Contextual Information Representation of Objects-Scenes in Deep CNNs: Effects of Training and Architectures

Rahul Ohlan (rohlan@gradcenter.cuny.edu)
PhD Computer Science, The Graduate Center, City University of New York
365 5th Avenue, NYC, NY 10016 U.S.A

Daniel D. Leeds (dleeds@fordham.edu)
Fordham CIS, 441 East Fordham Road
Bronx, NY 10458 U.S.A

Elissa M. Aminoff (eaminoff@fordham.edu)
Department of Psychology, Fordham University
441 East Fordham Road
Bronx, NY 10458 U.S.A

Abstract

This research provides a comprehensive analysis of contextual information representation in state-of-the-art convolutional neural networks (CNNs) trained on ImageNet and Places365 datasets for object and scene recognition tasks. While current CNN models excel at object detection and image classification, our study investigates how these networks capture relationships between objects and scenes. We demonstrate that objects within related scenes exhibit closer contextual associations compared to those in unrelated contexts. Moreover, we investigate the effects of training and different CNN architectures on this relationship, providing insights into the nuanced representation of contextual information in deep learning-based computer vision systems. Dataset collecting open-source images spanning 50 diverse contexts with each comprising of objects and related scenes images observed in that context was used in this analysis.

Keywords: Scene, Object-Recognition, Deep networks, Vision

Background

This paper explores how deep convolutional neural networks (CNNs) represent objects in scenes and whether single objects are represented similarly to the scenes in which they are typically found. In human vision, context facilitates object recognition and understanding (Biederman, Mezzanotte, & Rabinowitz, 1982). We explore whether context is also inherent within CNN processing of objects. Critically, we examine how training may affect object representations within the CNN and whether those representations include information about the context in which the objects are typically found. Previous research has shown that objects in similar contexts are represented more similarly within a CNN, suggesting CNNs do extract the contextual information of objects (Aminoff, Baror, Roginek, & Leeds, 2022), (Bracci, Mraz, Zeman, Leys, & Op de Beeck, 2023). Our study employs clustering of network responses to analyze object and scene representations after training on different datasets and tasks.

Methods

Stimuli

The dataset included RGB images of objects against a white background and pictures of complete scenes. 50 contexts were used to explore the relation between object and scene representations. Each context included images of a scene and two objects that commonly occur in the scene. Importantly, all scene images omitted the associated two objects. For example, a living room picture without a television, if television was an associated object. A total of 600 images obtained from the BOLD5000 dataset (which includes pictures from ImageNet, COCO, and the SuNS image dataset (Chang et al., 2019)) and from Google Image Search were used. The 600 pictures depicted 50 unique contexts, each featuring an object or the related scene with 4 distinct exemplars. We used the second object of each context to construct a second dataset to replicate findings from the analysis run on the first object. Figure 1 shows stimulus examples.

Architectures and Pretraining

The study primarily utilized AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) and ResNet (He, Zhang, Ren, & Sun, 2016) architectures (18 and 50 layers; represented in pytorch as 11 and 8 layers, each with multiple sub-layers) pre-trained on ImageNet (Deng et al., 2009) and Places365 (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017) datasets. Pearson's correlation computed image representation similarity for each layer for all images. The values of interest were the similarity of the network representation of the full scene, in comparison to the related key object across the layers.

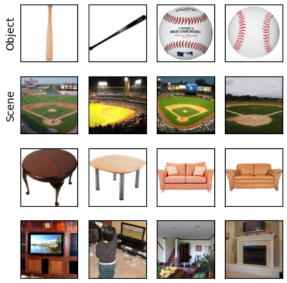


Figure 1: Sample images from Objects-Scenes dataset, including 4 images of objects (2 images from each of 2 object groups) and 4 images of related scenes in each of 2 contexts (labeled vertically in each row).

To assess this similarity, we used a ratio of within context similarities to out of context similarities. The out of context similarities were established using the correlations of the objects with unrelated scenes, e.g., a steering wheel and a kitchen. In-out ratios quantified representation evolution across layers, training, and network architectures.

$$In - Out\ Ratio = \frac{\frac{1}{N_{in-context}} \sum_{(i,j)\ \varepsilon\ C,\ i \neq j} sim(O_i,\ S_j)}{\frac{1}{N_{out-context}} \sum_{(i,j)\ \varepsilon\ C,\ j \varepsilon C'} sim(O_i,\ S_j)} \quad (1)$$

Results

A repeated measures ANOVA assessed how similarity between an individual object and a scene was modulated by architecture, training, or layer. The values at three different layers in each of the networks were extracted to represent low (layer 1), middle (layers 6 and 5) and top layers (layers 11 and 8) in AlexNet and ResNet architectures, respectively. Any ratio values across the 50 different contexts within a layer that were three times the standard deviation of the mean were considered outliers and removed from the analysis.

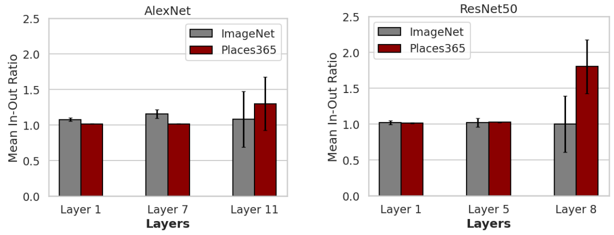
The first network comparison was between AlexNet and ResNet50. When examining the similarity between object 1

Table 1: Comparison between AlexNet and ResNet50

	Object 1				Object 2			
	df	F	Sig	Partial Eta Squared	df	F	Sig	Partial Eta Squared
Network	(1,41)	0.119	n.s.	0.003	(1,38)	0	n.s.	0
Training	(1,41)	5.385	0.025	0.116	(1,38)	0.64	n.s.	0.017
Layer	(2,82)	2.002	n.s.	0.047	(2,76)	1.369	n.s.	0.035
Network x Training	(1,41)	6.149	0.017	0.13	(1,38)	3.917	0.055	0.093
Network x Layer	(2,82)	0.558	n.s.	0.013	(2,76)	0.068	n.s.	0.002
Training x Layer	(2,82)	8.285	0.006	0.168	(2,76)	1.08	n.s.	0.028
Network x Training x Layer	(2,82)	2.078	n.s.	0.048	(2,76)	1.811	n.s.	0.045

and the related scene, we did not find a main effect of network or layer, however there was a significant effect of training ($p < .025$), Table 1. There was also a significant interaction of network × training ($p < .017$); and layer × training ($p < .006$). As can be seen in Figure 2, ResNet50 demonstrated an increase in similarity between the representation of the individual object and the respective scene in the high layer, but only when trained on Places365 ($p < .001$). In AlexNet, there was no significant difference in the high level between the two types of training. However, there was a slight increase in similarity when AlexNet was training on ImageNet for the low layer ($p < .007$) and the middle layer ($p < .016$).

Each context (N=50) had two different related objects. Between-group analysis demonstrated no difference between the objects, and thus the data from both objects were used in the analysis. We used the data analysis with the second object to replicate results that were found with the first object. In the ANOVA with the second object, there was only a significant interaction of network x training ($p < .055$), replicating the effect of training on network found with the first object. Again, the only comparison that reached significance was for ResNet50 in the high layers, where the similarity between object 2 and the scene was much higher when trained with Place365 compared with ImageNet ($p < .001$).



(a) **(b)**

Figure 2: Training and architecture effects on in-out ratios across the layers of **(a)** AlexNet and **(b)** ResNet50

In addition, we compared whether there would be a difference for more shallow (ResNet18) and deeper (ResNet50) architectures. We used an ANOVA to test how the similarity between the object and the scene representation changed as a function of network, training, and layer. Here, we found a significant main effect of training (object 1: $p < .052$; object 2: $p < .012$); and layer (object 1: $p < .068$; object 2: $p < .01$). We also found a significant interaction between training x layer (object 1: $p < .059$; object 2: $p < .011$). In this case, there did not seem to be an effect of depth of an architecture. However, the effect that the highest layer showed the greatest similarity between object and scene when trained with Places365 that was found in ResNet50 (object 1 $p < .001$; object 2 $p < .001$) was replicated in ResNet18 for the second object (object 2 $p < .028$; object 1: numerically higher, but not significant).

Discussion

Context is important for human vision to facilitate object recognition and understanding. We asked whether context was also captured and potentially utilized by CNNs. Prior work has shown object contextual relations implicitly learned at upper layers of diverse networks trained for object recognition (Aminoff et al., 2022). In the present work, we observe implicitly learned relation of objects to related scenes for ResNet trained on Places365. Notably, this object-scene relationship is not learned when training on ImageNet. While object surroundings are included in ImageNet, including typical nearby objects, these surroundings usually are not sufficiently broad in spatial scale to generalize associations between focused object pictures and broad scene images. Weaker context learning also may result from the training aims of ImageNet, labeling objects and not scenes. Places365 training aims to label scenes. Thus context information may be inherently learned in a network trained on Places365.

Object-scene relations may be learned from Place365 by AlexNet, but the heightened in-out-ratio is not significantly above the chance value of 1.0. The greater depth and complexity of ResNet may enable it to implicitly learn object-specific representations in earlier layers prior to generating the final scene label at the top layer. Alexnet's smaller size may not afford this power. Indeed, context learning does not appear to crystalize until the top layer of ResNet, similar to inter-object context relations found in (Aminoff et al., 2022).

The results of this study emphasize the importance of training and how results should be interpreted within the context of how the network was trained.

References

Aminoff, E. M., Baror, S., Roginek, E. W., & Leeds, D. D.

(2022). Contextual associations represented both in neural networks and human behavior. *Scientific reports*, *12*(1), 5570.

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

Bracci, S., Mraz, J., Zeman, A., Leys, G., & Op de Beeck, H. (2023). The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. *PLoS computational biology*, *19*(4), e1011086.

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, *6*(1), 49.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1452–1464.