Object representations in the human brain reflect the co-occurrence statistics of vision and language

Michael F. Bonner^{1, 2} and Russell A. Epstein²

¹Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA

²Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

Figures: 8
Supplemental figures: 4
Conflicts of interest: nothing to declare
Keywords: vision, objects, context, co-occurrence, scene-selective cortex, object-selective cortex, parahippocampal place area, fMRI, encoding model

Please address correspondence to:

Michael F. Bonner
Department of Cognitive Science
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218
Email: mfbonner@jhu.edu

ABSTRACT

A central regularity of visual perception is the co-occurrence of objects in the natural environment. Here we use machine learning and fMRI to test the hypothesis that object co-occurrence statistics are encoded in the human visual system and elicited by the perception of individual objects. We identified low-dimensional representations that capture the latent statistical structure of object co-occurrence in real-world scenes, and we mapped these statistical representations onto voxelwise fMRI responses during object viewing. We found that cortical responses to single objects were predicted by the statistical ensembles in which they typically occur, and that this link between objects and their visual contexts was made most strongly in the anterior portion of the scene-selective parahippocampal place area. In contrast, a language-based statistical model of the co-occurrence of object names in written text predicted responses in neighboring regions of object-selective visual cortex. Together, these findings show that the sensory coding of objects in the human brain reflects the latent statistics of object context in visual and linguistic experience.

INTRODUCTION

Many objects have a natural place in the world—a setting in which they and other co-occurring objects are typically encountered. For example, fire hydrants are often found with traffic lights and mailboxes on city sidewalks, while tea kettles are found with mugs and stoves in kitchens. This type of *contextual knowledge* can help people identify their surroundings and generate expectations for the other objects they might encounter. Previous behavioral work has demonstrated that the visual system uses contextual knowledge to facilitate object recognition^{1–3} and visual search⁴. Moreover, theoretical accounts suggest that contextual facilitation might be evidence for a general cognitive mechanism by which the mind makes predictions about the world in order to support adaptive behavior^{5–10}. Thus, identifying the brain mechanisms that support the representation of contextual knowledge is an important challenge for cognitive neuroscience.

Previous neuroimaging studies have attempted to meet this challenge^{11–14}. One common approach has been to operationalize context as a one-dimensional rating that reflected how strongly an object brought to mind a particular context^{11,15} or whether or not an object was associated with other objects or locations¹⁶. Objects with strong contextual associations by these definitions were found to elicit greater responses than objects with weak contextual associations in the parahippocampal place area (PPA) and retrosplenial complex (RSC), brain regions that are known to respond strongly during the visual perception of scenes^{17,18}. Given that scenes are contexts by definition, typically containing several co-occurring objects, these results are consistent with the view that PPA and RSC have a general role in both the perception and retrieval of contextual information¹⁹. However, because these conclusions were based on a univariate analysis of regional responses along a single stimulus dimension, they were vulnerable to alternative explanations. Indeed, subjective ratings of other high-level object properties, such as landmark suitability²⁰ and spatial definition²¹, have been found to explain similar (or greater) variance in PPA and RSC responses, and some of the effects of object context in scene-selective visual regions have not been consistently observed²². More importantly, this approach is limited insofar as it only examines whether an object has strong contextual associations or not; it does not attempt to analyze the nature of these associations (e.g., *which* objects are associated with each other). These limitations have led to uncertainty about whether scene regions play a central role in mediating contextual knowledge.

Here we attempt to resolve this issue by using a new approach that more directly tests for contextual representations. Specifically, we use fMRI to search for neural representations that reflect the multivariate statistical structure of object co-occurrence in the visual environment. To model this statistical structure, we first developed a novel adaptation of the word2vec machine-

learning algorithm from computational linguistics^{23,24}, which we label object2vec. Word2vec uses unsupervised learning to compress high-dimensional data about the co-occurrence of words within a large corpus of text into a relatively small number of highly informative dimensions. Object2vec performs the exact same operation for objects within a corpus of over 20,000 real-world scenes. Thus, object2vec creates a low-dimensional representational space that reflects the statistical regularities of object co-occurrence in images. We then scanned human observers while they viewed isolated single objects, and we used voxelwise encoding models to identify brain regions where the dimensions of object2vec predicted fMRI responses. For comparison, we performed the same fMRI analysis on the dimensions of language-based word2vec.

To anticipate, our results show that an anterior portion of the scene-selective PPA represents the statistical associations between objects and their visual contexts, and that these representations are elicited even when subjects view individual objects without any contextual information in the stimulus. In contrast, language-based word2vec explained variance in neighboring object-selective regions (including posterior fusiform; pFs), indicating that these regions represent object properties that are related to language-based co-occurrence statistics. Together these findings reveal the relationship between high-level object representations in ventral visual cortex and the latent manifolds of object context in vision and language. Our results also provide insight into the statistical properties of cortical representations—specifically, they suggest that high-level sensory regions utilize an efficient coding scheme that compresses a large number of behaviorally relevant stimulus properties into a relatively small number of statistically informative dimensions.

RESULTS

Object embeddings

We first sought to characterize the co-occurrence statistics of objects in the visual environment. To accomplish this, we needed a large dataset of natural images in which all object occurrences were labeled. We took advantage of the recently created ADE20K database, which contains 22,210 annotated scenes in which every object has been manually labeled by an expert human annotator²⁵. One approach for characterizing the co-occurrence statistics of this dataset would be to simply construct a matrix of co-occurrence frequencies for all pairwise comparisons of objects. However, this would produce a sparse and high-dimensional matrix, with many redundant dimensions. In the field of computational linguistics, there is a long history of modeling word co-occurrence data in language corpora with dense, lower dimensional representations²⁶. This modeling framework, known as distributional semantics, has proved highly useful because

Figure 1. Statistical model of object context. **A)** Object context was modeled using the ADE20K data set²⁵, which contains 22,210 images in which every pixel is associated with an object label provided by an expert human annotator. An adaptation of the word2vec machine-learning algorithm for distributional semantics—which we call *object2vec*—was applied to this corpus of image annotations to model the statistical regularities of object-label co-occurrence in a large sample of real-world scenes. Using this approach, we generated a set of 8-dimensional *object2vec* representations for the object labels in the ADE20K data set. These representations capture the co-occurrence statistics of objects in the visual environment, such that objects that occur in similar scene contexts will have similar *object2vec* representations. **B)** The panel on the left shows a two-dimensional visualization of the *object2vec* representations for the 81 object categories in the fMRI experiment. This visualization shows that the *object2vec* representations contain information about the scene contexts in which objects are typically encountered (e.g., indoor vs. outdoor, urban vs. natural, kitchen vs. office). The panel on the right shows a similar plot for the language-based *word2vec* representations of the same object categories. These plots were created using *t*-distributed stochastic neighbor embedding (*tSNE*). The colors reflect cluster assignments from *k*-means clustering and are included for illustration purposes only. *K*-means clustering was performed on the *tSNE* embeddings for *object2vec*, and the same cluster assignments were applied to the *word2vec* embeddings in the right panel for comparison.

the resulting representations capture information about the semantic properties of words, such that words with similar meanings tend to have similar representations. A leading technique for modeling distributional semantics is word2vec, which is a class of neural network models that are trained to predict target words from their surrounding linguistic contexts (or vice versa)^{23,24}. The learned internal weights of word2vec models (known as word embeddings) reflect the latent statistical structure of the linguistic contexts in which words are typically found. Here we adopt the general approach of the word2vec algorithm to model the statistical structure of object context in images. We named this model object2vec.

In the object2vec model, object labels are treated like word tokens in a corpus and their contexts are defined by the images they occur in. Using this approach, we trained an object2vec model on the image annotations from ADE20K and learned a set of 8-dimensional embeddings that represent statistical information about the natural contexts of objects (Fig. 1A). Through a parameter search and principal component analysis (PCA), we found that higher-dimensional embeddings did not substantially alter the representational geometry of object2vec, suggesting that 8 dimensions sufficed for learning representations of object context from this data set (see Methods for details). A two-dimensional visualization of the object2vec embeddings (for the objects in our experiment) shows that these representations contain meaningful information about the way in which objects are grouped together in the natural environment, with object clusters that can be intuitively interpreted as familiar scene categories (e.g., indoor, outdoor, kitchen, office).

We also examined a language-based word2vec model. This model was trained on the co-occurrence of words in a natural language corpus containing ~100 billion words (see Methods for details). A two-dimensional visualization of the word2vec representations suggests that they reflect a broader set of semantic associations than object2vec (Fig. 1B). These include information about the contexts that objects are found in, but also information about what kinds of taxonomic categories objects belong to, such as modes of transportation (e.g., airplane, car, bus, bicycle) and furniture for sitting on (e.g., swivel chair, armchair, sofa, stool).

Cortical representations and object co-occurrence in vision and language

With our models of vision- and language-based context in hand, we set out to test the hypothesis that object representations in visual cortex reflect the statistical regularities captured by these models. We were specifically interested in determining whether such contextual representations are elicited automatically when an object is viewed. Thus, it was important that our object stimuli be shown in isolation, without other co-occurring objects present in the image. We designed a stimulus set of 810 objects from 81 different categories (10 object tokens per category). Example

Figure 2. Experimental stimuli. **A)** In an fMRI experiment, subjects viewed isolated images of objects from 81 different categories. This panel shows one example image from each category. **B)** There were 10 unique tokens for each category (for a total of 810 unique stimuli). This panel shows all 10 stimuli for two categories. **C)** In the fMRI scanner, images were presented on complex, textured backgrounds to reduce the saliency of low-level features related to object shape and size. This panel shows four example stimuli on textured backgrounds.

stimuli are shown in Fig. 2. We measured fMRI responses while subjects viewed these objects, shown alone on meaningless textured backgrounds, and performed a simple perceptual task of responding by button press whenever they saw a “warped” object. Warped objects were created through diffeomorphic warping of object stimuli (Fig. S1; see Methods for details).

Figure 3. Encoding model of object context. Voxelwise encoding models were used to assess whether our statistical model of object context (object2vec) could reliably explain variance in the fMRI responses to the experimental object categories. Linear regression was used to map the representations of object2vec onto fMRI responses. We assessed the out-of-sample prediction accuracy of these regression models through a 9-fold cross-validation procedure. Each fold of the cross-validation design contains a set of object categories that do not appear in any other fold. These folds were shown in separate fMRI runs. Parameters for the voxelwise linear regression models were estimated using the fMRI data for 8 folds of the object categories and the learned regression weights were then applied to the held-out object categories in the remaining fold to generate a set of predicted fMRI responses. This procedure was repeated for each fold of the cross-validation design, and the predicted fMRI responses from each fold were concatenated together. Prediction accuracy was assessed by calculating the voxelwise correlations of the predicted and actual fMRI responses across all object categories.

Using a voxelwise modeling procedure, we examined whether fMRI responses to these object stimuli could be predicted from the co-occurrence statistics of objects in vision and language. We

first estimated the fMRI responses to each object category. We then fit voxelwise encoding models with the goal of predicting the fMRI responses to all object categories through a weighted linear sum of the object2vec or word2vec embeddings. Specifically, a set of linear regression models were estimated using the statistical embeddings as regressors and the voxelwise fMRI responses as predictands (Fig. 3). Through cross-validation, we assessed how well the estimated encoding models could predict fMRI responses to out-of-sample stimuli (see Methods for details). The cross-validation procedure was designed so that the training and test sets always contained objects from different categories, which allowed for a strong test of generalization to new semantic categories, rather than new stimuli from the same categories.

We first performed a region of interest (ROI) analysis of the PPA, which has been proposed to have a central role in the processing of object context^{19,27}. We looked separately at the anterior and posterior PPA, because previous work suggests that non-spatial contextual associations elicit responses in anterior parahippocampal cortex more than posterior parahippocampal cortex¹⁶ and because several lines of evidence suggest functional and connectivity differences across the anterior-posterior extent of the PPA^{28,29}. Figure 4 shows results for both the image-based object2vec model and language-based word2vec. As can be seen in the ROI plot, both the split-half reliability of our data and the prediction accuracy of our encoding models is quite high for an fMRI experiment. We believe this is likely attributable to the fact that we used a mini-block design, with a large number of repeated presentations for each stimulus (although another factor may be our use of a reliability mask, which would bias the reliability scores upward; see Methods). In terms of statistical significance, both object2vec and word2vec had significant prediction accuracy in both anterior and posterior segments of the PPA. Thus, at least some portion of the object-evoked responses throughout the PPA could be reliably predicted from representations of the visual and linguistic contexts of objects. However, in terms of the magnitude of the effects, the highest prediction accuracy was observed for the image-based statistics of the object2vec model in anterior PPA. In an interaction test, we found that the difference in accuracy for object2vec minus accuracy for word2vec was greater in anterior PPA compared to posterior PPA (permutation test, $p = 0.002$).

Figure 4. Voxelwise encoding models of object context in PPA. This plot shows the average prediction accuracies for encoding models in voxels from the anterior third and posterior third of the PPA using either image-based object2vec representations as regressors or language-based word2vec representations. The violin plots show the mean prediction accuracies (central black dots) and bootstrap standard deviations. The gray lines above each violin plot indicate the average voxelwise split-half reliability of the fMRI responses in each ROI. The image on the left shows the PPA parcel on a ventral view of a cortical surface rendering. PPA = parahippocampal place area. ** $p < 0.01$, *** $p < 0.001$.

We next performed ROI analyses for other functionally defined regions, including the scene-selective RSC, which has also been implicated in object context along with the PPA^{19,27}. We also analyzed scene-selective occipital place area (OPA), PPA (without segmenting into anterior and posterior), object-selective pFs³⁰, object-selective lateral occipital (LO)³¹, and early visual cortex (EVC). Results for these ROIs are shown in Figure 5. The image-based object2vec model generated significant prediction accuracies in all scene-selective ROIs and in one object-selective ROI (pFs). The language-based word2vec model generated significant prediction accuracies in all ROIs, with larger effects in the object-selective regions. These ROI analyses indicate that object context is not only represented in the PPA but also in other scene-selective and object-selective regions of high-level visual cortex. That said, it is notable that the effects for object2vec were numerically strongest in the PPA, supporting the idea that this region is particularly important for linking objects to their visual contexts.

Figure 5. Voxelwise encoding models of object context in functionally defined ROIs. This plot shows the average prediction accuracies for encoding models in voxels from multiple regions of interest using either image-based object2vec representations as regressors or language-based word2vec representations. Object-selective ROIs are plotted in orange and scene-selective ROIs are plotted in cyan. The violin plots show the mean prediction accuracies (central black dots) and bootstrap standard deviations. The gray lines above each violin plot indicate the average voxelwise split-half reliability of the fMRI responses in each ROI. EVC = early visual cortex, LO = lateral occipital, PFs = posterior fusiform, OPA = occipital place area, PPA = parahippocampal place area, RSC = retrosplenial complex. ** $p < 0.01$, *** $p < 0.001$.

Finally, we performed whole-brain analyses to examine the prediction accuracy of the object2vec and word2vec models across all voxels with reliable signal (see Methods for details). The results of these analyses are shown in Figure 6. The findings were generally consistent with the results of the ROI analyses, with a cluster of relatively high prediction accuracy for the object2vec model overlapping with anterior PPA. This cluster was located in the right hemisphere in a region of ventral visual cortex where lateral PPA overlaps with a medial part of the object-selective pFs. There were also several other smaller patches of significant prediction accuracy for object2vec, including the scene-selective OPA and RSC. For the word2vec model, a prominent cluster of significant prediction accuracy was observed in a region of ventral visual cortex, immediately adjacent to the largest cluster for the object2vec model in the right hemisphere. This cluster of significant prediction accuracy for word2vec appears to overlap with both pFs and portions of the PPA. Another large cluster of high prediction accuracy for word2vec was observed on the lateral surface near object-selective LO (see also ROI results in Figure 5). There were also several other smaller patches of significant prediction accuracy for word2vec, including the scene-selective OPA and PPA.

Figure 6. Whole-brain voxelwise encoding models of object context. Voxelwise encoding models were estimated and assessed using the 9-fold cross-validation procedure described in Figure 3. These analyses were performed for all voxels with split-half reliability scores greater than or equal to $r = 0.1841$, which corresponds to a p -value of 0.05 (see split-half reliability mask in Fig. S4). Encoding model accuracy scores are plotted for voxels that show significant effects ($p < 0.05$, FWE-corrected, permutation test). The left panel shows prediction accuracy for the image-based object2vec encoding model, and the right panel shows prediction accuracy for the language-based word2vec encoding model. There is a cluster of high prediction accuracy for the object2vec model overlapping with the anterior portion of the PPA, mostly in the right hemisphere. There are also several other smaller clusters of significant prediction accuracy throughout high-level visual cortex, including clusters in the OPA and RSC. The word2vec model produced a cluster of high prediction accuracy that partially overlapped with the significant voxels for the object2vec model in the right PPA but extended into a more lateral portion of ventral temporal cortex as well as lateral occipital cortex. ROI parcels are shown for scene-selective ROIs. PPA = parahippocampal place area, OPA = occipital place area, RSC = retrosplenial complex.

Category selectivity and co-occurrence models

The results of the ROI and whole-brain analyses suggested a general trend for the image-based object2vec to be associated with scene-selective regions and the language-based word2vec to be associated with object-selective regions. To quantify this trend, we performed a follow-up analysis that compared the voxelwise effects of the encoding models with measurements of

scene-selectivity and object-selectivity from the functional localizer data. Specifically, we extracted voxelwise responses to scenes and objects from a separate set of localizer data and computed activation difference scores for scenes minus objects. In the same voxels, we also computed accuracy difference scores for the encoding-model effects of object2vec minus the encoding model-effects of word2vec. We restricted our analyses to voxels that showed a significant effect in either the object2vec or word2vec encoding models, and we computed the correlation of the activation and accuracy difference scores across these voxels. This analysis showed that a simple difference score of scene and object selectivity could account for 28% of the variance in the differential effects of object2vec and word2vec across voxels (Fig. 7; $r^2 = 0.28$; results for single subjects are shown in Figure S2). Specifically, this analysis showed that voxels that were better explained by a model based on the co-occurrence statistics of objects in images tended to have greater scene selectivity, whereas voxels that were better explained by a model based on the co-occurrence statistics of object names in language tended to have greater object selectivity.

Figure 7. Relationship between encoding-model accuracy and category selectivity. This scatter plot shows how the voxelwise encoding-model effects are related to category selectivity. Each point is a voxel in standard space, and this plot includes all voxels with significant effects for either the image-based object2vec encoding model or the language-based word2vec encoding model. The x-axis plots the difference in prediction accuracy between the object2vec encoding model and the word2vec encoding model (averaged across participants). The y-axis plots the difference in activation between scenes and objects, based on data from a separate set of functional localizer runs. The positive linear relationship shows that voxels that are better predicted by object2vec than word2vec tend to be more scene-

selective, whereas voxels that are better predicted by word2vec than object2vec tend to be more object-selective (Pearson $r = 0.53$).

Principal components of voxel tuning

We next performed an exploratory analysis to visualize how tuning for the object2vec and word2vec models varied across voxels. For each model (object2vec and word2vec), we applied PCA to the regression weights for all voxels that had significant prediction accuracies. We visualized the first four principal components (PCs), by projecting each object category onto these PCs. Figure 8 shows the object categories color coded according to their score on each PC. A common theme across most PCs is the broad distinction between indoor and outdoor environments. Additionally, each PC emphasizes different subsets of indoor and outdoor objects. For example, PC2 differentiates electronics and appliances from other indoor objects (e.g., microwave and computer vs. soap and napkin), and it differentiates between man-made and natural items found in outdoor scenes (e.g., sea and sand vs. sidewalk and car). However, these PCs also contain complex mixtures of object clusters that are not easily interpretable as representing any single semantic dimension. The semantic complexity of these PCs may reflect the mixed-selectivity that arises from encoding a large set of object contexts in a small number of representational dimensions (i.e., the 8 dimensions of object2vec). A similar visualization is provided for the word2vec model in Figure S3. These exploratory visualization analyses can help us to develop intuitions about cortical tuning for object context, but as with many data-driven analyses, it is important to keep in mind that our interpretations of these visualizations are *post hoc* and have not been statistically tested.

Figure 8. Principal components of voxel tuning for image-based object context. Principal components analysis was used to examine variance in encoding-model regression weights across voxels. This plot illustrates the first four principal components (PCs) of the regression weights for the image-based object2vec encoding model, using all voxels with significant prediction accuracies. The 81 object categories from the fMRI experiment were projected onto each PC (as indicated by the color coding from blue to red). The spatial arrangement of the object categories is the same as the tSNE plot in Figure 1. At a coarse level, most PCs broadly distinguish between items found in indoor and outdoor environments. At a more fine-grained level, each PC organizes the objects into distinct sets of clusters, some of which appear to be semantically interpretable. For example, PC2 differentiates electronics and appliances from other indoor objects (e.g. microwave and computer vs. soap and napkin), and it also differentiates between man-made and natural items found in outdoor scenes (e.g., sea and sand vs. sidewalk and car). However, these PCs also contain complex clusters of objects that do not appear to have intuitive semantic interpretations. This semantic complexity may reflect the mixed-selectivity that arises from encoding a large set of object contexts in a small number of representational dimensions (i.e., the 8 dimensions of object2vec).

DISCUSSION

The goal of this study was to determine if object representations in visual cortex reflect the statistical regularities of object co-occurrence in the visual environment. We first developed an objective model of the latent statistical structure of object co-occurrence using an unsupervised machine-learning algorithm applied to a large set of annotated images of real-world scenes. We then compared the representations of this model with fMRI activity evoked by single objects shown in isolation. We found that responses of voxels in scene-selective visual cortex—most strongly, the anterior PPA—were predicted by the co-occurrence statistics of objects in visual scenes. In contrast, the representations of a language-based co-occurrence model for object names predicted the responses of voxels in object-selective cortex. These findings indicate that scene-selective cortex automatically links objects with representations of their visual contexts and that anterior PPA is a particularly important locus for performing this operation. They further suggest that the statistical regularities of vision and language map onto partially distinct components of cortical object processing. Below we consider the implications of these results, both for our understanding of how visual and semantic context is represented in the brain, and for our understanding of cortical function more generally.

Contextual representations in scene regions

Our results provide new insights into how visual context is represented in the brain, a longstanding issue in visual neuroscience. Previous behavioral studies have shown that the associations between objects and their typical contexts can influence performance on perceptual tasks, including object recognition, scene recognition, and visual search^{1–4}. For example, when an object is shown in an unusual context (e.g., a priest on a football field), subjects are slower and less accurate at recognizing both the foreground object and the background scene¹. One possible mechanism that could underlie these behavioral effects is the automatic representation of an object’s statistical associations to other objects and scene elements. Previous neuroimaging work on contextual associations has found evidence consistent with this hypothesis. Specifically, it has been shown that objects with strongly associated scene contexts (e.g., stoves are strongly associated with kitchen contexts), elicit an increased mean fMRI signal in the PPA and RSC, two regions of scene-selective visual cortex^{19,27}. However, the implication of these results has been unclear, because several other subjective ratings of object properties, such as landmark suitability²⁰ and spatial definition²¹, have been found to covary with contextual-association ratings and explain similar or more variance in fMRI responses. Moreover, previous studies have not investigated the fundamental question of whether object-evoked responses in visual cortex represent the underlying multidimensional structure of object context (i.e., which objects are associated with each other).

Our work directly addresses this issue by explicitly modeling the statistics of object co-occurrence in the visual environment and relating these statistics to object-evoked fMRI responses. We overcame the challenge of objectively quantifying the real-world statistics of object context by leveraging a large dataset of densely labeled images that was originally developed for training semantic segmentation models in computer vision²⁵. We analyzed the annotations for these images using *object2vec*, which is a modified version of the *word2vec* algorithm from computational linguistics^{23,24}. An important aspect of this modeling procedure is the fact that we explicitly analyzed the co-occurrence of object categories in scenes. This was possible because we applied a text-based machine-learning algorithm to image annotations. In comparison, a learning algorithm that took pixel values as inputs, such as a convolutional neural network, might also learn the natural statistics of object co-occurrence³², but this co-occurrence information would be latent in the model's internal representations and would thus be difficult to observe and quantify.

A similar approach of combining methods from computational linguistics with image annotations was previously used to model scene categorization behaviors³³ and to show that the contextual information derived from object co-occurrence statistics is at least partially correlated with semantic information derived from the co-occurrence statistics of natural language³⁴. A similar method was also used in an earlier fMRI study to model the influence of object co-occurrence statistics on scene representation³⁵. The authors in this last study used a topic-modeling procedure to identify sets of scene “categories” that were defined by the co-occurrence statistics of objects in images. They found that scene-evoked fMRI responses throughout much of high-level visual cortex could be predicted from a set of co-occurrence-defined scene categories, which suggests that scene representations in the human brain are shaped by the co-occurrence of objects in the visual environment. Our work uses a similar approach to address a different set of questions. We set out to understand how visual cortex links objects with representations of their contexts. For this purpose, it was crucial that we used isolated single objects as stimuli, rather than full scenes (for which contextual information would already be present in the visual display). This approach allowed us to identify a set of cortical regions that encode the contextual information associated with visual objects, even in the absence of a surrounding scene.

It is notable that the strongest evidence for contextual coding was found in the anterior PPA. Previous fMRI studies of visual context typically found results throughout the PPA and other scene-responsive regions¹¹; although one study found a more anterior parahippocampal locus (outside of the PPA) for non-spatial associations between objects¹⁶. Our findings are consistent with other data implicating anterior PPA in the coding of abstract information related to scenes and contexts. For example, in a previous study we found that multivoxel activation codes in this

region generalized across interior and exterior views of the same building, indicating some degree of abstraction related to building identity³⁶. Functional connectivity data also suggest the possibility of a division between anterior and posterior PPA, with posterior PPA more connected to areas implicated in scene perception, such as the OPA and early visual cortex, and anterior PPA more connected to regions involved in spatial perception and memory, such as RSC, caudal inferior parietal lobule, and medial temporal lobe^{28,37}. The fact that anterior PPA encodes the high-level statistics of object co-occurrence is consistent with previous findings of functional differences along the anterior-posterior axis of PPA. Beyond the anterior PPA, we found weaker though significant object context effects in posterior PPA, scene-selective RSC and OPA, and in object-selective pFs. Thus, although anterior PPA may be most crucial, contextual processing appears to rely on a broad network of high-level visual regions that extends beyond PPA.

Statistical properties of vision and language

In addition to looking at representations of object co-occurrence statistics in visual scenes, we also examined the co-occurrence statistics of object names in linguistic corpora. We found an interesting and unexpected relationship between the modality of the distributional models (i.e., vision or language) and the category-selectivity of the regions that these models best predicted. Specifically, we found that language-based word2vec was more strongly associated with object-selective regions, whereas image-based object2vec was more strongly associated with scene-selective regions. Previous work has shown that language-based models of distributional statistics are partially similar to object-evoked representations in both fMRI and magnetoencephalography^{38,39}, but no previous studies have examined the possibility that responses in object-selective cortex might be better predicted by the distributional statistics of language than the distributional statistics of vision.

What might account for this pattern of findings? One relevant consideration is that language-based word2vec likely contains more *taxonomic* information (e.g., What class of object is this?), whereas image-based object2vec emphasizes *thematic* information (e.g., What places and events is this object found in?). Indeed, previous findings from computational linguistics suggest that standard models of distributional semantics often perform better on tests of taxonomic knowledge than tests of thematic knowledge⁴⁰. Thus, one possibility is that the representations of word2vec and object2vec reflect different domains of abstract semantic associations. Some previous work that is consistent with this interpretation has shown that the posterior fusiform gyrus and parahippocampal cortex are involved in the processing of object semantics, even when the experimental stimuli are words rather than images^{41,42}. However, a different, but not mutually exclusive, possibility follows from the fact that objects from similar taxonomic categories tend to have similar shapes. Insofar as the word2vec embeddings for object terms in

language capture information about object shape, they may be better predictors of perceptual responses in object-selective visual cortex. Our findings cannot adjudicate between these possibilities. Thus, the role of abstract associations and perceptual features in explaining these contextual representations should be explored in future work.

Natural statistics and cortical representation

Our results have implications for understanding cortical function more generally. A fundamental theory in visual neuroscience is that neurons in the visual system are tuned to the natural statistics of the sensory environment^{43,44}. Specifically, it is thought that neural response preferences are shaped through evolutionary and developmental processes to optimize “efficiency”—that is, to convey as much useful information as possible with the limited computational resources of the brain. One potential way to improve the efficiency of a neural population code is to match neural tuning properties to the natural statistics of sensory stimuli, such that the proportion of neurons tuned to a particular stimulus is related to the frequency of that stimulus in sensory experience. These ideas have been investigated in the context of low-level vision by calculating sensory statistics from readily available image properties, such as pixel intensities^{45–47}. The natural statistics of higher-level sensory information are also likely to be relevant for understanding the functions of visual cortex, but a major challenge in testing this idea is the need for large-scale annotated stimulus sets and quantitative approaches for modeling their statistical structure. By combining annotated image datasets with tools from computational linguistics, as we do here, vision researchers can benefit from decades of work that has sought to model the latent structure of language from the distributional statistics of words in text²⁶.

The approach of examining the high-level statistical regularities of objects and scenes may provide insight into longstanding questions about cortical function. In the case of scene-selective cortex, previous work has shown that scene regions respond to several high-level properties of objects, including contextual-association strength, real-world size, landmark suitability, spatial definition, and interaction envelope^{11,21,48–50}. Scene-selective regions also respond to several mid-level properties of images, including high spatial frequencies, rectilinearity, and cardinal orientations^{51–54}. Many of these high-level object properties appear to covary in the visual environment and explain similar variance in fMRI responses²⁰, and they may also covary with mid-level image properties^{55,56}. One question that is often raised is whether any of these high-level or mid-level stimulus properties is of primary importance in explaining the functions of scene-selective cortex. Although it is possible that cognitive demands require scene-selective regions to be specialized for processing a specific and limited class of scene features, an alternative possibility is that these regions are tuned to the covariance of features in real-world scenes, and should thus show responses to a variety of stimulus properties, ranging from high-

level to low-level. In this view, the representations of scene-selective cortex may reflect a statistically efficient basis set from which a large number of behaviorally relevant aspects of scenes can be decoded. Our current data support this idea by showing that a large portion of explainable variance in PPA can be accounted for by a low-dimensional model based on the natural statistics of objects in scenes. We predict that a number of other high- and mid-level object and scene properties are linked to these statistics, which would be consistent with the hypothesis that cortical tuning to statistical regularities produces representations that are informative to a broad range of stimulus properties. Similar statistical principles may underlie the cortical representations that support other high-level cognitive functions, such as object categorization⁵⁷, face recognition⁵⁸, and reinforcement learning⁵⁹.

This perspective suggests several questions for future investigations. One open question is whether visual cortex contains representations that reflect the natural statistics of object locations in images. Here we examined object co-occurrence independent of object location, but it is known that the visual system also takes advantage of associations between objects and spatial locations⁶⁰ and that perceptual behaviors and cortical representations are sensitive to violations of the typical positions of objects^{2,61–63}. Future investigations could utilize annotated image databases to explicitly quantify these spatial statistics and to examine how they relate to the object representations of visual cortex. A second open question relates to how the statistics of object co-occurrence are learned and incorporated into the representations of visual cortex. One possibility is that these statistics are initially encoded in memory regions of the medial temporal lobe and then transferred to visual cortex⁶⁴. A third open question is whether object2vec or similar image-based approaches could be merged with word embeddings from computational linguistics to create richer representational models of object semantics⁶⁵. Our findings suggest that object2vec could complement language-based representations by bringing in perceptual information that is not easily learned from language corpora.

Our findings demonstrate that the statistical properties of scenes have strong explanatory power in modeling the representations of scene-selective visual cortex. Future research on high-level vision can benefit from a greater emphasis on the statistical basis of cortical representations and their utility for behaviorally relevant computational goals.

Conclusions

We adapted a machine-learning technique from computational linguistics to model the latent statistics of object context in scenes. Using this model, we found that scene-selective visual cortex represents the natural statistics of object context, as previously suggested, but never conclusively demonstrated. Overall, our findings are consistent with the hypothesis that the

coding mechanisms of sensory systems leverage the latent statistical structure of the environment, which may reflect a fundamental strategy for optimizing representational efficiency in the brain.

METHODS

Subjects. Four healthy subjects (2 female, ages 30, 32, 32, and 34) from the University of Pennsylvania community participated in the experiment. All had normal or corrected-to-normal vision and provided written informed consent in compliance with procedures approved by the University of Pennsylvania Institutional Review Board.

MRI acquisition. Participants were scanned on a Siemens 3.0 T Prisma scanner using a 64-channel head coil. We acquired T1-weighted structural images using an MPRAGE protocol (TR = 2200 ms, TE = 4.67 ms, flip angle = 8°, matrix size = 192 x 256 x 160, voxel size = 0.9 x 0.9 x 1 mm). We acquired T2*-weighted functional images sensitive to blood oxygenation level-dependent contrasts using a multiband acquisition sequence (TR = 2000 ms for main experimental scans and 3000 ms for localizer scans, TE = 25 ms, flip angle = 70°, multiband factor = 3, matrix size = 96 x 96 x 81, voxel size = 2 x 2 x 2 mm).

Stimuli. Subjects viewed images of individual objects presented on textured backgrounds. The objects were from 81 different categories, with 10 unique images per category, resulting in a total of 810 experimental images. Images of objects were obtained by searching for the object category terms on Google Images and downloading high-quality images in which the object was prominent. We then manually isolated the object from the image background using Adobe Photoshop. All images were cropped around the isolated object and then resized to 600 x 600 pixels with the original aspect ratio preserved (thus, the shorter axis of the image was padded with transparent pixels).

We placed the objects over complex, textured backgrounds (Fig. 2). These background textures were generated using a procedure that was designed to systematically reduce the similarity between our models of interest (i.e., object2vec and word2vec) and a model of low-to-mid-level perceptual features (i.e., AlexNet trained on ImageNet). We used representational similarity analysis (RSA) to assess the similarity of these models. First, representations were generated for all stimuli as the vectorized outputs of each model (i.e., object2vec embeddings, word2vec embeddings, and unit activations of AlexNet). For AlexNet, representations for the 10 images within each object category were averaged to create a single representation per category. Next, representational dissimilarity matrices (RDMs) were created by calculating all pairwise comparisons of the representational vectors for the 81 object categories. Representational dissimilarities between categories were measured using Pearson distance (one minus the Pearson correlation coefficient). We then assessed the similarity between models by calculating the Spearman correlation of their RDMs, which we refer to as an RSA correlation.

Background textures were created from composites of DeepDream visualizations for all units in convolutional layer 5 of AlexNet using the MATLAB function `deepDreamImage`. DeepDream visualizations were initiated with random pixel values and optimized through gradient ascent using a multi-resolution image pyramid and Laplacian Pyramid Gradient Normalization (3 pyramid levels, scaling of 1.4 between pyramid levels, 10 iterations per level). We created composite textures by averaging the visualizations from three channels sampled randomly with replacement and then adjusting the overall luminance to 90 in Lab color space. We used a stochastic optimization procedure to synthesize composite background textures for each object image that minimized the RSA correlations of the CNN layers with both the `object2vec` and `word2vec` models. Our optimization criterion, which we sought to minimize, was the maximum correlation coefficient ($r_{\max}$) across all RSA comparisons between the models of interest (i.e., `object2vec` and `word2vec`) and the convolutional layers of AlexNet (layers 1-5). We began by randomly synthesizing 3,000 composite background textures and generating 1,000 random assignments of these backgrounds to the 810 object images. From these, we selected the set of object-background pairings with the lowest $r_{\max}$. We then implemented an iterative optimization procedure in which the image backgrounds were individually altered to gradually reduce $r_{\max}$. On each iteration, a new randomly synthesized background texture was generated for a single randomly selected object and $r_{\max}$ was re-calculated. If $r_{\max}$ was lower than the current best value, the new object-background pairing from that iteration was retained; otherwise, it was discarded. This procedure was stopped shortly after $r_{\max}$ dropped below 0.05.

We also created a set of warped object stimuli by applying a diffeomorphic transformation to the 810 isolated object images⁶⁶. The diffeomorphic warping was implemented over 5 iterations with a maximum distortion of 30. These warped objects served as target stimuli for a category-detection task in the fMRI scanner (see Fig. S1 for examples).

Procedure. We collected fMRI data while subjects viewed images of objects on textured backgrounds and performed a category-detection task. Subjects were asked to fixate on a central cross that remained on the screen throughout the scan and press a button whenever the stimulus was a warped object. Successful performance on this task required subjects to attend to each stimulus but did not require them to think about the mnemonic associations of the objects. Stimuli were presented in a mini-block design. In each block, 5 images from the same object category were presented in rapid succession. Each image was shown for 500 ms, and consecutive images were separated by a 500 ms inter-stimulus interval, during which a random background texture was shown, which differed from the backgrounds that the objects were shown on. After the fifth image in each block, a gray screen with a central cross appeared for 1.5 s before the start of the next block. Each block was thus 4.5 s in length, with a minimum inter-block interval of 1.5

s. Every run also included 6 randomly placed null events of 4 s each, which always occurred between blocks.

Consistent with previous studies using encoding model analysis methods, we sought to obtain a large amount of fMRI data from each participant, rather than data from a large number of participants as is typical in standard fMRI studies. The experiment was designed to be completed in 4 sessions, each taking place on different days; however, two subjects chose to split up the scanning sessions over more than 4 days. For each planned session, we randomly divided the 81 object categories into 9 folds of 9 categories each. These folds were presented in separate runs, randomly ordered. Runs contained 6 mini-blocks for each of the 9 categories; images were assigned to mini-blocks by creating 3 different 2-fold splits for the 10 images in each category. Each run also contained 1 to 3 warped-object mini-blocks, in which a single warped object image was shown among 4 normal images from the same object category. The categories for the warped-object mini-blocks were randomly chosen from the 9 categories in the run. The order of the mini-blocks within the run was randomized, as was the order of the images within each mini-block, subject to the restriction that mini-blocks from the same category were never shown more than twice in a row. Runs were ~6.3 min in length, with the exact length depending on the variable number of warped-object mini-blocks (i.e., 1-3).

In two of the scanning sessions, additional fMRI data were collected for defining functional ROIs (described in the section Functional localizer and regions of interest). In another two scanning sessions, we also collected 11 minutes of fMRI data while subjects viewed a continuous natural movie (from the show Planet Earth). These movie data were collected for use in hyperalignment⁶⁷, and they were not analyzed for the current project.

fMRI preprocessing. fMRI data were processed and modeled using SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK) and MATLAB (R2019b Mathworks). Functional images for each participant were realigned to the first image, co-registered to the structural image, warped to Montreal Neurological Institute (MNI) standard space, and spatially smoothed with a 6-mm full-width-half-maximum isotropic Gaussian kernel. Voxelwise responses to each object category were estimated using a general linear model, which included regressors for each of the object categories and the target (warped-object) mini-blocks in each run, and a regressor for each run. Each regressor consisted of a boxcar function convolved with a standard hemodynamic response function. Low-frequency drifts were removed using a high-pass filter with a cutoff period of 128 s, and temporal autocorrelations were modeled with a first-order autoregressive model. Voxelwise beta values were z-scored across the 9 object categories within each run. These beta values were used as object category responses for all additional analyses.

Voxelwise reliability estimates. We calculated noise-ceiling estimates based on the split-half reliability of voxelwise responses to the 81 object categories. For each voxel, split-half reliability was calculated as the Pearson correlation between the mean responses to the 81 object categories in odd and even runs. The resulting reliability estimates were used to select voxels for further analyses⁶⁸. Only voxels whose average split-half reliability across subjects was greater than or equal to $r = 0.1841$, corresponding to an uncorrected p-value of 0.05, were included in ROIs (see next section), or plotted in whole brain maps (Fig. S4).

Functional localizer and regions of interest. BOLD responses in two additional scan runs were used to define functional ROIs. The localizer protocol included 4.5-s blocks of scenes, faces, objects, and scrambled objects. Images were presented for 225 ms with a 75-ms inter-stimulus interval, and subjects performed a one-back repetition-detection task on the images. We identified the three scene-selective ROIs (PPA, OPA, and RSC) based on a contrast of scenes > objects implemented in a general linear model with a group-based anatomical constraint of category-selective activation derived from a large number of localizer subjects in our laboratory ($n = 42$)^{36,69}. Using a similar parcel-based approach, we defined two object-selective ROIs (LO and pFs) based on a contrast of objects > scrambled objects, and we defined EVC with the contrast of scrambled objects > scenes. Bilateral ROIs were defined by selecting the top 50 voxels from the parcel in each hemisphere based on the localizer contrast (i.e., the difference in beta values across conditions), excluding voxels that did not meet the split-half reliability criterion described above. For the anterior and posterior PPA ROIs, we restricted the full PPA parcel to either its anterior or posterior third and selected 25 voxels per hemisphere from each third.

Representational models. Our analyses made use of two representational models, object2vec and word2vec, which are described below.

Object2vec. Using tools from computational linguistics, we developed an approach for characterizing the latent statistical structure of object co-occurrence in natural images, which we refer to as object2vec. To train this model, we used annotated images from the ADE20K dataset: <http://groups.csail.mit.edu/vision/datasets/ADE20K/>. This dataset contains 22,210 scene images that have been densely annotated by human observers such that every object within every scene is segmented and labelled. Across the entire scene image set, there are 3,148 unique object segmentation labels²⁵. ADE20K is ideal for our purposes because it provides a large and diverse set of image annotations that we can use to model the co-occurrence statistics of objects in natural images.

Our modeling approach was inspired by the field of computational linguistics, which has a long history of modeling the lexical-semantic properties of words based on their co-occurrence

statistics in natural linguistic contexts²⁶. These co-occurrence-based models are referred to as distributional-semantic models because they reflect the distribution of words in language (often written language). The general goal of these models is to learn word representations (i.e., vectors) that capture the contextual similarity of words—that is, representations in which words that occur in similar contexts are similar. A leading algorithm for learning these distributional-semantic representations from text corpora is word2vec^{23,24}. Object2vec is based on the continuous bag of words (CBOW) version of word2vec, which seeks to identify word representations that can be used to predict a missing target word from its surrounding context words (e.g., the other words around it in a sentence).

We adapted the word2vec CBOW algorithm, as implemented in fastText⁷⁰, and applied it to the image annotations from ADE20K. Our implementation modifies how the context window is defined for word2vec. Typically, each word’s context is defined as a weighted window spanning forwards and backwards from the position of the word in a sentence. This approach to defining the context window is not applicable to our image annotations, which can either be analyzed as a bag of words, containing object labels in an arbitrary order, or as a 2D segmentation map. Here we implemented the simpler bag-of-words approach by defining the context window for each object so that it includes all other objects in the same image (with equal weights for all context objects). Note that fastText allows for the possibility of learning representations for character n-grams (i.e., groups of letters within words), but this option was not applicable to our image annotations and was thus not used.

We first created an annotation dataset to use as our training corpus. We converted each image into a list of object labels, without including repeats for objects that occurred multiple times in the same image. The object labels were cleaned up and simplified in several ways. In ADE20K, each object can be associated with a list of multiple possible names (e.g., “bathtub, bathing tub, bath, tub”), which we converted into a single name using the first term in the list (e.g., “bathtub”). For the 81 object categories in our experimental stimuli, we manually inspected the object labels in ADE20K to identify other possible labels that could be applied to these objects (e.g., “bowl” and “bowls”, “notebook” and “notepad”). For categories with multiple associated labels, we converted all instances to one identical label (e.g., all instances of “notepad” were converted to “notebook”).

We created the object2vec embeddings by running our modified version of the fastText CBOW algorithm on the ADE20K annotation corpus. These embeddings are learned through stochastic gradient descent and the dimensionality of the embeddings is set by hand before model training. We first created embeddings of different sizes, spanning from 10 to 90 dimensions in steps of 20. For each dimensionality, we created 100 versions using different random initializations. The

models were trained for 1000 epochs with a negative sampling parameter of 20. We sought to identify the simplest possible model that captured the contextual representations of the objects in ADE20K. By converting the resulting object2vec representations into RDMs, we could perform RSA to compare the information contained in the embeddings at different dimensionalities. We averaged the RDMs for all 100 versions of each dimensionality and then performed pairwise correlations of the RDMs for all dimensionalities. We found that the lowest-dimensional model (10 dimensions) was highly correlated with each of the higher-dimensional models (mean r -value = 0.92, sd = 0.02). Because of the strong representational similarity of these models, we used the simpler 10-dimensional embeddings for all further analyses. We then performed PCA on the 10-dimensional embeddings to determine if they could be further reduced in dimensionality. We found that on average across all 100 versions, 8 dimensions were sufficient to explain over 90% of the variance. Finally, we concatenated all versions of the 10-dimensional embeddings and used PCA to extract latent dimensions that were commonly identified across different random initializations. We retained 8 principal components (PCs), which served as our object2vec representations.

Word2vec. In addition to our object2vec embeddings, which were trained on image annotations, we also examined word2vec embeddings trained on written language. We used a set of 300-dimensional word2vec embeddings that had been previously trained on the Google News dataset (~100 billion words): <https://code.google.com/archive/p/word2vec/>. We downloaded the word2vec embeddings here: https://github.com/chrisjmccormick/word2vec_matlab. These embeddings were filtered to include only words that are present in WordNet, which removes many entries that are not useful for our analysis of object concepts, including misspellings and multi-word phrases. We then used PCA to reduce the dimensionality of these representations. We inspected the scree plot of explained variance for each PC and found an elbow at ~30 PCs, where explained variance began to level off. We therefore retained 30 PCs of the word2vec embeddings for all further analyses. We later ensured that our findings were not contingent on the specific number of PCs retained. We observed similar results in the range of 20 to 50 PCs. Beyond 50 PCs, cross-validated accuracy of the encoding models decreased, due to overfitting. For each of the 81 object categories in our experiment, we created a list of its associated names. This list included all names in the full ADE20K label for the object (e.g., "bathtub, bathing tub, bath, tub") as well as plural forms and any additional names that were detected through manual inspection of the WordNet-filtered word2vec vocabulary (207,147 words). For each object category, we averaged the embeddings for all of its associated names. The resulting 30-dimensional vectors served as our word2vec representations.

Voxelwise encoding models. We used voxelwise encoding models to determine if the fMRI responses to our 81 object categories could be predicted from the representations of object2vec

and word2vec. For each subject, voxelwise responses to the 81 object categories were determined by averaging responses across all runs. Ordinary least squares regression was used to estimate weights that map representational features (e.g., object2vec) to these fMRI responses. We used 9-fold cross-validation to quantify the out-of-sample prediction accuracy for each voxelwise model. This cross-validation design was built into our fMRI protocol. As described above (Procedure), the stimuli for each subject were randomly split into 9 folds of 9 object categories each, and the object categories from different folds were always shown in different runs. For each iteration of the cross-validation, voxelwise regression weights for the object2vec and word2vec dimensions were estimated using object categories from 8 of the 9 folds, and the estimated regression weights were applied to the object categories in the held-out fold to generate predicted fMRI responses.

We included nuisance regressors in our encoding models to account for low-level stimulus properties that could influence the estimated weights for our regressors of interest (i.e., object2vec and word2vec). Nuisance regressors were generated using the five convolutional layers of a convolutional neural network (AlexNet) trained on ImageNet⁷¹: <http://www.vlfeat.org/matconvnet/models/imagenet-caffe-ref.mat>. The goal of this procedure was to create a small set of nuisance regressors that explain prominent low-level image features while not drastically increasing the overall number of regressors and thus causing problems with overfitting. We ran the experimental stimuli through AlexNet to obtain activations from the final output of each convolutional layer and then concatenated these activations into a single vector for each stimulus. We averaged these AlexNet activation vectors across all 10 images for each object category and used PCA to reduce their dimensionality. Based on inspection of the scree plot, we retained 20 PCs, which is at the point where explained variance leveled off. In the cross-validation procedure, the nuisance regressors were only included in the training folds for the purpose of estimating regression weights—they were not included when applying the regression weights to generate predictions for the object categories in the held-out test fold. Thus, all encoding-model predictions were computed using only the representations of object2vec or word2vec.

We used permutation tests to assess the statistical significance of prediction accuracy scores. To estimate the distribution of effects within each voxel under the null hypothesis, we randomly permuted the object category labels within each fold of the cross-validation design and then computed the Pearson correlation between the actual fMRI responses and the permuted version of the predicted responses. We repeated this procedure 5,000 times. For whole-brain analyses, we averaged the actual prediction accuracy scores in each voxel across subjects and did the same for the r-values from all 5,000 iterations of the permutation procedure. We then logged the maximum r-value across all voxels on each iteration, which produced a null distribution

controlled for family-wise error rate. We compared the actual r -value in each voxel to this null distribution to generate p -values. For ROI analyses, we calculated the average r -value across all voxels in all subjects and did the same for the r -values from all 5,000 iterations of the permutation procedure. We compared the actual r -value for each ROI to the permutation-based null distribution to generate p -values. For the ROI analyses, we also obtained bootstrap standard errors by randomly resampling the actual and predicted fMRI responses within each fold and then re-computing the voxelwise correlations and ROI means. This bootstrap resampling procedure was performed 5,000 times.

We also performed a permutation-based interaction test on the encoding-model accuracy scores for object2vec and word2vec in anterior and posterior PPA. For each subject, we subtracted the voxelwise accuracy for word2vec from the accuracy for object2vec. We calculated the average of these differences scores across all voxels in all subjects for each ROI, and we subtracted the average in posterior PPA from the average in anterior PPA. The resulting statistic reflects the degree to which object2vec produced higher prediction accuracy than word2vec in anterior PPA relative to posterior PPA. We calculated the same interaction statistic for 5,000 iterations of our permutation procedure, and we compared the actual interaction statistic to the permutation-based null distribution to generate a p -value.

Principal component analysis of encoding-model regression weights. We performed a *post hoc* exploratory analysis to examine variance in the encoding model regression weights across voxels. Each voxel has a set of corresponding regression weights for each encoding model (e.g., 8 regression weights per voxel for object2vec). We calculated the average regression weights across subjects for each voxel, and we then performed PCA on these average regression weights across all voxels that had significant effects in the whole-brain analysis. We performed this procedure separately for the object2vec and word2vec encoding models. We applied the loadings for the first four principal components to the original regressors and generated PC scores for each object category in our experiment. To visualize these PCs, we color-coded the t-distributed stochastic neighbor embedding (tSNE) plot of object2vec from Figure 1 according to the scores for each PC (see Fig. 8 and Fig. S3).

REFERENCES

1. Davenport, J. L. & Potter, M. C. Scene Consistency in Object and Background Perception. *Psychol Sci* **15**, 559–564 (2004).
2. Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology* **14**, 143–177 (1982).
3. Palmer, S. E. The effects of contextual scenes on the identification of objects. *Memory & Cognition* **3**, 519–526 (1975).
4. Chun, M. M. & Jiang, Y. Top-Down Attentional Guidance Based on Implicit Learning of Visual Covariation. *Psychol Sci* **10**, 360–365 (1999).
5. Bar, M. The proactive brain: memory for predictions. *Phil. Trans. R. Soc. B* **364**, 1235–1243 (2009).
6. Davachi, L. & DuBrow, S. How the hippocampus preserves order: the role of prediction and context. *Trends in Cognitive Sciences* **19**, 92–99 (2015).
7. Hindy, N. C., Ng, F. Y. & Turk-Browne, N. B. Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat Neurosci* **19**, 665–667 (2016).
8. Friston, K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci* **11**, 127–138 (2010).
9. Henderson, J. M. Gaze Control as Prediction. *Trends in Cognitive Sciences* **21**, 15–23 (2017).
10. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat Neurosci* **20**, 1643–1653 (2017).
11. Bar, M. & Aminoff, E. Cortical Analysis of Visual Context. *Neuron* **38**, 347–358 (2003).

12. Preston, T. J., Guo, F., Das, K., Giesbrecht, B. & Eckstein, M. P. Neural Representations of Contextual Guidance in Visual Search of Real-World Scenes. *Journal of Neuroscience* **33**, 7846–7855 (2013).
13. Brandman, T. & Peelen, M. V. Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. *J. Neurosci.* **37**, 7700–7710 (2017).
14. Brandman, T. & Peelen, M. V. Signposts in the Fog: Objects Facilitate Scene Representations in Left Scene-selective Cortex. *Journal of Cognitive Neuroscience* **31**, 390–400 (2019).
15. Bar, M., Aminoff, E. & Schacter, D. L. Scenes Unseen: The Parahippocampal Cortex Intrinsically Suberves Contextual Associations, Not Scenes or Places Per Se. *Journal of Neuroscience* **28**, 8539–8544 (2008).
16. Aminoff, E., Gronau, N. & Bar, M. The Parahippocampal Cortex Mediates Spatial and Nonspatial Associations. *Cerebral Cortex* **17**, 1493–1503 (2007).
17. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
18. Epstein, R. A. & Baker, C. I. Scene Perception in the Human Brain. *Annu. Rev. Vis. Sci.* **5**, 373–397 (2019).
19. Aminoff, E. M., Kveraga, K. & Bar, M. The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences* **17**, 379–390 (2013).
20. Troiani, V., Stigliani, A., Smith, M. E. & Epstein, R. A. Multiple Object Properties Drive Scene-Selective Regions. *Cerebral Cortex* **24**, 883–897 (2014).
21. Mullally, S. L. & Maguire, E. A. A New Role for the Parahippocampal Cortex in Representing Space. *Journal of Neuroscience* **31**, 7441–7449 (2011).

22. Epstein, R. A. & Ward, E. J. How Reliable Are Visual Context Effects in the Parahippocampal Place Area? *Cerebral Cortex* **20**, 294–303 (2010).
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* **26**, 3111–3119 (2013).
24. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (2013).
25. Zhou, B. *et al.* Semantic Understanding of Scenes Through the ADE20K Dataset. *Int J Comput Vis* **127**, 302–321 (2019).
26. Lenci, A. Distributional Models of Word Meaning. *Annu. Rev. Linguist.* **4**, 151–171 (2018).
27. Bar, M. Visual objects in context. *Nat Rev Neurosci* **5**, 617–629 (2004).
28. Baldassano, C., Beck, D. M. & Fei-Fei, L. Differential connectivity within the Parahippocampal Place Area. *NeuroImage* **75**, 228–237 (2013).
29. Baldassano, C., Esteva, A., Fei-Fei, L. & Beck, D. M. Two Distinct Scene-Processing Networks Connecting Vision and Memory. *eneuro* **3**, ENEURO.0178-16.2016 (2016).
30. Grill-Spector, K. *et al.* Differential Processing of Objects under Various Viewing Conditions in the Human Lateral Occipital Complex. *Neuron* **24**, 187–203 (1999).
31. Malach, R. *et al.* Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences* **92**, 8135–8139 (1995).
32. Zhu, Z., Xie, L. & Yuille, A. Object Recognition with and without Objects. *Proceedings of the 26th International Joint Conference on Artificial Intelligence* 3609–3615.

33. Greene, M. R. Statistics of high-level scene context. *Front. Psychol.* **4**, (2013).
34. Sadeghi, Z., McClelland, J. L. & Hoffman, P. You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia* **76**, 52–61 (2015).
35. Stansbury, D. E., Naselaris, T. & Gallant, J. L. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron* **79**, 1025–1034 (2013).
36. Marchette, S. A., Vass, L. K., Ryan, J. & Epstein, R. A. Outside Looking In: Landmark Generalization in the Human Navigational System. *Journal of Neuroscience* **35**, 14896–14908 (2015).
37. Silson, E. H. *et al.* A Posterior–Anterior Distinction between Scene Perception and Scene Construction in Human Medial Parietal Cortex. *J. Neurosci.* **39**, 705–717 (2019).
38. Carlson, T. A., Simmons, R. A., Kriegeskorte, N. & Slevc, L. R. The Emergence of Semantic Meaning in the Ventral Temporal Pathway. *Journal of Cognitive Neuroscience* **26**, 120–131 (2014).
39. Bankson, B. B., Hebart, M. N., Groen, I. I. A. & Baker, C. I. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage* **178**, 172–182 (2018).
40. Kacmajor, M. & Kelleher, J. D. Capturing and measuring thematic relatedness. *Lang Resources & Evaluation* (2019) doi:10.1007/s10579-019-09452-w.
41. Martin, A. The Representation of Object Concepts in the Brain. *Annu. Rev. Psychol.* **58**, 25–45 (2007).

42. Bonner, M. F., Price, A. R., Peelle, J. E. & Grossman, M. Semantics of the Visual Environment Encoded in Parahippocampal Cortex. *Journal of Cognitive Neuroscience* **28**, 361–378 (2016).
43. Simoncelli, E. P. & Olshausen, B. A. Natural Image Statistics and Neural Representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
44. Geisler, W. S. Visual Perception and the Statistical Properties of Natural Scenes. *Annu. Rev. Psychol.* **59**, 167–192 (2008).
45. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
46. Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nat Neurosci* **16**, 974–981 (2013).
47. Okazawa, G., Tajima, S. & Komatsu, H. Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc Natl Acad Sci USA* **112**, E351–E360 (2015).
48. Konkle, T. & Oliva, A. A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron* **74**, 1114–1124 (2012).
49. Julian, J. B., Ryan, J. & Epstein, R. A. Coding of Object Size and Object Category in Human Visual Cortex. *Cereb. Cortex* bhw150 (2016) doi:10.1093/cercor/bhw150.
50. Bainbridge, W. A. & Oliva, A. Interaction envelope: Local spatial representations of objects at all scales in scene-selective regions. *NeuroImage* **122**, 408–416 (2015).
51. Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C. & Tootell, R. B. H. The “Parahippocampal Place Area” Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. *PLoS Biol* **9**, e1000608 (2011).

52. Nasr, S. & Tootell, R. B. H. A Cardinal Orientation Bias in Scene-Selective Visual Cortex.
Journal of Neuroscience **32**, 14921–14926 (2012).
53. Nasr, S., Echavarria, C. E. & Tootell, R. B. H. Thinking Outside the Box: Rectilinear Shapes Selectively Activate Scene-Selective Cortex. *Journal of Neuroscience* **34**, 6721–6735 (2014).
54. Bryan, P. B., Julian, J. B. & Epstein, R. A. Rectilinear Edge Selectivity Is Insufficient to Explain the Category Selectivity of the Parahippocampal Place Area. *Front. Hum. Neurosci.* **10**, (2016).
55. Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc Natl Acad Sci USA* **115**, E9015–E9024 (2018).
56. Berman, D., Golomb, J. D. & Walther, D. B. Scene content is predominantly conveyed by high spatial frequencies in scene-selective visual cortex. *PLoS ONE* **12**, e0189828 (2017).
57. Bracci, S., Ritchie, J. B. & de Beeck, H. O. On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia* **105**, 153–164 (2017).
58. Chang, L. & Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013–1028.e14 (2017).
59. Botvinick, M., Weinstein, A., Solway, A. & Barto, A. Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences* **5**, 71–77 (2015).
60. Kaiser, D., Quek, G. L., Cichy, R. M. & Peelen, M. V. Object Vision in a Structured World. *Trends in Cognitive Sciences* **23**, 672–685 (2019).
61. Kaiser, D. & Cichy, R. M. Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology* **120**, 848–853 (2018).

62. Chan, A. W.-Y., Kravitz, D. J., Truong, S., Arizpe, J. & Baker, C. I. Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nat Neurosci* **13**, 417–418 (2010).
63. Kaiser, D. & Cichy, R. M. Typical visual-field locations facilitate access to awareness for everyday objects. *Cognition* **180**, 118–122 (2018).
64. Turk-Browne, N. B. The hippocampus as a visual area organized by space and time: A spatiotemporal similarity hypothesis. *Vision Research* **165**, 123–130 (2019).
65. Bruni, E., Tran, N. K. & Baroni, M. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research* **49**, 1–47 (2014).
66. Stojanoski, B. & Cusack, R. Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision* **14**, 6–6 (2014).
67. Haxby, J. V. *et al.* A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron* **72**, 404–416 (2011).
68. Tarhan, L. & Konkle, T. Reliability-based voxel selection. *NeuroImage* 116350 (2019) doi:10.1016/j.neuroimage.2019.116350.
69. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage* **60**, 2357–2364 (2012).
70. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]* (2017).

71. J. Deng *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).

doi:10.1109/CVPR.2009.5206848.