Implicit Scene Segmentation in Deeper Convolutional Neural Networks

Noor Seijdel (n.seijdel@uva.nl)
Department of Psychology, University of Amsterdam,
Nieuwe Achtergracht 129, 1018 WS Amsterdam, the Netherlands

Nikos Tsakmakidis (nikostsakmakidis@gmail.com)
Machine Learning Group, Centrum Wiskunde & Informatica,
Science Park 123, 1098 XG Amsterdam, the Netherlands

Edward de Haan (e.h.f.dehaan@uva.nl)
Department of Psychology, University of Amsterdam,
Nieuwe Achtergracht 129, 1018 WS Amsterdam, the Netherlands

Sander Bohté (s.m.bohte@cwi.nl)
Machine Learning Group, Centrum Wiskunde & Informatica,
Science Park 123, 1098 XG Amsterdam, the Netherlands

Steven Scholte (h.s.scholte@uva.nl)
Nieuwe Achtergracht 129, 1018 WS Amsterdam, the Netherlands

Abstract:

Feedforward deep convolutional neural networks (DCNNs) are matching and even surpassing human performance on object recognition. This performance suggests that activation of a loose collection of image features could support the recognition of natural object categories, without dedicated systems to solve specific visual subtasks. Recent findings in humans however, suggest that while feedforward activity may suffice for sparse scenes with isolated objects, additional visual operations ('routines') that aid the recognition process (e.g. segmentation or grouping) are needed for more complex scenes. Linking human visual processing to performance of DCNNs with increasing depth, we here explored if, how, and when object information is differentiated from the backgrounds they appear on. To this end, we controlled the information in both objects and backgrounds, as well as the relationship between them by adding noise, manipulating background congruence and systematically occluding parts of the image. Results indicated less distinction between object- and background features for more shallow networks. For those networks, we observed a benefit of training on segmented objects (as compared to unsegmented objects). Overall, deeper networks trained on natural (unsegmented) scenes seem to perform implicit 'segmentation' of the objects from their background, possibly by improved selection of relevant features.

Keywords: Scene segmentation; Object recognition; Deep Convolutional Neural Network;

Introduction

When performing an object recognition task, the visual input elicits a feedforward drive that rapidly extracts basic image features through feedforward connections (Lamme & Roelfsema, 2000). For sparse scenes with isolated objects, this set of features might be enough for successful recognition. For more complex scenes, however, the jumble of visual information ('clutter') may be so great that object recognition cannot rely on having access to a reliable set of features, effectively working as pre-segmented objects. For those images, extra visual operations ('visual routines') that aid the recognition process, such as scene segmentation and perceptual grouping, might require the feedforward activity to be modulated by recurrent loops of activity (Roelfsema, 2006; Groen et al., 2018).

While this view seems to suggest that object recognition only depends on the features that belong to the object, many studies have shown that features from the background can also influence the recognition process. For example, objects appearing in a congruent background are detected more accurately and quickly than objects in an incongruent environment (Davenport & Potter 2004), and many computational models of object recognition use features both from the object and from the background (Riesenhuber & Poggio 1999).

In the current study, we explore how the number of layers (depth) in a DCNN influences object segmentation and how this compares to human vision. We use deep residual networks (ResNets; He, Zhang, Ren & Sun, 2016) to systematically manipulate network depth, because they can be up-scaled by adding their basic building blocks without altering the architecture in another way. We presented seven DCNNs (with increasing depth) and 38 human participants with images of segmented and unsegmented objects. To investigate the influence of features from the background on object recognition, we generated stimuli in which objects were placed on top of congruent or incongruent scenes. Thereby we ask to what extent DCNNs exhibit the same sensitivity to scene properties (i.e. context) as human observers. To complement our findings, we further explore the role of segmentation on learning by training ResNets on a dataset with segmented objects, and a dataset in which objects were embedded in a scene.

Experiment 1: background congruence

Methods

Stimuli Images of 27 different object categories were generated by placing cut-out objects from the ImageNet validation set onto white (segmented), congruent and incongruent backgrounds. There were ten exemplars for every category, and backgrounds were sampled from the SUN2012 database (512x512 pixels, full-color). For each category, three congruent backgrounds were selected using the five most common places where this object was found within the database. Three incongruent backgrounds were manually chosen.

Participants and networks 38 participants (9 males) aged between 18 and 30 years (*M* = 22.03, *SD* = 3.02) took part in the experiment. To investigate the effect of depth on scene segmentation in DCNNs, tests were conducted on ResNets with increasing number of layers (10, 18, 34, 50, 101, 152), using the fb.resnet.torch implementation by Gross & Wilber (2016). Input images from the ImageNet dataset (Russakovsky et al., 2015) were 224x224 randomly cropped from a resized image using the scale and aspect ratio augmentation of Szegedy et al. (2015). Downsampling was done by stride-2 convolutions in the 3x3 layer of the first block in each stage (instead of the first 1x1 convolution) and weight decay was applied to all weights and biases (instead of just the weights of the convolution layers). ResNet-10 was trained on ImageNet with 1 GPU. We used pre-trained versions for the other ResNets.

Human performance

A repeated-measures ANOVA, with factor Background differentiated accuracy across the three conditions, $F(2,74) = 366.2$, $p < .001$, $\eta^{2par} = .91$ (Figure 1D). Participants made fewer errors for segmented objects than for the congruent, $t(37) = 15.655$, $p < .001$, and incongruent condition, $t(37) = 27.6$, $p < .001$. Additionally, participants made fewer errors for congruent than for incongruent, $t(37) = 9.376$, $p < .001$ (Bonferroni corrected). Overall, results indicate that when a scene is glanced briefly (32 ms), objects are not (always) completely segregated from their background and semantic consistency information influences object perception.

Network performance

For human participants, results indicated that features from the background influenced object perception. Do DCNNS show a similar pattern and how is this influenced by network depth?

Experiment 1 showed both a substantial overlap and difference in performance between human participants and DCNNs. Both were better in recognizing an object

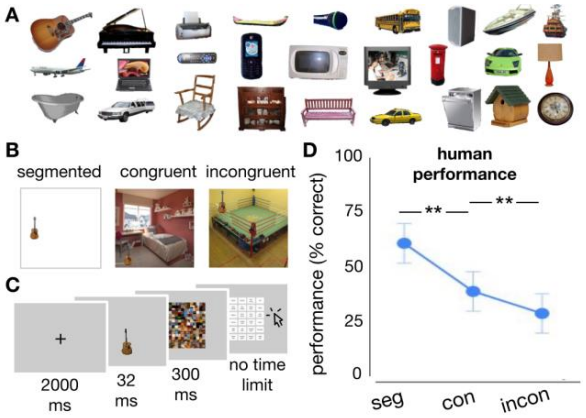

Figure 1: Stimuli and experimental design. A) Exemplars of the 27 object categories. B) Stimuli were generated by placing the object onto white, congruent and incongruent backgrounds. C) Participants performed on an object recognition task. D) Human performance (% correct).

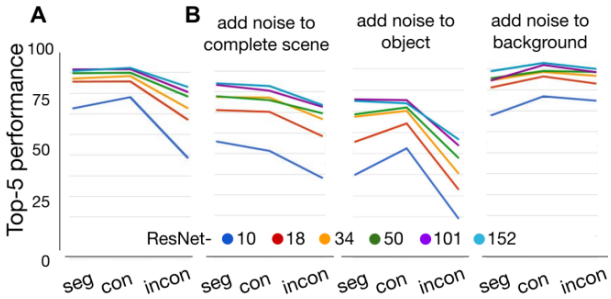

Figure 2: DCNN performance on the object recognition task. A) Top-5 performance of the different ResNets on segmented, congruent and incongruent images B) Top-5 performance after adding noise to the object, background, or both.

on a congruent versus an incongruent background. However, whereas human participants performed best in the segmented condition, DCNNs performed equally well (or better) for the congruent condition. Performance for the incongruent condition was lowest. This effect was particularly strong for more shallow networks.

To further investigate the degree to which the networks are using features from the object and/or background, we systematically occluded different parts of the image and evaluated the changes in activation of the correct class, before the softmax activation function (Zeiler & Fergus, 2014). We quantified the importance of features in the object vs. background by averaging the change across pixels belonging either to the object or the background. For this analysis, positive values indicate that pixels are helping classification (higher values indicating a higher importance). For example, figure 3A shows that the network is localizing the object in the scene, as the activity drops significantly when the object (china cabinet in this example) is occluded.

To evaluate whether deeper networks are better at localizing the objects in the scene, while ignoring

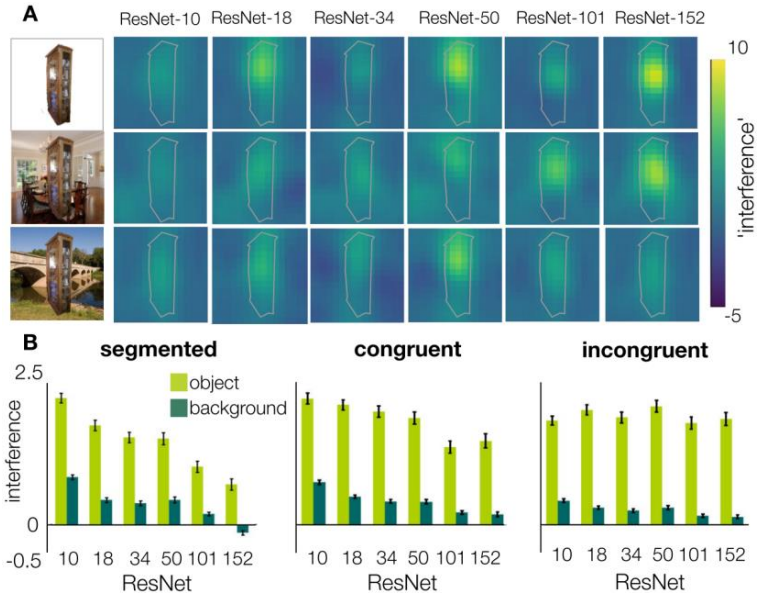

Figure 3: Systematic occlusion of parts of the image. A) Examples where we occluded different portions of the scene, and visualize how the classifier output for the correct class changed (before the softmax activation function). Images were occluded by a gray patch of 128x128 pixels, sliding across the image in 30 pixel steps. B) The relative change in activation (compared to the original image), after occluding pixels of either the object or the background, for the different conditions (segmented, congruent, incongruent). Error bars represent 1 SEM.

irrelevant background information, we computed the relative drop in performance when pixels of the background vs. pixels of the object were occluded. Results indicated a larger influence of background pixels on classification for more shallow networks, for all conditions. For those models, pixels from the object had a larger impact as well, for the segmented and congruent condition.

Experiment 2: Training

Next, we investigated how training is influenced by network depth. If deeper networks indeed implicitly learn to segment object from background, we expect them to show a smaller difference in learning speed, when trained with segmented vs. unsegmented stimuli (as compared to shallow networks).

Methods

Stimuli To train the models, images from 10 categories were selected from ImageNet. We used 10 categories to obtain a reasonable mixture of ease of computing and performance gradients that show a substantial difference from untrained to trained. With the selected images, we generated two training sets: one in which the objects were segmented, and one with the original images (objects embedded in scenes). Objects were segmented using a DCNN trained on the MS COCO dataset (Lin et al. 2014), using the Mask R-CNN method (He, Gkioxari, Dollár & Girshick, 2017). Images with object probability scores lower than 0.98 were discarded, to minimize the risk of selecting wrongly classified or low quality images. Images were resized to 128x128 pixels. In total, the set contained ~9000 images, 80% was used for training, 20% was used for validation.

Networks As in experiment 1, we used ResNets with increasing number of layers (6, 10, 18, and 34). Deeper networks generated overfitting problems and were not included.

Network convergence

Accuracy of the ResNets was evaluated after each epoch (100 total) on the validation set. Results indicated a higher classification accuracy in the early stages of training for the networks trained on segmented objects, compared to those trained on unsegmented objects (Figure 4A). Additionally, these networks converged (accuracy constant >10 epochs) in less epochs. In later epochs, accuracy between the two types of networks was similar. Shallow networks trained on segmented stimuli converged earlier than those trained on unsegmented stimuli. The difference in epochs until convergence decreased as the network depth increased. These results confirm that networks need to learn to segment objects from their background for optimal performance.

Discussion

Classic models of grouping and segmentation presume an explicit process in which certain elements of an image are grouped, whilst other are segregated from each other, by a labelling process. Our results from behavioral experiments with segmented and unsegmented objects indicate that recognition can take place without an explicit segmentation step. Furthermore, we show that segmentation can, and for DCNNs does, arise implicitly as a function of network depth.

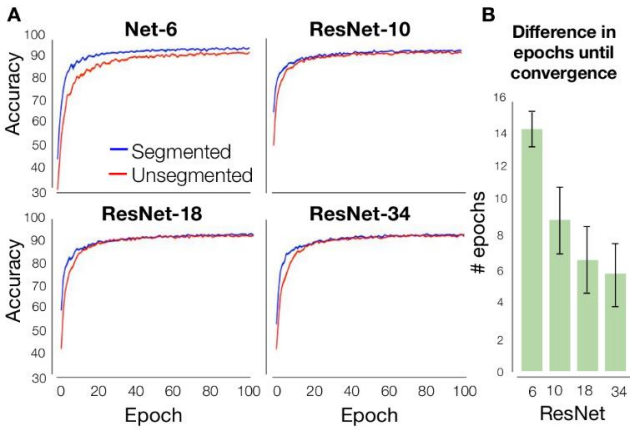

Figure 4: Influence of training on segmented vs. unsegmented objects. A) Accuracy during training. B) Difference in epochs until convergence. Networks were reinitialized and tested with 10 different seeds to obtain statistical results. Error bars represent 1 SEM.

Different accounts of object recognition in scenes propose different loci for contextual effects (Oliva & Torralba, 2007). It has been argued that a bottom-up visual analysis is sufficient to discriminate between basic level object categories, after which context may influence this process in a top-down manner, by priming relevant semantic representations, or by constraining the search space of most likely objects (e.g. Bar, 2003). The current results show that context features may impact object recognition in a bottom-up fashion, even for objects in a spatially incongruent location.

Instead of being an ultra-deep feedforward network, the brain might employ recurrent connections for object recognition in complex natural environments. The interpretation that deeper networks are better at object recognition, because they are capable of limiting their analysis to (mostly) the object –when necessary– is consistent with the idea that deeper networks are solving the challenges that are resolved by recurrent computations in the brain (Liao & Poggio, 2016).

Conclusion

We investigated the extent to which object and context information, and the interplay between them, impacts object recognition for both DCNNs and human observers. Combined, the current findings show that with an increase in network depth there is better selection of the features that belong to the object category. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and might be one of the ways in which scene segmentation is implemented in the brain.

Acknowledgments

We thank Yannick Vinkesteijn for help with data collection (human participants).

References

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience, 15*(4), 600-609.

Groen, I.I.A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V.A.F., & Scholte, H.S. (2018). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS computational biology, 14*(12), e1006690.

Gross S, Wilber M. Training and investigating residual nets. (2016). *Facebook AI Research*, CA [Online] Available: http://torch.ch/blog/2016/02/04/resnets.html

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

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

Lamme, V.A.F., & Roelfsema, P.R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, *23*(11),571-579

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

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in cognitive sciences, 11*(12), 520-527.

Riesenhuber, M., & Poggio, T. (1999). Are cortical models really bound by the "binding problem*"?. Neuron, 24*(1), 87-93.

Roelfsema, P.R. (2006). Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, *29*,203-227.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A.C. (2015). Image-net large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Zeiler M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision (*pp. 818–833).