can be found in Bonner and Epstein (2021). We computed RDMs for word2vec and object2vec in the same manner as our CNN models, by first z-scoring across items and channels and then computing pairwise squared Euclidean distances.

2.3. Representational similarity analysis

To compare the representations of our computational models to the behavioral similarity data, we performed representational similarity analysis (RSA) by calculating correlations (Spearman's ρ) of the model and behavioral RDMs. Bootstrap-resampling distributions of these correlations were calculated over 5000 iterations in which the rows and columns of the RDMs were randomly resampled. This is effectively a resampling of the stimulus labels in the RDM. Resampling was performed without replacement by subsampling 90% of the rows and columns of the RDMs. Statistical significance was assessed through a permutation test over 5000 iterations in which the rows and columns of one of the RDMs were randomly permuted and a correlation coefficient was calculated to generate a null distribution. The p-value for an observed RSA correlation was computed as the ranked percentile of the correlation coefficient relative to this null distribution.

2.3.1. Noise ceiling estimation

Noise ceilings for the RDMs from the behavioral tasks were computed by performing an iterative (n = 5000) split-half reliability analysis. At each iteration, participants were split into two random groups and their average RDMs were correlated using Spearman's rank correlation coefficient (ρ), with the correlation value corrected using the Spearman–Brown prophecy formula. We found that the RDMs were highly reliable across participants for all four behavioral experiments (average split-half reliability across 5000 iterations: $\rho = 0.68$ for ADE20K intuitive task, $\rho = 0.68$ for COCO intuitive task, $\rho = 0.64$ for ADE20K shape task, $\rho = 0.65$ for COCO shape task).

2.3.2. Model comparisons

We tested differences in the RSA correlations for different models using a permutation procedure. For each pair of models, we computed the true difference in their RSA correlations with a behavioral RDM. We then randomly shuffled the behavioral RDM, correlated it with the two model RDMs, and computed the difference between the two RSA correlations. This operation was repeated 5000 times to build a null distribution of difference scores. We then calculated the p-value of the true difference relative to this null distribution.

2.3.3. Variance partitioning analyses

We used multiple linear regression and a variance-partitioning procedure to quantify the overlap of explained variance for the predictor RDMs. The multiple linear regression analyses included regressors for the predictor RDMs and a constant term. These regressors were used to explain variance in the behavioral RDMs, which served as dependent variables. Thus, the data points for the dependent and independent variables were the pairwise distance measurements of the RDMs. We quantified the overlap of explained variance for the predictor RDMs using a procedure known as commonality analysis (Bonner & Epstein, 2018; Nimon et al., 2008; Nimon & Oswald, 2013). This procedure partitions the explained variance of the regression model into the shared and unique components contributed by the predictor RDMs. The variance partitions were calculated using standard partitioning formulas for three predictor variables (Nimon et al., 2008, 2017). To express the variance partitions as percentages, we divided each partition by the total explained variance for all three predictor RDMs combined.

3. Results

3.1. Contextual prototypes are strong predictors of intuitive similarity judgments

To determine whether the object similarities in our representational models matched the similarities obtained from human behavioral data, we computed correlations between the RDMs for our models and the RDMs for the multi-arrangement tasks. We first examined data from the intuitive-similarity task, in which participants spontaneously grouped objects based on their intuitions about similarity (Fig. 1). For our initial experiment using objects from ADE20K, we found that the contextual-prototype RDM was significantly correlated with the behavioral RDM for the intuitive similarity task and outperformed both of the object-based CNN models (Fig. 3A; see Supplementary Fig. S3A for participant-level results). Further, we observed that the object-prototype model performed worse than both the object-image and contextual-prototype models. We also examined layerwise results for all models by using individual layers of AlexNet as feature vectors (Supplementary Fig. S4A). The contextual prototype model was highly consistent across all layers, which aligns with previous work suggesting that contextual information can be captured by low-level scene statistics (Oliva & Torralba, 2007), whereas the object models generally showed a slight improvement at higher layers.

While we hypothesized that the contextual-prototype model would be competitive with object-based models, we were surprised to observe such a strong advantage for the contextual prototypes. We wondered if this result could be related to two properties of the ADE20K database. The first is that the definition of "object" in ADE20K is broad and includes not only discrete objects, like mug and car, but also extended surfaces, like road and grass, which may have much stronger contextual associations than discrete objects. The second is that many of the objects in ADE20K are visually small, given that they occur in the context of large-scale scenes (e.g., a pen is a small visual element of an office scene). As a result, the performance of the object-prototype model could potentially be hampered by the use of zoomed-out object images from ADE20K. To address these issues, we ran a separate experiment using a new set of stimuli that were all discrete objects with high-quality, close-up images in the COCO image database. In this follow-up experiment, we found that the contextual-prototype RDM was once again significantly correlated with the behavioral RDM, but, importantly, the performance of the contextual-prototype model was equivalent to the object-based models rather than being strongly superior, suggesting that contextual prototypes match object-based models but do not generally outperform them (Fig. 3A and Supplementary Fig. S3A). When performing the analyses layerwise, the results were similar to those in the ADE20K data, with consistent effects for the contextual prototype model across layers and a slight improvement at higher layers for the object models (Supplementary Fig. S4A).

We next performed variance partitioning analyses to quantify the shared and unique explained variance of contextual and object-related information. In both the ADE20K and COCO data, we found that while the contextual and object-based models explained some degree of shared variance, they also explained substantial amounts of disjoint variance (Fig. 4A and see Supplementary Table 2 for explained variance values for all partitions). This suggests that contextual and object-related information account for partially distinct aspects of intuitive similarity judgments.

In sum, across two experiments using different sets of objects sampled from different image databases, we found that contextual prototypes performed as well as or better than models of the objects themselves. Furthermore, in follow-up analyses, we observed similar results using a different CNN that was pre-trained on the Places dataset rather than ImageNet (Supplementary Fig. S5) (Zhou et al., 2014), demonstrating that these findings generalize beyond the specific CNN presented here. Together, these results demonstrate a striking phenomenon of object representation: intuitions about object similarity can be explained equally well by models that directly represent the objects themselves or by models that ignore the objects and simply represent the scene contexts in which they are typically encountered.

3.2. Contextual prototypes are weak predictors of shape similarity judgments

Although contextual prototypes are highly informative for explaining spontaneous judgments of similarity, we predicted that they would show a clear dissociation from object-based models when explaining judgments about the visual properties of objects, such as their shapes, which should be better captured by an object model. To test this hypothesis, we performed a new set of experiments in which participants were asked to perform a multi-arrangement task on the same sets of stimuli, but with the key difference that participants were now specifically instructed to judge the similarity of the objects based on their shapes (Fig. 1). We computed RSA correlations between our model RDMs and the behavioral RDM from the shape task. Across two experiments using the ADE20K and COCO stimulus sets, we found that object-based CNN models strongly outperformed the contextual-prototype model when explaining shape similarity judgments (Fig. 3B and Supplementary Fig. S3B). Furthermore, the performance of the contextual-prototype model on the shape task was dramatically lower than its performance on the intuitive task. When performing the analyses layerwise, the effects for the contextual prototype model were consistently low across all layers, whereas the object models were generally better at higher layers (Supplementary Fig. S4B). Moreover, variance partitioning analyses showed that nearly all the explained variance of the contextual prototypes on the shape task was shared with the object-based models (Fig. 4 and Supplementary Table 2). Thus, not all dimensions of object similarity are well explained by context. For the case of shape similarity, there is a striking dissociation between context- and object-based models.

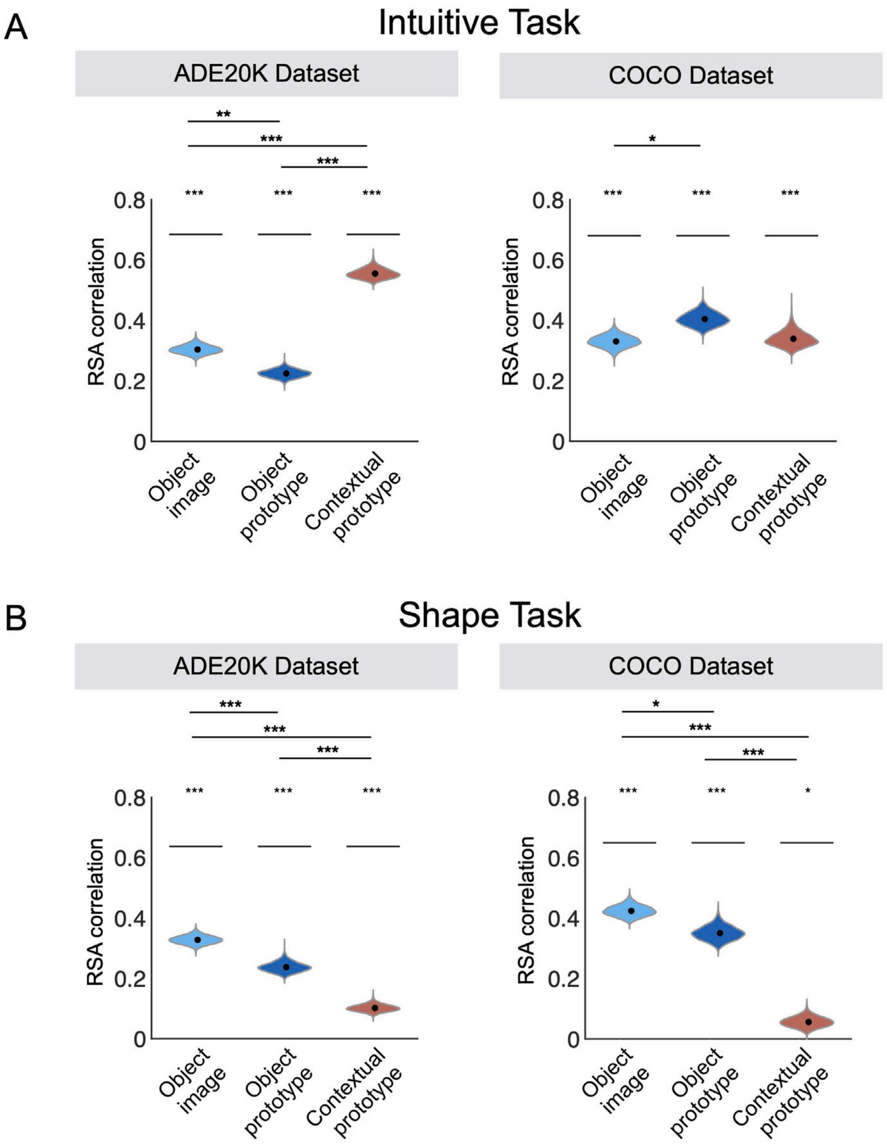

Fig. 3. RSA correlations of representational models and behavioral similarity. The top panel shows RSA correlations of model RDMs and the behavioral RDM for the intuitive similarity task. The bottom panel shows RSA correlations for the shape similarity task. Results are shown for both the ADE20K and COCO experiments. (A) RSA results show that the contextual-prototype model was similar to or better than object-based models when explaining intuitive similarity judgments in both the ADE20K and COCO experiments (top panel). (B) In contrast, the object-based models exhibited a strong advantage over the contextual-prototype model when explaining similarity judgments based on object shape. The violin plots show the mean RSA correlations (central black dots) and bootstrap resampling distributions. The gray lines above each violin plot indicate the noise-ceiling estimates based on Spearman–Brown-corrected split-half correlations of the behavioral RDMs. *p< 0.05, **p< 0.01, ***p< 0.001.

3.3. *Contextual prototypes rival models of distributional semantics*

We next sought to determine how our contextual prototypes compare with models of distributional semantics based on the co-occurrence statistics of objects in images and object names in language, which have previously been examined as models of contextual representation in high-level visual cortex (Bonner & Epstein, 2021). To do this, we examined a model based on object co-occurrences in images, known as object2vec (Bonner & Epstein, 2021) and another model based on word co-occurrences in written language, known as word2vec (Mikolov, Sutskever, et al., 2013). For the intuitive task in both the ADE20K and COCO experiments, we found that object2vec and word2vec exhibited strongly significant correlations with the behavioral RDMs and were either lower than or similar to the performance of the contextual-prototype model (Fig. 5; see Supplementary Fig. S6 for performance on the shape task and Supplementary Fig. S7 for participant-level results). Interestingly, the contextual-prototype model was slightly better than object2vec in both experiments (though this difference was not significant in the participant-level analyses). This suggests that even though object2vec is based on the co-occurrence of object labels in annotated images, there may be additional contextual information to be learned from modeling the entire scene context in which an object is encountered. For example, some contextual information, like the spatial geometry of a scene, may not be represented in a co-occurrence model that treats images as bags of objects.

4. Discussion

We developed a computational approach to learn contextual prototypes that capture the scene associations of objects. Using this approach, we found that the natural statistics of object context are predictive of how humans intuitively judge the similarities between real-world objects. Remarkably, as a computational model of object similarity, contextual prototypes rivaled the performance of CNN representations of object images and word embeddings of object names.

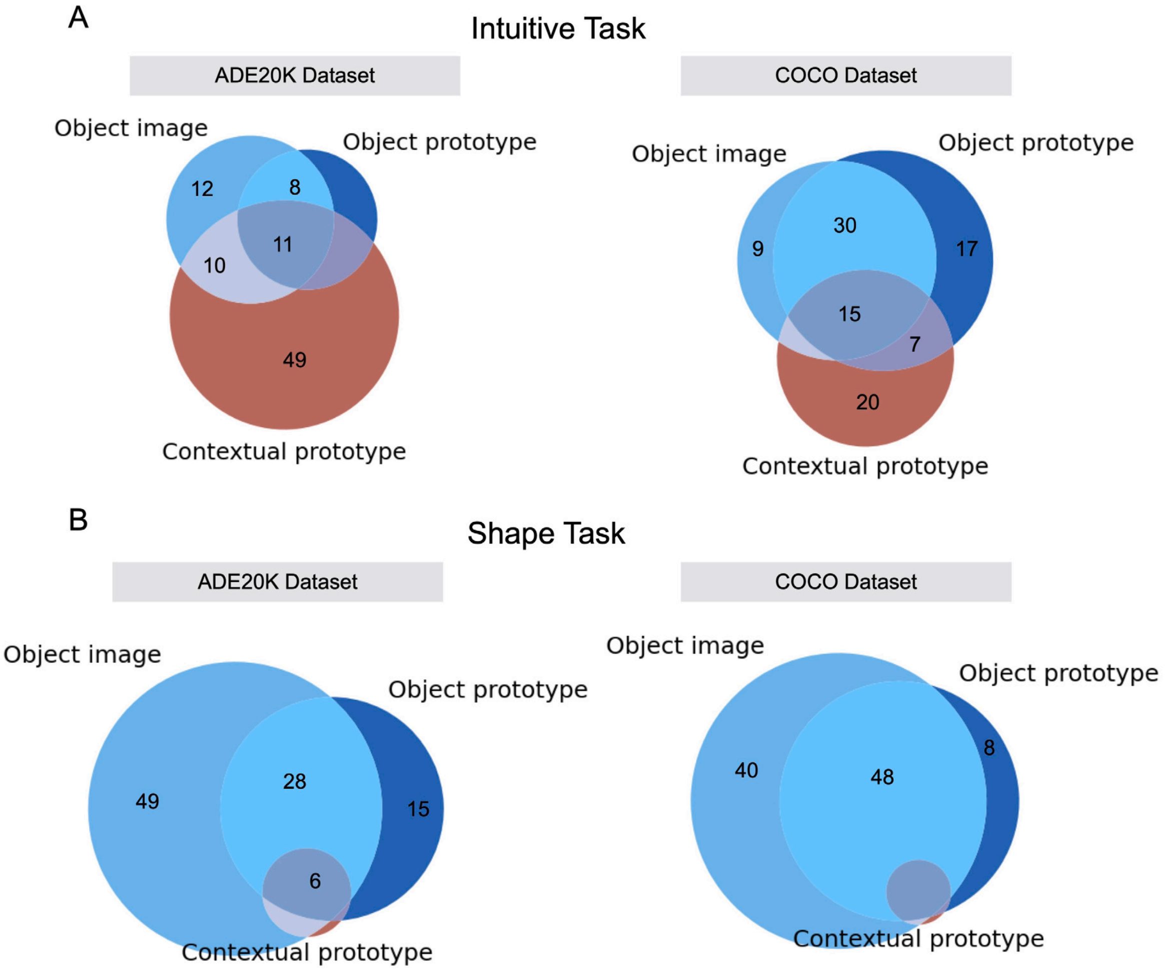

Fig. 4. Variance partitioning analysis of representational models. Variance partitioning was used to quantify the shared and unique explained variance for the object image, contextual prototype, and object-prototype models. Results are shown for both the ADE20K and COCO experiments. (A) Partitioning the explained variance in the intuitive task showed that the contextual and object-based models explained some portion of shared variance but additionally accounted for a substantial amount of disjoint variance. (B) In contrast, partitioning the explained variance in the shape task showed that nearly all explained variance for the contextual prototypes was shared with the object-based models and, furthermore, that the object image model accounted for the most unique variance. Note that variance partition values below 5% are not annotated in this figure but can be seen in Supplementary Table 2. Total r^2 values: ADE20K Intuitive = 0.434, COCO Intuitive = 0.365, ADE20K Shape = 0.213, and COCO Shape = 0.295.

Together, this work develops a novel computational approach to rigorously characterize the statistical associations between objects and their natural scene contexts, and it shows that these contextual associations are tightly linked to the core dimensions of object similarity in human behavior.

4.1. Contextual associations and similarity judgments

Similarity judgments have long been considered a window into how objects are represented in the mind. They are used in cognitive science to characterize the representational geometry of objects and other naturalistic stimuli (Greene et al., 2016; Hebart et al., 2019, 2020; Jozwik et al., 2017; Mur et al., 2013; Peterson et al., 2018; Tarhan et al., 2021), and they are used in neuroscience to identify brain regions that support the high-level processing of these stimuli (Dima et al., 2022; Groen et al., 2018; King et al., 2019; Mur et al., 2013). While similarity judgments can provide insights into the underlying dimensions of object representation, our work suggests that the most important dimensions may emerge not from the objects themselves but rather from their latent statistical associations. Moreover, our findings suggest that in neuroscience research, intuitive similarity judgments of objects may be better suited for identifying brain regions that process scene contexts rather than the visual forms of objects (Bar & Aminoff, 2003; Bonner & Epstein, 2021; Groen et al., 2018). Consistent with this, evidence from fMRI has shown that similarity judgments for objects are correlated most strongly with scene-selective regions, such as the parahippocampal place area (King et al., 2019), which have previously been linked to representations of object co-occurrence and contextual associations (Bar & Aminoff, 2003; Bonner & Epstein, 2021) rather than representations of object form.

Other recent work has found that CNNs trained on object classification implicitly learn the associations between objects in scenes and are highly predictive of human judgments of contextual similarity (Aminoff et al., 2022; Bracci et al., 2022). These findings corroborate the motivation for our study, which is that CNNs are well-suited for modeling the rich contextual information associated with objects and, furthermore, that the contextual information in CNNs is relevant to human behavior. However, our study addresses two important issues that have not been explored in previous work. First, we developed a new computational method for modeling the scene associations of objects without modeling the features of the objects themselves. The development of such methods is crucial given that there are currently few computational tools for characterizing the natural statistics of object context (Bonner & Epstein, 2021). Second, our work goes beyond the study of explicit ratings of contextual association and shows that contextual information strongly predicts similarity judgments even when participants are not explicitly cued to think about context. This suggests that commonly used unguided similarity tasks may reveal more about

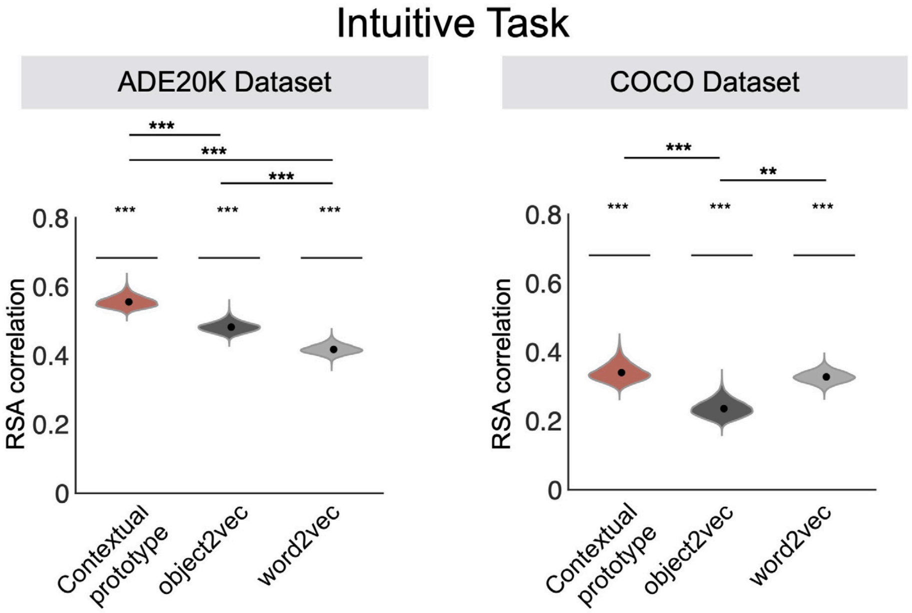

Fig. 5. RSA correlations for contextual prototypes and models of distributional semantics in vision and language. Two models of distributional semantics were examined: object2vec and word2vec. The object2vec model contains information about the co-occurrence statistics of objects in a database of densely annotated scenes (Bonner & Epstein, 2021). The word2vec model contains information about word co-occurrence statistics in a large corpus of written language. Results are shown for both the ADE20K and COCO experiments. The violin plots show the mean RSA correlations (central black dots) and bootstrap resampling distributions. The gray lines above each violin plot indicate the noise-ceiling estimates based on Spearman–Brown-corrected split-half correlations of the behavioral RDMs. *$p< 0.05$, **$p< 0.01$, ***$p< 0.001$.

the co-occurrence of objects in scenes than about the appearance of the objects themselves, such as the appearance of their shapes (Jozwik et al., 2017; Mur et al., 2013; Peterson et al., 2018).

Our findings have connections to an important theoretical distinction in the semantic memory literature between taxonomic and thematic similarity (Mirman et al., 2017). Objects that are taxonomically similar come from similar categories (such as a dog and a cat, or an apple and a pear), whereas objects that are thematically similar come from similar scenes and events (such as a dog and a leash, or an apple and a peeler). Evidence from semantic memory studies has shown that taxonomic and thematic knowledge reflect different modes of reasoning about the relationships among objects and rely on different neural substrates (Schwartz et al., 2011). Although this distinction between taxonomic and thematic knowledge has long been appreciated in the semantic memory literature, there has been far less consideration of this distinction in the literature on visual object processing. Our contextual-prototype model naturally captures the thematic associations of visual objects using images of their co-occurring scene contexts, and it may be useful in future work for determining whether the visual system contains different processing pathways for encoding the thematic and taxonomic properties of objects. Furthermore, our findings suggest that when observers evaluate the intuitive similarities among visual objects, they may default to thematic reasoning.

4.2. Computational modeling of similarity judgments

Our results highlight the need for better computational approaches to characterize the latent associations of objects. Previous efforts to model the representations underlying object similarity have largely focused on CNN representations of object images, word embeddings for object names, or human-annotated feature lists (Hebart et al., 2019; Jozwik et al., 2017; Peterson et al., 2018). Recent work has developed a data-driven approach to induce object embeddings from behavioral similarity data and has shown that interpretable dimensions corresponding to perceptual and conceptual object properties emerge (Hebart et al., 2020). Nonetheless, even with such data-driven embeddings, one still needs representational theories to make sense of them. If one seeks to test a representational theory of contextual or thematic associations, there are few existing models for doing so (Bonner & Epstein, 2021). This dearth of scene-context models is particularly striking in contrast to the many models that exist for testing representational theories of object form based on CNNs or theories of semantics based on word embeddings. Our procedure for building contextual prototypes based on CNN representations of scene contexts demonstrates a powerful new approach for modeling the latent associations of objects in natural images. This contextual-prototype model could be useful for future studies seeking to characterize how object processing is influenced by the natural statistics of context.

4.3. Varied approaches for assessing similarity

Our work and other recent studies of unguided similarity judgments raise the question of how participants rank or weight the many possible dimensions that could be used to evaluate similarity. In unguided tasks, participants are free to evaluate objects along any dimensions that come to mind. For example, a participant could evaluate objects along dimensions related to contexts, shapes, textures, colors, affordances, material properties, and so on. Our findings and previous work demonstrate that despite the many possible dimensions that one could consider, the primary dimensions that come to mind in unguided tasks are generally consistent across individuals (Hebart et al., 2019; Mur et al., 2013; Peterson et al., 2018). Thus, there appears to be a default ranking or weighting of the representational dimensions that underlie human intuitions about object similarity. While our work suggests that contextual dimensions may be high in this ranking, we know that there are myriad other types of object similarity that participants could alternatively consider. A goal for future work is to catalogue these many possible dimensions of object similarity and to characterize their relative importance for object-related behaviors.

One approach to cataloging the dimensions of object similarity is to consider a set of experimenter-defined dimensions and to instruct participants to specifically evaluate these dimensions. For example, in our own data, we found that when participants were instructed to judge similarity based on object shape, the resulting similarity space was consistent across participants and differed from that obtained in the unguided task. However, one does not necessarily have to use experimenter-defined dimensions. An alternative approach is to ask participants to evaluate similarity for small groups of objects, as in the triplet odd-one-out task in Hebart et al. (2019). In such a design, there will be trials in which contextual similarity is either irrelevant (e.g., because all three objects are from the same context) or less salient (e.g., a trial in which two of the three objects are bright red). Another possibility is to perform a version of the multi-arrangement task in which participants are asked to group objects in as many distinct ways as possible. For example, a participant might first group the objects by contextual similarity, and then on the second iteration, the objects might be grouped by shape similarity, and then color similarity, and texture similarity, and so on. Using such a design, it may be possible to discover a rich variety of similarity dimensions and to determine their relative rankings.

Our findings suggest that an important direction for future work is to better understand how similarity judgments might be affected by the design parameters of behavioral tasks. In our work, we found that when participants evaluated similarity in a multi-arrangement task—which is one of the most commonly used methods in the field—their responses resembled contextual similarities. However, an interesting question is whether the multi-arrangement task might implicitly draw attention to contextual information. There are at least two properties of the multi-arrangement task that could potentially call attention to context. The first property is the spatial nature of the task, in which participants arrange objects on a 2D spatial display so that nearby objects are more similar. Perhaps when participants are asked to arrange objects in space, they naturally think about the spatial contexts in which the objects typically occur. The second property is the large number of stimuli on the display. When participants view an array of many objects, they may be inclined to think about how these objects would typically be grouped if they were encountered under natural conditions. Thus, it is an open question whether contextual information would drive similarity judgments as strongly if the task were not spatial in nature and if the number of items on the display were decreased. A task that fits this description is the odd-one-out task Hebart et al. (2019), and it would, thus, be informative in future work to directly compare unguided similarity judgments on the same objects when using the multi-arrangement and odd-one-out tasks.

4.4. Future directions and limitations

A goal for future work is to elucidate the specific contextual regularities that underlie the performance of our contextual-prototype model. We know that contextual prototypes capture the latent statistical regularities of the natural scene contexts in which objects are encountered. What we do not yet know is what specific aspects of scene context are most important for these contextual prototypes. For example, are contextual prototypes best explained in terms of semantic scene categories, object co-occurrence statistics (Bonner & Epstein, 2021), functions and affordances (Bonner & Epstein, 2017; Greene et al., 2016), spatial layout (Brady et al., 2017; Oliva & Torralba, 2007) or other possible factors? Future work could seek to quantify these properties in natural images and to determine which, if any, best explains behavioral similarity judgments.

Another exciting direction for future work is to understand the level of specificity of contextual effects in object processing. For example, when we view objects, does the brain represent contextual information at the level of individual items (e.g., your own car), basic-level object categories (e.g., all cars), or superordinate categories (e.g., all vehicles)? Furthermore, what level of contextual specificity best explains behavior on the unguided similarity judgments investigated here? To systematically address this question, we would need a dataset in which the stimuli were selected so that one could model contextual information at different levels of specificity. While our data has some examples of sets of stimuli that can be naturally grouped into superordinate categories, such as cars and trucks, this is not generally true for most of our stimuli (e.g., clock, handbag, scissors). Thus, it remains an open question whether we could obtain similar results if our contextual prototype representations were modeled at a more superordinate category level or, alternatively, whether the contextual prototypes might perform even better if they were modeled at a more subordinate level.

Our findings also highlight an important methodological consideration when investigating the statistical regularities of objects in natural images. Namely, there is a tradeoff between image databases that are best suited for characterizing contextual regularities versus those that are best suited for characterizing object regularities. This is because scene-centric databases contain rich contextual information, but the objects in scene images are often zoomed-out, making it difficult to capture their detailed properties. In contrast, object-centric databases contain high-quality, close-up images of objects but lack rich surrounding contexts. We saw such a tradeoff in our own analyses. Specifically, we found that the relative performance of the object-prototype model was substantially better when using images from the more object-centric COCO database rather than the more scene-centric ADE20K database. We suggest that attaining a complete picture of the statistical regularities of objects in natural images will require approaches that combine the strengths of both object-centric and scene-centric datasets.

There are important limitations to our findings to keep in mind. First, our findings do not account for all aspects of object similarity. Indeed, similarity is a task-dependent construct, and there are many possible dimensions along which object similarity could be evaluated. Our work focuses on the primary dimensions of object similarity that come to mind in an unguided task. Second, the multi-arrangement task used here would be challenging to scale up to more than a few hundred stimuli. In the multi-arrangement task, all stimuli are displayed on the screen simultaneously, which allows participants to get a global view of the stimulus space and to consider its most important organizing dimensions. However, the downside of this task is that it becomes overwhelming and impractical beyond a few hundred stimuli. Other similarity tasks, like the triplet-odd-one-out task, also face challenges with data scaling. When considering datasets on the order of thousands of stimuli or more, a promising alternative is to build computational models that infer the latent representational dimensions of object similarity from limited behavioral data (Hebart et al., 2020).

4.5. Conclusion

In sum, we developed a computational approach to characterize the high-level scene associations of objects in natural images, and we found that these scene associations predicted how participants spontaneously judged the similarities between objects. Surprisingly, this means that when examining computational models of object similarity in unguided tasks (Jozwik et al., 2017; Mur et al., 2013; Peterson et al., 2018), the models do not need to see the objects themselves—all they need are the contexts in which the objects are typically encountered. These findings support the theory that object representations are fundamentally intertwined with representations of their associated contexts (Aminoff et al., 2022; Aminoff & Tarr, 2015; Bar & Aminoff, 2003; Bonner & Epstein, 2021; Davenport & Potter, 2004; Oliva & Torralba, 2007). Moreover, our novel method for modeling the contextual associations of objects opens new avenues of research on the statistical underpinnings of object representations in brains and machines.

CRediT authorship contribution statement

Caterina Magri: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Eric Elmoznino**: Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Michael F. Bonner**: Conceptualization, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105535.

References

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to voxels: Modeling visual representation in the human brain. arXiv:1407.5104 [Cs, Q-Bio], arXiv:1407.5104.

Aminoff, E. M., Baror, S., Roginek, E. W., & Leeds, D. D. (2022). Contextual associations represented both in neural networks and human behavior. *Scientific Reports*, *12*(1), 5570.

Aminoff, E. M., & Tarr, M. J. (2015). Associative processing is inherent in scene perception. *PLoS One*, *10*(6), 1–19.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*(2), 347–358.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 238–247). Baltimore, Maryland: Association for Computational Linguistics.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177.

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology*, *14*(4), Article e1006111.

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, *12*(1), 4081.

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *The Journal of Neuroscience*, *36*(2), 432–444.

Bracci, S., Mraz, J., Zeman, A., Leys, G., & de Beeck, H. O. (2022). The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. http://dx.doi.org/10.1101/2021.08.13.456197, BioRxiv, Cold Spring Harbor Laboratory.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160–1176.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.

Dima, D. C., Tomita, T. M., Honey, C. J., & Isik, L. (2022). Social-affective features drive human representations of observed actions. In C. I. Baker, A. Lingnau, & M. Lescroart (Eds.), *ELife*, *11*, Article e75027.

Goldstone, R. L., & Son, J. Y. (2012). *Similarity*. Oxford University Press.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82–94.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *ELife*, *7*, Article e32962.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1854 object concepts and more than 26,000 naturalistic object images. *PLoS One*, *14*(10), 1–24.

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*.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. In J. Diedrichsen (Ed.), *PLoS Computational Biology*, *10*(11), Article e1003915.

King, M. L., Groen, I. I., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, *197*, 368–382.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *The Journal of Neuroscience*, *20*(9), 3310–3318.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. In M. Bethge (Ed.), *PLoS Computational Biology*, *12*(4), Article e1004896.

Lauer, T., Schmidt, F., & Võ, M. L.-H. (2021). The role of contextual materials in object recognition. *Scientific Reports*, *11*(1), 1–12.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft COCO: Common objects in context. arXiv:1405.0312 [Cs], arXiv:1405.0312.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781 [Cs], arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological Bulletin*, *143*(5), 499–520.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*.

Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods*, *40*, 457–466.

Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2017). Erratum to: An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example (Behav Res,(2008) 40, 2,(457-466), 10.3758/BRM. 40.2. 457). *Behavior Research Methods*, *49*(6), 2275.

Nimon, K. F., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, *16*(4), 650–674.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling representations of object shape and object category in human visual cortex: The animate–inanimate distinction. *Journal of Cognitive Neuroscience*, *28*(5), 680–692.

Roads, B. D., & Love, B. C. (2020). Enriching ImageNet with human similarity judgments and psychological embeddings.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., Mirman, D., & Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences*, *108*(20), 8520–8524.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123, Publisher: American Psychological Association.

Tarhan, L., De Freitas, J., & Konkle, T. (2021). Behavioral and neural representations en route to intuitive action understanding. BioRxiv, Cold Spring Harbor Laboratory.

Zeman, A. A., Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific Reports*, *10*(1), 2453.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems, vol. 27*. Curran Associates, Inc..

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, *127*(3), 302–321.