scientific reports

Check for updates

OPEN

Contextual associations represented both in neural networks and human behavior

Elissa M. Aminoff[1]✉, Shira Baror[1,2], Eric W. Roginek[3] & Daniel D. Leeds[3]

Contextual associations facilitate object recognition in human vision. However, the role of context in artificial vision remains elusive as does the characteristics that humans use to define context. We investigated whether contextually related objects (bicycle-helmet) are represented more similarly in convolutional neural networks (CNNs) used for image understanding than unrelated objects (bicycle-fork). Stimuli were of objects against a white background and consisted of a diverse set of contexts (N = 73). CNN representations of contextually related objects were more similar to one another than to unrelated objects across all CNN layers. Critically, the similarity found in CNNs correlated with human behavior across multiple experiments assessing contextual relatedness, emerging significant only in the later layers. The results demonstrate that context is inherently represented in CNNs as a result of object recognition training, and that the representation in the later layers of the network tap into the contextual regularities that predict human behavior.

Objects do not appear in isolation, but rather embedded within a context. The context of an object includes the regularities of the scene in which it is found, the cluster of other objects it is typically found with, and the spatial relationships between all of these components. These contextual relationships have repeatedly been shown to facilitate human cognition and perception[1–6]. For example, faster reaction times and more accurate responses in recognizing an object are found when the object is either primed by a contextual association (e.g., contextually related scene[7]), or when it is embedded in a congruent context compared with an incongruent context[1,2]. Thus, contextual associations are a strong cue for understanding our visual world and recognizing objects. However, what is the nature of these contextual associations and what is the relation between the contextual associations and object representations? For example, can contextual associations be extracted purely by exposure in the visual domain? And if so, are these types of contextual associations of objects learned specifically to enhance object recognition and incorporated in object perception even in the absence of visual background cues? One way to address the nature and role of contextual associations is to examine whether context is represented in artificial visual models.

Building on decades of work, especially in the most recent decade, computer vision has excelled to the level of human performance in recognizing objects[8–10]. This is largely due to the development of deep convolutional neural networks (CNNs) trained to identify objects in images. CNNs are trained on thousands to millions of images to recognize the statistical regularities that indicate an object's identity. However, it is unknown whether a CNN inherently learns and utilizes contextual associations to do this, even though it is not explicitly trained to do so. Moreover, if CNNs do learn contextual associations, it is unknown whether these contextual associations relate to the ones utilized in human perception.

Some computer vision work has developed models to integrate scene context into object perception explicitly. Several studies have incorporated context through object spatial position and camera pose, with small to modest improvements in object detection and recognition[11,12]. Bell et al. extracted context and object representations through distinct CNNs and merged these two representations to improve object classification nearly twofold, particularly excelling at detecting small objects[13]. These studies suggest that contextual information aids object recognition when context is visually apparent. However, whether context is inherently embedded within the object representation, even when not visually presented, remains an open question. Answering this question is important to understand the similarities and differences between human and computer vision. While multiple studies show that long-term contextual knowledge aids object recognition by human observers, even when objects are presented independent from context, the question remains whether context inherently facilitates

[1]Department of Psychology, Fordham University, Bronx, NY, USA. [2]Neuroscience Institute, New York University School of Medicine, New York, NY, USA. [3]Department of Computer and Information Sciences, Fordham University, Bronx, NY, USA. ✉email: eaminoff@fordham.edu

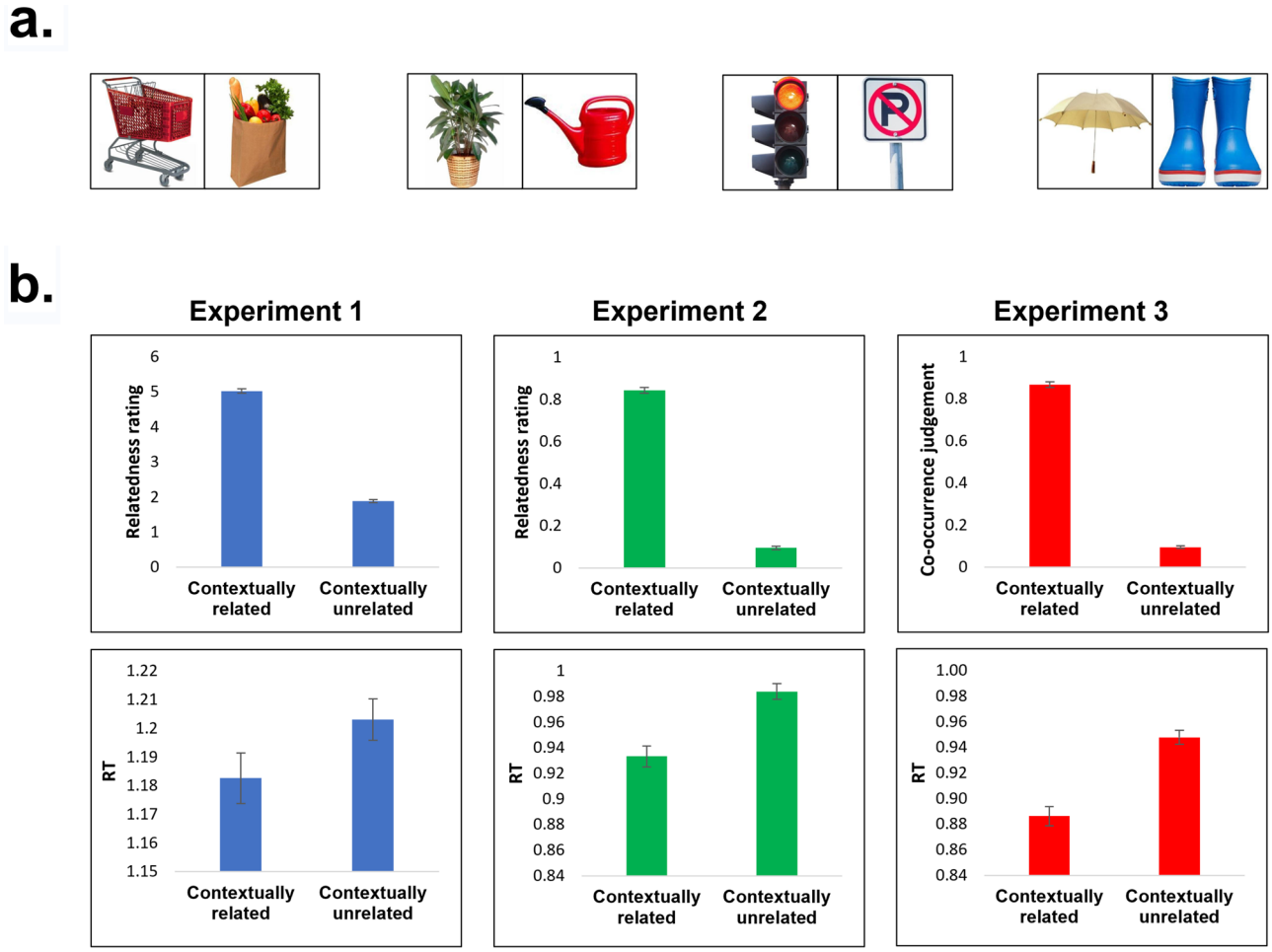

Figure 1. **(a)** Examples of the stimuli used. 73 contexts were chosen, each with 2 paired objects. All pictures were photographs of objects against a white background. **(b)** Behavioral validation results. Three experiments were used to asses relatedness across contextually related object pairs. These experiments differed based on the instructions (Experiment 1 and 2 asked about relatedness; Experiment 3 asked whether the two objects would be found in the same picture); and the response choice (Experiment 1 used a rating scale of 6 points; Experiments 2 and 3 used a two alternative forced choice), see Methods. Across these three experiments, contextually related objects were judged as related significantly more than unrelated objects (top row). This is reflected in reaction time as well, such that related judgements of contextually related objects are made significantly faster than judgements of contextual unrelated objects. Error bars represent standard error of the mean.

computer vision, without being explicitly modeled. And if so, whether context is covertly represented within the representation of objects in a CNN. Extraction of contextual associations during training of artificial object recognition algorithms would further support the functional utility of contextual associations for object recognition. It would also demonstrate that context can be defined through regular object and scene co-occurrences and provide a framework for predicting the influence and strength of specific contexts on object recognition.

The current study took a multi-part approach to investigating the role of contextual associations in CNNs and how this role relates to human perception. We focused on examining contextual associations between individual objects (e.g., table-chair) and asked whether these associations are represented in a CNN. To accomplish this, the object representation across the units at each layer of the CNN was first examined. We then looked at whether the representations of contextually related objects (bike—helmet) were more similar to one another than to objects that were unrelated (bike—fork). If so, this would indicate that contextual associations are inherently represented in the network even though the network was not explicitly trained to do so. In order to survey a wide range of potential contextual associations, 73 pairs of contextually related objects were tested (see Fig. 1a for examples). Critically, we extracted the CNN representation of objects depicted in images against a white background to prevent confounding the input stimulus with additional contextual information and/or scenery. To address our second question—whether the representations of contextual associations in a CNN were related to the representations that humans use, the contextual similarity represented in a CNN across the related pairs of objects was compared with human performance when rating whether these object pairs were associated with one another. A significant correlation between these measures would indicate the similarity depicted in the CNN was related to contextual associations used in human perception.

Results

Validation of stimuli. Contextual associations between objects were first assessed in three human behavioral experiments. The three experiments were treated as replications in demonstrating the effect across slight variations in instructions as well as independent, non-overlapping groups of participants. In these experiments, two images belonging to two object categories (e.g., an easel and a palette) were presented simultaneously and participants were asked to judge whether the objects were related. In all experiments, 73 pairs of contextually related object categories were assessed. To make sure judgements generalize beyond exemplar-specific attributes, for each object category (146 object categories, comprising the 73 pairs) five different exemplars of the object were employed. For example, five different pictures of hairdryers and five different pictures of barber chairs comprised the two object categories that as a pair form the barber context. In the experiments, each context was represented across four different trials: two trials in which both objects were contextually related; and two trials in which the objects were unrelated. Unrelated pairs consisted of swapping object categories from the contextually related pairs (e.g., hairdryer—bird). Relatedness was assessed using three different tasks while presenting two objects simultaneously, side by side: in Experiment 1 (N = 32), relatedness was assessed on a 6 point scale, such that each pair of objects was rated from unrelated (1) to related (6). In Experiment 2 (N = 20), a two-alternative-forced-choice was employed, asking participants to press the button (s) if the objects were related, and another button (d) if the objects were unrelated. In Experiment 3 (N = 20), participants pressed a button (s) if they expected to find the two objects in the same picture, and another button (f) if they did not. For all three tasks, participants were asked to respond as quickly and as accurately as they could.

Results validated that contextually related objects were indeed more strongly related to one another than unrelated objects (see Fig. 1b). In Experiment 1, contextually related object categories were rated as more related (mean 5.02) than unrelated object categories (mean 1.88; $t(72) = 44.48$, $p < 4.09 \times 10^{-54}$). Participants were also faster at rating the contextually related objects as related (mean 1.18 s) compared with the unrelated object categories (mean 1.2 s; $t(72) = -2.09$, $p < 0.04$). When examining whether these ratings were stable, we performed a split-half analysis and compared the ratings across each half of the participants. Ratings were significantly correlated suggesting a stable reflection of context ($r(71) = 0.691$, $p < 8.5 \times 10^{-9}$). These results replicated across the two additional experiments. In Experiment 2, when tasked with a two alternative forced choice, participants rated contextually related object categories faster and as more related (mean response: 84% related responses, reaction time (RT): 0.93 s) compared with contextually unrelated pairs of object categories (mean response: 9% related responses, RT: 0.98 s; paired response t-test: $t(72) = 47.22$, $p < 6.35 \times 10^{-56}$; paired RT t-test: $t(72) = -6.28$, $p < 2.20 \times 10^{-8}$). In Experiment 3, when asked whether the objects belonged in the same picture, related object categories were more predicted to appear in the same picture (mean response: 86% same picture) and were responded to faster (mean RT = 0.88 s) compared with unrelated objects (mean response: 9% same picture, mean RT: 0.94 s; paired response t-test: $t(72) = 55.59$, $p < 6.91 \times 10^{-61}$; paired RT t-test: $t(72) = -7.6$, $p < 8.29 \times 10^{-11}$). Thus, given our behavioral data we can confirm that humans do perceive the contextually related objects as strongly related to one another compared with the unrelated objects.

Neural network representation of context. Would a CNN also treat contextually related pairs of objects as related, even though it was not explicitly trained to do so? To test CNN context-integration, the representation of an object across units in a layer was compared to the representation of a contextually related object and to the representation of an unrelated object. Stronger similarity to the contextually related object compared to the unrelated object would indicate that contextual associations were included in the object's representation. We first tested context-integration with a popular benchmark CNN, VGG 16[10], trained on image recognition with the ImageNet dataset[14]. To provide a framework in which to interpret the results of the contextually related pairs, we first compared the representations of objects that belong to the same category. This comparison demonstrates similarity of representations in the CNN based on object category membership (see Fig. 2), as the network was explicitly trained to link together objects in the same category. As mentioned above, each object was depicted in five different exemplars. To look at categorical representation, the similarity of the CNN representations across the different exemplars of the same category was evaluated. Similarity is measured as the ratio of average pairwise similarity of stimuli within a category (bird 1 vs bird 2, bird 1 vs bird 3, etc.) versus the average similarity outside of the category (bird 1 vs palette 1, bird 3 vs palette 5), even contrasting stimuli from the same context but still different category (bird 1 vs birdhouse 1). Ratios above 1 would indicate that there is greater similarity across categorically related objects than unrelated objects; thus, we expected similarity ratios above 1. Results showed that objects that are categorically related to one another (e.g., two hairdryers) are represented more similarly in every layer of the CNN studied above the first layer (Fig. 2c). Ratios were consistently above 1.1 starting at layer 4 ($t(141) > 14.8$, $p < 2 \times 10^{-29}$) and grew larger with each subsequent layer (the output layer of the network was not included in the analysis). The maximum similarity was found in the last layers (ratios 2.23 and 3.05, in layers 25 and 27); this heightened similarity was expected since a more invariant representation of category is thought to be represented higher in the network. In addition, we asked whether the category similarity also included visual similarity. To assess this, we analyzed the similarity of pictures of objects within the same category compared to outside of the category in the pixel domain, as well as using Histogram of Gradient (HOG[15]) visual features. To this end, we did indeed find that pictures of objects within the same category (e.g., bird 1, bird 2, bird 3, etc.) were more visually similar than pictures of objects across categories ($t(141) > 8.81$, $p < 1 \times 10^{-14}$).

Next, in comparison to categorical relationships, contextual relationships were investigated. These relationships are not explicitly trained but are rather implicitly learned based on statistical regularities found in the training stimuli. We found contextually related pairs of objects were represented more similarly in the network (despite not looking like one another, e.g., lamp-chair) than unrelated pairs (e.g., lamp-stroller). Two contextually

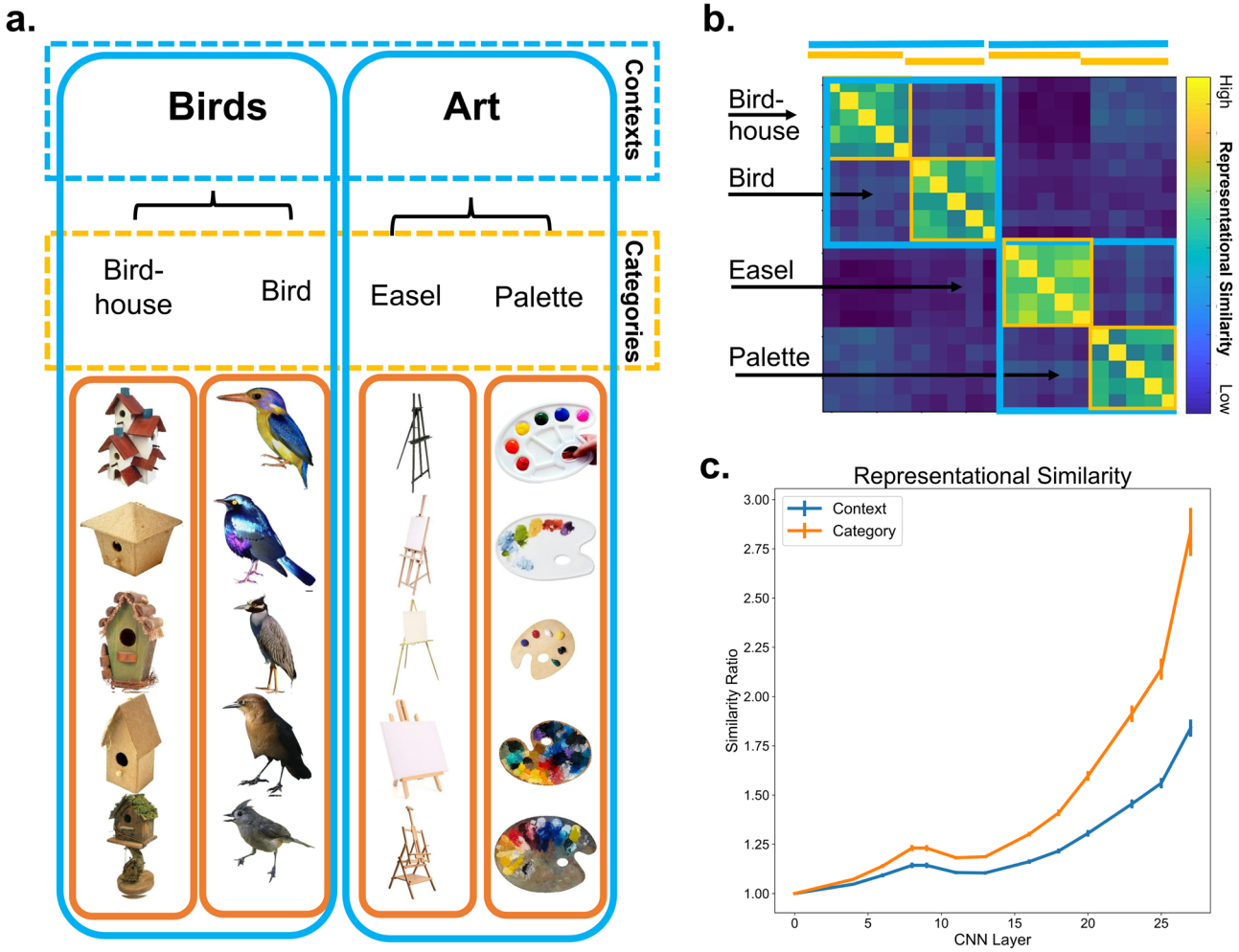

Figure 2. **(a)** Examples of the contrasts of interest. Five different exemplars of each object were used. Category similarity was assessed by comparing representations in the CNN across exemplars of the same category (orange outlines). Context similarity was assessed by comparing the representations of exemplars across the object paired categories (e.g., one easel and one palette exemplar, blue outlines). **(b)** Example of an ideal similarity matrix for each comparison using simulated data. **(c)** Representational similarity of images in the same category (orange) and same context (blue) at each layer of VGG16 CNN. Similarity computed as ratio of average in-group versus out-group similarities for each group. Error bars represent standard error of the mean.

related pairs had maximum contextual similarity ratios several magnitudes above the mean. These were considered outliers and were removed from subsequent analyses to prevent false inflation of the results. Thus, all analyses henceforth included 71 contexts. Surprisingly, the level of context-based similarity was significant at every layer studied above the first layer. Ratios were consistently above 1.01 and significantly above 1 starting at layer 4 ($t(70) > 8.5$, $p < 3 \times 10^{-12}$). The magnitude and significance of the ratio grew with each subsequent layer, reaching the maximum similarity in the last hidden layers of the network (ratios 1.62 and 2.01, in layers 25 and 27). Naturally, the degree of context similarity was less than the similarity exhibited for categorical relationships (Fig. 2c), as would be expected given the network is trained to provide explicit output identifying the object category. Nevertheless, although the network was trained to recognize individual object categories, the network also implicitly represented contextual associations between objects. In support of this analysis, we also assessed the visual similarity of object pairs using pixel similarity and HOG features to determine whether contextually related pairs were more visually similar than unrelated pairs. We found that contextually related pairs of objects were not significantly more visually similar than unrelated pairs of objects ($t(70) < 0.81$, n.s.).

Correlating contextual processing in human behavior and in neural network representation. The critical test in our analysis was to determine whether the representation of contextual associations in a CNN had any relation to the influence of contextual associations on human behavior. To address this, Pearson correlations between the behavioral performance and the similarity between the CNN representations were computed. The similarity measure used to represent context representations in the CNN were the ratio of the similarity between paired objects used in the behavioral experiment over the similarity of unrelated, unpaired objects. To do this, we used the maximum similarity ratio value extracted from the CNN. Comparing across all

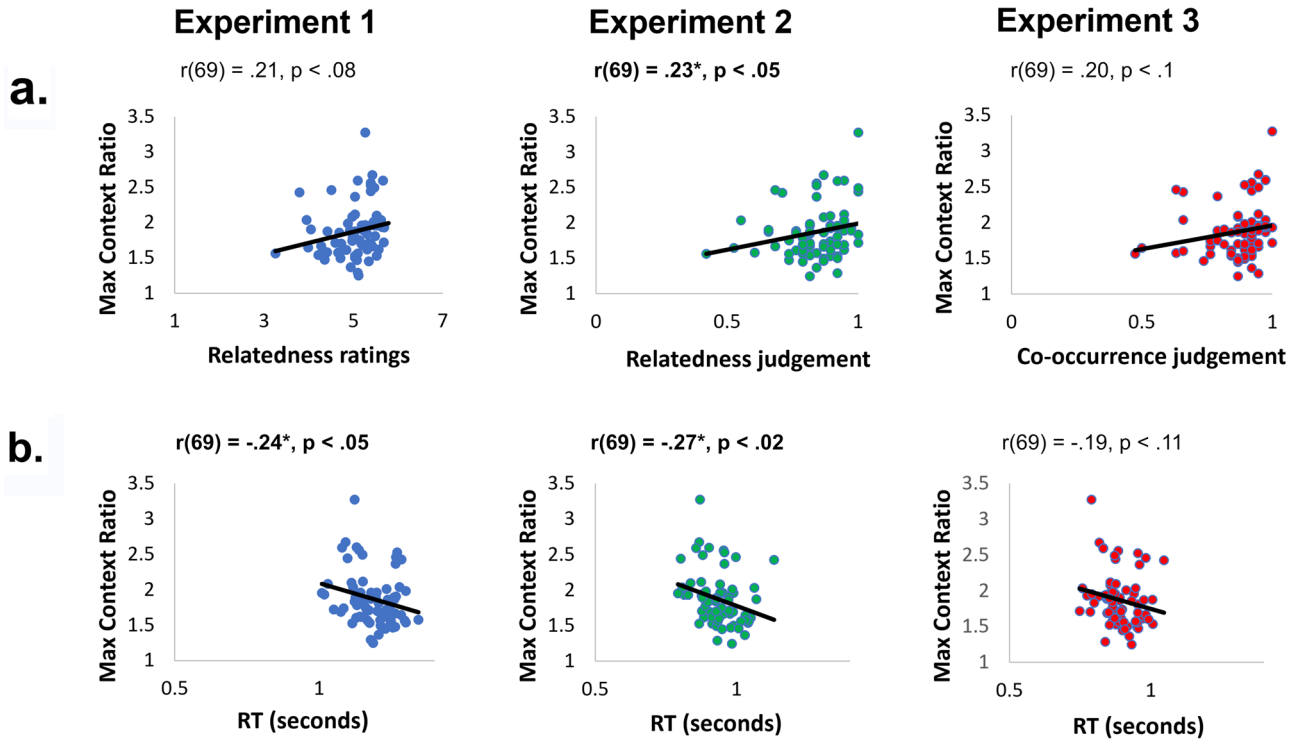

Figure 3. Correlations between the representation of context in VGG16 and contextual performance in humans. (**a**) Positive correlations between human ratings of contextual relatedness and context similarity in the CNN; significant in Experiment 2, marginal correlation in Experiment 1 & 3. (**b**) Negative correlations between human reaction time and context similarity in the CNN; significant in Experiment 1 & 2. Correlations are computed over 71 contexts.

network layers, the maximum ratio for the contextually related pairs of objects was found at the highest hidden layer (layer 27) in all but 1 of the 71 context pairs (the defiant context's maximum ratio was found at the penultimate layer, layer 25). The ratio used for the unrelated pairs was also extracted from the highest hidden layer to be consistent with values extracted for the related trials. This correlation was computed separately for each behavioral experiment. The purpose of looking at all three experiments was to reflect the ability to replicate the findings across tasks and independent sets of participants. We first ran this analysis examining the correlation across both contextually related and unrelated trials, with the expectation that the more contextually related the pairs of objects were judged, the more similar the object representations would be in the CNN. A robust positive correlation was indeed found across all three behavioral experiments (Ex. 1: $r(140) = 0.79$, $p < 2.5 \times 10^{-31}$; Ex. 2: $r(140) = 0.80$, $p < 3.9 \times 10^{-32}$; Ex. 3: $r(140) = 0.80$, $p < 2.02 \times 10^{-32}$).

However, to provide a more stringent test of the results, we examined this correlation only including the contextually related objects. Because rating responses to related and unrelated trials were significantly different, and largely non-overlapping, we were concerned the correlation would be driven by this distinction and wanted to investigate a more sensitive measure looking at the variability only with contextually related objects. In addition, we wanted to examine the correlation with an even more sensitive nuanced measure of contextual facilitation—reaction time—which we could not do when including the unrelated trials. When making predictions about reaction time and similarity, the prediction is different for related on unrelated pairs. For the related pairs, we expected faster reaction time the more similar the pair is in the CNN; however, for the unrelated pairs, we expected faster reaction times (e.g., faster at saying unrelated) the more dissimilar the pair is in the CNN. Thus, to hone in on the relationship between contextual associations used in human behavior, and those represented in a CNN we concentrated our investigation to only the related trials.

Specifically, behavioral performance, both relatedness judgements and reaction time, were correlated with contextual representations in the CNN (i.e., the maximum similarity ratio for the related objects). The resulting analysis found a correlation between the level of relatedness as indicated by human performance and similarity in the CNN network. This was significant in the second experiment ($r(69) = 0.23$, $p < 0.05$), but not significant in the first and third experiment (Ex. 1: $r(69) = 0.21$, $p < 0.08$; Ex. 3: $r(69) = 0.20$, $p < 0.1$). The positive direction of the correlation demonstrated the more participants found the pair of objects related to one another, the more similar the objects were represented across units in a CNN (Fig. 3a). Furthermore, to use a more stringent test, the more implicit and sensitive measure of reaction time was compared with the similarity of representation in a CNN. The resulting correlation demonstrated a significant negative correlation in experiments 1 and 2 (Ex. 1: $r(69) = -0.24$, $p < 0.05$; Ex. 2: $r(69) = -0.27$, $p < 0.02$). The negative correlation demonstrated that the faster participants were able to judge the relatedness of the objects, the more similar contextually related objects were

VGG16	Exp. 1 RT	Exp. 2. RT	Exp. 2 response
Layer 0	0.197	0.096	−0.158
Layer 4	−0.132	−0.004	0.048
Layer 6	−0.179	0.046	−0.035
Layer 8	−0.137	0.019	0.026
Layer 9	−0.137	0.019	0.026
Layer 11	−0.085	0.051	0.053
Layer 13	−0.076	0.04	0.037
Layer 16	−0.083	−0.027	0.036
Layer 18	−0.176	−0.174	0.18
Layer 20	**−0.253***	**−0.266***	0.195
Layer 23	**−0.266***	**−0.270***	0.201
Layer 25	**−0.261***	**−0.257***	0.208
Layer 27	**−0.237***	**−0.269***	**0.230**

Table 1. Pearson correlation between human behavioral performance and similarity in the VGG16 CNN network layers. Pearson r values between behavior (Experiment 1 reaction time, Experiment 2 reaction time, and Experiment 2 relatedness response) are listed in each cell. Cells marked with bold indicates correlations that were significant at $p < .05$, * indicates p values that withstood a false discovery rate correction of multiple comparisons across the layers. Significance of the correlation was only found in the later layers of VGG16.

represented in the CNN (Fig. 3b). This further supports the proposal that contextual associations are represented in a CNN, and those representations are related to human behavior.

The hierarchical nature of the layers (lowest layer 0, and highest hidden layer 27) in a CNN can also provide additional insight into understanding when the similarity of contextually related objects in a CNN is most relevant to human behavior. To find out which layers were most related to human behavior, we correlated behavioral responses of the contextually related trials with the similarity of the contextually related object representations in each extracted VGG 16 layer (N = 13). Correlation across the different layers was only computed for those comparisons that demonstrated a significant correlation with the maximum similarity ratio discussed above (Experiment 1 RT, Experiment 2 response and reaction time). Correlations were significant only in the later layers of the network (see Table 1). Significance between reaction time and similarity of CNN representations first appeared in layer 20 (Experiment 1 RT: $r(69) = -0.25$, $p < 0.03$, Experiment 2 RT: $r(69) = -0.27$, $p < 0.03$) and continued to be significant through layer 27 (Experiment 1 RT: $r(69) = -0.24$, $p < 0.05$, Experiment 2 RT: $r(69) = -0.27$, $p < 0.02$). Significance between relatedness responses and similarity of CNN representations was only significant at the last hidden layer, layer 27 (Experiment 2 relatedness response: $r(69) = 0.23$, $p < 0.05$), however this did not withstand correction for multiple comparisons across the layers.

This analysis revealed that although contextually related objects have a significantly similar representation across almost all layers of the CNN, only those representations in the later layers are related to the contextual associations that humans use.

Exploratory analysis of context representation in other CNNs. Now that we established that contextual associations are represented in one CNN, we investigated whether these associations were unique to the VGG16 network, or whether contextual associations were represented in a variety of CNNs. We therefore studied CNNs that varied by number of layers, by architecture (computational components for image classification and learning), and by whether they were trained on ImageNet (i.e., object based) or Places365[16] (i.e., scene based). Similar to the procedure carried out for VGG16, for each network, in−group versus out−group similarity ratios were measured across 71 Contexts and 142 Categories at the maximum ratio layer, comparing these ratios to the null hypothesis (ratio = 1, Fig. 4). The results show that the representation of contextual associations exists significantly in each network studied (p's $< 2 \times 10^{-23}$). For full results, please see Supplemental Materials Fig. 1. Some trends were found, however, differentiating between networks. More traditional, or shallow, networks showed substantially higher in−out ratios for both category and context than did networks with more novel architectures, which are considerably deeper in layers, ($t(70) = 21.4$, $p < 3 \times 10^{-32}$), suggesting simpler representations are more effectively fit to object and context properties. ImageNet−trained networks showed higher in−out context ratios than did Places365-trained networks ($t(70) = 23.6$, $p < 6 \times 10^{-35}$), suggesting that the representational similarity of objects from the same context is more effectively captured when learning to distinguish objects rather than when learning to distinguish overall scenes. Or alternatively, that context may be facilitative in processes related to object recognition, rather than scene categorization.

Discussion

This study used a CNN framework to examine whether contextual associations between objects are inherently extracted while learning to perform object classification, and whether the CNN-learned representational space for objects and contexts was related to human behavior. Our findings revealed that objects that were contextually related to one another were more similarly represented at each layer of the CNN compared to unrelated counter

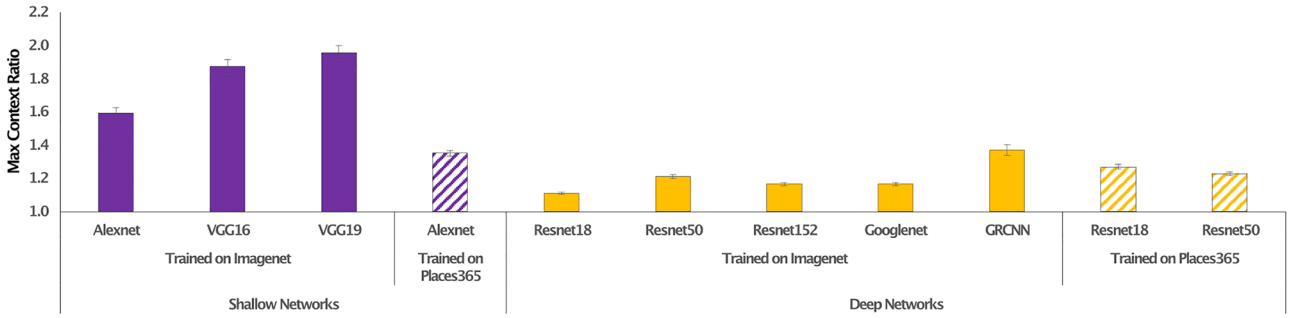

Figure 4. Contextual representations across a variety of networks. Maximum similarity ratio for ten CNN architectures trained on ImageNet (solid bars) and Place365 (striped bars) data sets. Each network tested demonstrated a significant effect of context (p's $< 2 \times 10^{-23}$). This effect is significantly greater in traditional (shallow) network architectures (purple bars) compared with more novel, deeper, network architectures (yellow bars).

objects. This indicates that commonalities related to context are integrated in the object CNN representation. Thus, despite being dissimilar in their visual features (e.g., umbrella and rainboots) and despite being presented on a white background with no visual context depicted, the network recognized a commonality between contextually related pairs of objects (bike helmet—bike) compared with unrelated objects (e.g., bike—football). This finding reveals the novel result that contextual associations are inherently coded in a CNN that is designed to recognize objects, despite it not being explicitly trained to capture context. Furthermore, we found that the extent to which the representations of contextually related objects were similar in the CNN correlated with human performance, both with how fast human participants responded in judging whether these object pairs were related, as well as with the degree of relatedness judgment itself.

Our study focused on the VGG16 network, a traditional convolutional neural network trained on ImageNet with the purpose of recognizing objects within an image, ultimately identifying each image as one of a possible 1000 objects. The network's successful optimization for object categorization is evident in the high similarity in the representation of objects that share the same category (e.g., two different birdhouses) compared with objects that do not share the same category. The significant similarity between objects of the same category was evident across all studied layers of the network and, as expected, increased with each subsequent layer. This analysis of category-based similarity was used as control and validated the use of this parameter as evaluating the representations encapsulated in the network. We then used it as a benchmark to compare the representation of context in the network, and found that belonging to a shared context, not only to a shared category, increased object representational similarity in the network as well. Analysis using pixel and HOG features further underscores the lack of basic visual similarities between different categories in the same context, indicating the VGG16 network learns additional co-occurrences of context-related categories within the training data without being explicitly trained to do so.

Furthermore, in our investigation of context it was critical to present the objects against a white background in order to prevent confounding the presented stimulus with any visual contextual information (distinguishing from other recent related work[17]). Even though objects were presented in isolation, the similarity related to a shared context was evident in every layer of the network after the first layer, increasing with each subsequent layer. As later layers are more tuned to the ultimate goal of object categorization[18], it is very possible that achieving the goal of object categorization relies on contextual information with every additional layer. As expected, due to the targeted task of the network, the similarity based on object category was significantly higher than the similarity based on context. However, our results suggest regularities associated with context were correlated with object categories, and thus were integrated into the object representations to facilitate category recognition. This relation was true not only for VGG16, but for each of the ten other CNNs that were studied, trained on both ImageNet and Places365. Thus, networks using both traditional or shallow and more-novel, deeper, architectures, including residual and recurrent networks, inherently integrated context within object representations. Interestingly, those networks trained to recognize scene categories (i.e., those trained on Places365) did not have significantly stronger context representations, suggesting that context is important for scene categories, but also critical to the task of object recognition itself. Previous studies, as highlighted in the introduction, were able to improve the network's performance by explicitly modeling context. Our results are novel in showing that even without explicit training, context is inherently captured in the network. Context, therefore, is not just evident in facilitating human perception, but also inherently facilitates artificial perception in computer vision.

The degree of similarity in the CNN representations of contextually related objects correlated with human behavior both with regards to rating the degree to which two objects are related, as well as with how quickly these ratings were made, suggesting that a CNN can be used to predict context effects on human behavior, such as the strength and weight a particular context can have on facilitating object recognition. More generally, this also supports that the context represented in a CNN is relevant to understanding and modeling how context is derived in human behavior. Context facilitation of human perception is typically framed as utilizing prior knowledge to generate predictions and expectations on our perception in a top-down manner. In a CNN, however, the computations that are applied to determine the representation at each layer of the network are driven by bottom-up statistical regularities found in the training images. The results of the current study support the idea

that contextual relations between objects are derived bottom-up from the statistical regularities of co-occurrence, which is supported by previous neuroimaging results[19]. Thus, the correlation to human behavior supports a model that signifies at least some of the contextual associations that humans utilize in perception derive from the statistical regularities one experiences in the environment.

Interestingly, the significant correlations between CNN and human behavior were only found in the later layers of the network, even though context was significantly represented at almost each layer of the CNN. Higher network layers are typically associated with less simple visual properties and more with higher-level semantic properties[20,21], indicating the learned regularities relevant to human behavior are more likely to be at a conceptual level rather than a visual level. Thus, relying just on statistical regularities to explain context in human cognition is too simplistic. Moreover, running a visual similarity analysis helped verify that low-level similarities between objects in the contextually related pairs does not account for our results. Further research is needed to uncover the differences of context representation at low levels versus high levels of a CNN. This will provide more insight into how context is derived, utilized, and updated in humans.

Naturally, unlike humans who experience the world in a multi-sensory way, CNNs can only learn from the visual information fed to them. Thus, although the correlations between human behavior and CNN were significant, the shared variance between the systems is relatively small, hinting that there are yet many differences between human and artificial vision. This is not surprising, given that humans interact with the perceptual environment in a much richer way than a CNN, acquiring inter-object associations that go beyond the visual domain (e.g., action-related, temporal), which are unlikely to be represented in a network trained for a single purpose of object recognition. However, the small effect size may speak more to the small, but significant, role visual statistical co-occurrences play in contextual associations between objects in humans and helps paint the picture of how we can define contextual associations.

The results of this study demonstrate the importance of taking into account contextual associations in models of object recognition and image understanding in both human and artificial vision. And indeed, recent work incorporates context more readily in the deep networks used for image understanding. For example, scene graphs have played an important role in recent computer vision approaches to modeling static and dynamic scenes. Co-occurring objects in a given image are recognized and characterized through their relations with one-another[22–25]. However, patterns of inter-object association are not explicitly tied to the underlying scene in each image, and scenes can vary with inter-object association. These models that incorporate a cluster of contextual information are a promising avenue for further work in biological and computational studies as they demonstrated the strong role for context in modeling vision. Further development and utilization of convolutional neural networks that employ feedback from higher layers to lower layers during classification may further aid in the discovery and incorporation of context during object recognition. In contrast, most current recurrent network models focus only on feedback of each layer to itself, failing to obtain potential benefits of top-down processing, as observed by the Wang and Hu model's inferior contextual groupings compared to even a simpler feedforward network model[26].

Computer vision can achieve human level performance in recognizing objects of a scene (e.g., [27]), however, there is still a gap between how humans understand an image and how computer vision understands an image. This difference may be fueled by a divergence in learning experiences—humans assemble visual and contextual knowledge across a lifetime of linearly evolving experiences that build off one another, while CNNs typically train on a static training set of images mixed together in less determined order and more limited in overall diversity. It is possible that further context learning may be key to bridging the gap between the biological and artificial systems, and that the more neural networks utilize context in ways similar to those of humans, for example, relying more on the context representation at later layers, the more computer vision may understand unusual or unique images in the same way humans effortlessly do. Ultimately, the more we bridge the gaps between human and artificial vision, the more computer vision can be applied to aiding and working in concert with human vision, for example supplementing and aiding people with visual impairments.

In conclusion, our study shows that context is inherently encapsulated in neural networks without being explicitly trained to do so, and that this representation of context correlates with human perception. Thus, object recognition trained CNNs represent context, emphasizing those contexts with the most impact on human cognition. Understanding the shared and unique ways artificial and human systems utilize context is therefore a promising direction in enhancing performance in both realms.

Methods

All methods were carried out in accordance with relevant guidelines and regulations.

Stimuli: Stimuli used in this study were photographs of objects against a white background. The objects were selected with a white background to remove confounding background variables and standardize background luminance. Images were obtained from a dataset by Brady et al.[28], as well as from google image search. Stimuli were 375×375 pixels. A total of 730 images were used in this experiment. This was composed of 146 object categories, where each object had five different exemplars (e.g., five different tractors). The 146 object categories were then grouped into pairs of objects that belong to the same context, of which there were 73 contexts. Two contexts, and their corresponding four object categories, were removed from analysis due to their unusually high context and category similarity ratios in VGG16 CNN analysis—over three standard deviations above the mean. The remaining 71 contextually related pairs of objects were used in all remaining CNN related analyses.

Human behavioral experiments. *Participants.* Participants were recruited online via Prolific (https://www.prolific.co/). Participants self-reported they had normal or corrected to normal vision, fluent in English, and were located in the USA. Participants were financially compensated for their time. All study procedures

were approved by the Institutional Review Board of Fordham University. Informed consent was obtained from all participants. In Experiment 1, there were a total of 34 participants (17 females, mean age 36.67, 21–68 range). Two participants were excluded from analysis for not performing the task (did not press any key); in Experiment 2 there were a total of 20 participants (10 females, mean age 34.9, 20–58 range); and in Experiment 3 there were a total of 22 participants (8 females, mean age 28.14, 18–50 age range). Two participants were excluded from analysis in this experiment for not performing the task (one did not press any key and the other constantly pressed the same key).

Procedure. All experiments were presented using PsychoPy software[29] and hosted through the Pavlovia website (https://pavlovia.org). Participants were only permitted to participate in the experiment from a desktop/laptop computer (i.e., no mobile devices).

In all three experiments, a trial began with a fixation cross for 100 ms that remained on the screen for the entire trial. Afterwards, two pictures of objects were presented side by side until the participant responded, or up to 1000 ms. Participants had up to 3 s to make a response. Before the participants began the experiment, they were given 16 practice trials. Participants were asked to respond as quickly and as accurately as they could.

The pair of objects presented were either of the same context or of different contexts. The experiment involved a total of 292 trials, which consisted of four presentations of each object (using four different exemplars). Two trials of each object were presented with a contextually related object, and two trials were presented with a contextually unrelated object (swapped from other contexts). Thus, half of the trials depicted contextually related objects (with two trials per context), and half of which depicted contextually unrelated objects. Specific exemplars of objects were balanced across the conditions across participants.

In Experiment 1, participants were asked to rate how related they found the pair of objects. They responded using a 6-point scale from 1: very dissimilar contexts to 6: very similar contexts. The scale was present on the screen during the duration of the experiment and participants responded using keys 1–6.

In Experiment 2, we wanted to use a paradigm that more accurately reflected reaction time differences across the trials. To accomplish this, participants were asked to make a two alternative forced choice and judge whether the two objects were of the same context (key s) or were of different contexts (key d). Instructions were displayed on the screen for the duration of the experiment. Experiment 2 also included ten catch trials in which two identical objects were presented and the participant had to respond that they were of the same context. This was to increase the quality of data collection.

In Experiment 3, we wanted to use a task more closely related to what a CNN might be picking up on—those objects appear together in the same scene. To accomplish this, in this last study we asked participants to judge whether the two objects would be found in the same photograph ('s' for same; 'd' for different). Instructions were displayed on the screen for the duration of the experiment. Like Experiment 2, catch trials were also included in this experiment to increase the quality of data collection.

Analysis. In all three experiments, we averaged participants' responses and reaction time for each context (e.g., bike riding) in the related trials (e.g., bike-helmet) and in the unrelated trials (bike-fork). We then ran paired t-test analyses in all experiments to examine whether there were significant differences between related and unrelated responses, both in terms of the relatedness response (depending on the experiment and task) and its reaction time (RT).

To examine the stability of the ratings responses in Experiment 1, we ran a split-half analysis by breaking the participants into two groups and correlating the averaged ratings across the contexts. To break the participants into two groups, a random ordering was determined and then the participants were split in half (15 participants in one group; 16 in the other). The ratings for each group were averaged, and then the ratings of the two groups were correlated. This was done 1000 times, each time with a random ordering of participants, and the averaged r value and p values were reported.

Neural network analysis. We focused our study on the VGG16 convolutional neural network, used for its high performance in computer vision and simplicity as an analog to biological neural networks[10]. We focused on the network as pretrained on the ImageNet data set[14,30], selected for its wide usage in computer vision due to its large size and its diversity of object classes. Network implementation and analysis was conducted using the Python PyTorch library[31].

The VGG16 network consists of 31 total Layers, each with 64 to 512 units to represent the image; the top layer (layer 31) contains 1000 units, corresponding to the 1000 object categories in ImageNet. Unit responses were extracted at the layers at the end of each processing block of the CNN architecture, where each block begins with a convolution and ends with rectification or max pooling. For VGG16, we studied thirteen layers, specifically: layers 0, 4, 6, 8, 9, 11, 13, 16, 18, 20, 23, 25, and 27. At each layer, representational similarity of image pairs were computed using Pearson's Correlation of unit responses.

Similarity ratios were computed for each category and context. Specifically for categories, similarities were pooled among image pairs within each category and among image pairs with one image in-category and one image out-of-category. These relative similarities were measured by $SimRatio^C$, computed as follows:

$$SimRatio^C = \frac{MeanInSim^C}{MeanOutSim^C} = \frac{\frac{1}{N^C_{inGroup}} \sum_{(i,j)\in C, i\neq j} sim(p_i, p_j)}{\frac{1}{N^C_{outGroup}} \sum_{i\in C, j\in C''} sim(p_i, p_j)}$$

p_i, p_j are two images to be compared; in the numerator we consider every pair of distinct images in the category C, i.e., $(i,j) \in C, i \neq j$; in the denominator, one image is in the category and the other is outside the category, i.e., $i \in C, j \in C'$. For several categories C, a set of confounds were removed from the corresponding outside set C', designated for those objects that share a super ordinate category, e.g., a bike helmet was not considered in the "out" group for football helmet. In both the numerator and denominator, the average similarity ratio is computed by dividing the summed ratio by the total number of image pairs in the summation, $N^C_{inGroup}$ and $N^C_{outGroup}$. This ratio $SimRatio^C$ was computed for each layer to measure the evolution of category representations through the network layers.

Along the same lines, similarities were pooled among image pairs within each context and among image pairs with one image in-context and one image out-of-context. These relative similarities were measured also by $SimRatio^C$ as computed above. Now, C denotes a context rather than a category. Again, a set of confound contexts are removed from C' when computing the denominator. For example, there were multiple object pairs from the kitchen context (e.g., pot—oven mitt; oven—fridge) that would not be considered in the "out" category for one another. For the context ratio, the in-context pairs considered in the numerator exclude all pairs in the same category. For both category and context comparisons, a ratio of 1 indicates no difference in pictures inside and outside the group.

For example, we can consider five easel images representing the "easel" category, five palette images representing the "palette" category, and both categories representing the "art" context. Our category ratio for easel is obtained by dividing the average similarity of distinct easel pictures in-category by the average similarity of easel pictures to non-easel pictures. Our context ratio for art is obtained by dividing the average similarity of easel pictures to palette pictures by the average similarity of easel-or-palette pictures to any other picture in our data set.

Two out of the 73 contexts revealed a context ratio that was several orders of magnitude above the ratios of the other categories. These were removed from further analysis, to assure that these two categories were not inflating the results. Thus, all following analyses include 71 contexts.

To assess visual similarity, we repeated our category and context representation analyses comparing pairs of pictures based on Euclidean distances between their pixel representations and Euclidean distances between their Histogram of Gradient (HOG) descriptors[15].

We further repeated our category and context representation analyses on several additional CNN models. We studied additional traditional architectures, VGG 19 and Alexnet, which use the same computational building blocks as VGG16 arranged in different quantities and orders[8,10]; we also studied more novel network architectures designed to enable learning on deeper networks using "skip connections" for residual learning in Resnet50 and Resnet152, and "inception blocks" in GoogLeNet, all pretrained for object recognition using ImageNet[27,32]. We further studied a recent recurrent network architecture (GRCNN[26], in which units in each layer incorporated their own previous activity in computing their current output, intended to model lateral connections in biological networks and to simulate deeper network architectures while maintaining a relatively small number of learned network parameters. This recurrent network also was trained on ImageNet. We also studied several architectures pretrained for scene recognition using Places365[16]. For each network, the layer containing the maximum similarity ratio for category and context was studied. (In each network, the same layer had the highest ratio for both context and category.) Traditional networks (VGG and AlexNet) were compared to more novel network architectures; and ImageNet-trained networks were compared to Places365-trained networks through T-tests on the average maximum similarity ratio for each of the 142 categories and for each of the 71 contexts.

Human behavior—CNN correlations. To examine whether the representation of contextual information in CNNs related to human behavior, we computed Pearson correlation between the behavioral results and the CNN results. Initially ran on all trials, both the related and unrelated trials, comparing relatedness ratings and context similarity in the CNN. Subsequent analyses focused only on those trials in which the two objects were related. Two behavioral measures were used when analyzing just the related trials: the relatedness response and the RT for each of the 71 contexts. These behavioral measures were then correlated with the context similarity ratio in the VGG16 CNN network. First, we correlated behavioral performance with the maximum context ratio found at any layer (typically found at layer 27 for all but one context, which had the maximum ratio at layer 25). We then also examined the relationship between behavior and the representation of each layer of VGG16, and correlated with behavior with the context similarity ratio extracted for each layer of VGG16. The correlations were computed separately for each behavioral experiment.

Correcting for multiple comparisons. All p values from the analyses that included a test for each layer of the network were assessed for significance using a false discovery rate correction of multiple comparisons across the layers. If the significance of the p value did not withstand the correction, it was noted in the text.

Received: 29 November 2021; Accepted: 21 March 2022
Published online: 02 April 2022

References

1. Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: Detecting and judging objects undergoing relational violations. *Cognit. Psychol.* **14**, 143–177 (1982).
2. Davenport, J. L. & Potter, M. C. Scene consistency in object and background perception. *Psychol. Sci.* **15**, 559–564 (2004).

3. Koehler, K. & Eckstein, M. P. Scene inversion slows the rejection of false positives through saccade exploration during search. *Proc. Annu. Meet. Cogn. Sci. Soc.* **6**, 1 (2015).
4. Lauer, T., Willenbockel, V., Maffongelli, L. & Võ, M.L.-H. The influence of scene and object orientation on the scene consistency effect. *Behav. Brain Res.* **394**, 112812 (2020).
5. Mudrik, L., Lamy, D. & Deouell, L. Y. ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia* **48**, 507–517 (2010).
6. Welbourne, L. E., Jonnalagadda, A., Giesbrecht, B. & Eckstein, M. P. The transverse occipital sulcus and intraparietal sulcus show neural selectivity to object-scene size relationships. *Commun. Biol.* **4**, 768 (2021).
7. Palmer, S. E. The effects of contextual scenes on the identification of objects. *Mem. Cognit.* **3**, 519–526 (1975).
8. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
9. Rosenblatt, F. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms.* (Cornell Aeronautical Lab Inc, 1961).
10. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:14091556 Cs (2014).
11. Beery, S., Wu, G., Rathod, V., Votel, R. & Huang, J. Context R-CNN: Long term temporal context for per-camera object detection. in *Proceedings of the IEEECVF Conference on Computer Vision on Pattern Recognition.* 13075–13085 (2020).
12. Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. & Hebert, M. An empirical study of context in object detection. in *IEEE Conference on Computer Vision on Pattern Recognition.* 1271–1278 (2009).
13. Bell, S., Zitnick, C. L., Bala, K. & Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2874–2883. https://doi.org/10.1109/CVPR.2016.314. (IEEE, 2016).
14. Deng, J. *et al*. ImageNet: A large-scale hierarchical image database. in *IEEE Conference on Computer Vision Pattern Recognition*. 248–255. (2009).
15. Dalal, N. & Triggs, B. Histograms of oriented gradients for human detection. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 886–893. (IEEE, 2005).
16. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 1452–1464 (2018).
17. Bracci, S., Mraz, J., Zeman, A., Leys, G. & de Beeck, H. O. *Object-Scene Conceptual Regularities Reveal Fundamental Differences Between Biological and Artificial Object Vision*. http://biorxiv.org/lookup/doi/https://doi.org/10.1101/2021.08.13.456197 (2021).
18. Rafegas, I., Vanrell, M., Alexandre, L. A. & Arias, G. Understanding trained CNNs by indexing neuron selectivity. *Pattern Recognit. Lett.* **136**, 318–325 (2020).
19. Aminoff, E. M. & Tarr, M. J. Associative processing is inherent in scene perception. *PLoS ONE* **10**, e0128840 (2015).
20. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3319–3327. https://doi.org/10.1109/CVPR.2017.354 (IEEE, 2017).
21. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. in *Computer Vision—ECCV 2014* (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.). Vol. 8689. 818–833. (Springer, 2014).
22. Ost, J., Mannan, F., Thuerey, N., Knodt, J. & Heide, F. Neural scene graphs for dynamic scenes. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2855–2864. https://doi.org/10.1109/CVPR46437.2021.00288 (IEEE, 2021).
23. Xu, D., Zhu, Y., Choy, C. B. & Fei-Fei, L. Scene graph generation by iterative message passing. in *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*. 5410–5419. (2017).
24. Yang, J., Lu, J., Lee, S., Batra, D. & Parikh, D. Graph R-CNN for scene graph generation. in *Computer Vision—ECCV 2018* (eds. Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.). Vol. 11205. 690–706. (Springer, 2018).
25. Zhang, L., Xu, D., Arnab, A. & Torr, P. H. S. Dynamic graph message passing networks. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3723–3732. https://doi.org/10.1109/CVPR42600.2020.00378 (IEEE, 2020)..
26. Wang, J. & Hu, X. Convolutional neural networks with gated recurrent connections. *IEEE Trans. Pattern Anal. Mach. Intell.* https://doi.org/10.1109/TPAMI.2021.3054614 (2021).
27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (2016).
28. Brady, T. F., Konkle, T., Alvarez, G. A. & Oliva, A. Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci.* **105**, 14325–14329 (2008).
29. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).
30. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
31. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **12**, 32 (2019).
32. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594 (IEEE, 2015).

Acknowledgements

This work was supported by an interdisciplinary grant from the Office of Research at Fordham University.

Author contributions

E.A. designed the project. E.A. & S.B. designed the behavioral studies, collected the behavioral data, and analyzed the data. D.L. & E.R. designed and executed the CNN analysis. E.A., S.B., & D.L. wrote the paper. E.R. contributed to writing the CNN methods. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09451-y.

Correspondence and requests for materials should be addressed to E.M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.