Deep Neural Networks Represent Semantic Category in Object Images Independently from Low-level Shape

Astrid Zeman (astrid.zeman@kuleuven.be)
J Brendan Ritchie (brendan.ritchie@kuleuven.be)
Stefania Bracci (stefania.bracci@kuleuven.be)
Hans Op de Beeck (hans.opdebeeck@kuleuven.be)
Brain and Cognition Department, 102 Tiensestraat
Leuven 3000, Flemish Brabant, Belgium

Abstract

Deep Neural Networks (DNNs) categorize object images with extremely high levels of accuracy, with performance that is able to match, or even surpass, humans. In natural images, category is often confounded with shape information, therefore it is possible that DNNs rely heavily upon visual shape, rather than semantics, in order to discriminate between categories. Using two datasets that explicitly dissociate shape from category, we quantify the extent to which DNNs represent semantic information independently from shape. One dataset defines shape as a high-level property, namely low versus high aspect ratio. The second dataset defines shape as 9 different types that best represent low-level, retinotopic shape. We discover that DNNs are able to encode semantic information independently from low-level shape, peaking at the final fully connected layer in multiple DNN architectures. The final layer of multiple DNNs represents high-level shape to the same level of correlation as category. This work suggests that DNNs are able to bridge the semantic gap, by representing category independently from low-level shape.

Keywords: deep networks; object images; shape; category

Introduction

In recent years, the performance of Deep Neural Networks (DNNs) has improved significantly, such that they are able to meet (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015; Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke & Rabinovich, 2015), and even surpass human performance in classifying objects. There are increasing comparisons between these artificial networks and their biological counterparts, building up a wealth of evidence for their use as a benchmark model of visual object recognition (Kietzmann, McClure, & Kriegeskorte, 2017). While recent performance of DNNs for object classification tasks is impressive, it is unclear to what extent DNNs are representing the true semantics of categories. Deriving high-level semantic meaning from low-level feature descriptions is commonly referred to as the “semantic gap” in computer vision literature (Markowska-Kaczmar & Kwaśnicka, 2018). To establish the level at which DNNs bridge the semantic gap and extract meaningful information from images, they require testing with images that reduce any possible reliance upon low-level features that could be exploited to improve performance. Shape and category information significantly interact in natural images (Bracci & Op de Beeck, 2016). Given that these networks represent shape information (Kubilius, Bracci, & Op de Beeck, 2016), it is possible that these architectures are exploiting shape features for classification, without learning any underlying category semantics. In this paper we test network performance on carefully designed images that minimize potential dependencies between category and influencing features.

Methods

Using Representational Similarity Analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008), we test object shape and category information in each layer of multiple DNNs. We analyze 4 different deep networks: GoogLeNet, VGG-16, VGG-19 and CaffeNet.

DNN architectures

We take four leading DNN architectures: GoogLeNet (Szegedy, et al., 2015), CaffeNet, which is an implementation of AlexNet as described in Krizhevsky, Sutskever, & Hinton (2012), and VGG-16 and VGG-19 (Simonyan & Zisserman, 2015). CaffeNet, VGG-16 and VGG-19 all have layers stacked in a single column with increasing depth of 8, 16 and 19 layers respectively. These three architectures have chained convolutional operations followed by max pooling. GoogLeNet diverges from this standard architectural arrangement with the addition of miniature networks embedded within the global architecture, referred to as "inception modules", which are multi-sized convolutional operations configured in parallel. GoogLeNet has a maximum depth of 22 parameterized layers.

Stimulus Sets

We use two stimulus sets that are designed to dissociate shape from category information. Both stimulus sets are greyscale images of objects on a

white background, centered at the origin and presented at a normal viewing angle (see Figure 1). Stimulus Set A (top row) contains 32 unique images, divided into 2 equal-sized categories (animal vs non-animal) and 2 equal-sized shapes (low and high aspect ratio). Stimulus Set B (bottom row) contains 54 images divided into 6 object categories (minerals, animals, fruit/veg, music, sport and tools) and 9 shape types. All stimuli are balanced across common shape information (circled in dashed grey). Each category division is highlighted by a distinct color. Details on Set B are found in (Bracci & Op de Beeck, 2016).

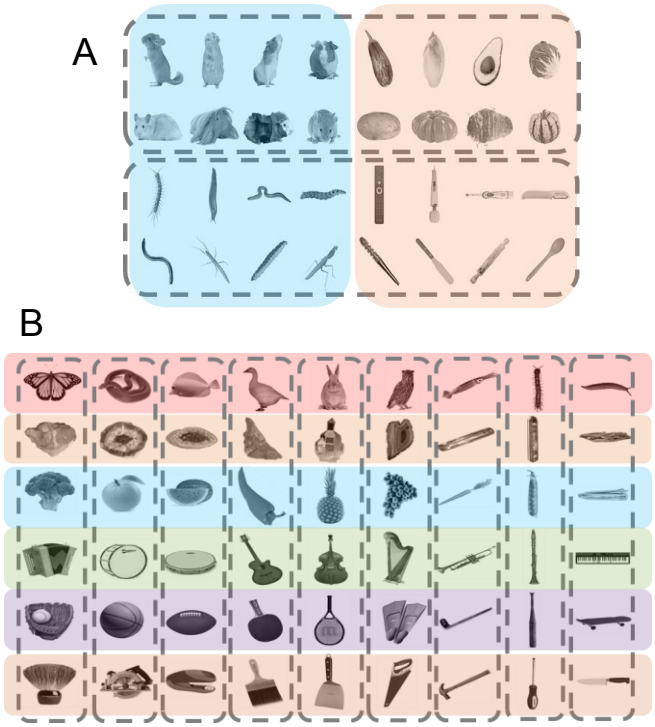

Figure 1: Stimulus Sets. Top (A): 32 stimuli in 2 categories (animal and non-animal), Bottom (B) 54 stimuli in 6 categories (animals, minerals, fruit/vegetables, music, sports equipment, tools).

Shape and Category Models

Shape and category models are represented as binary value Representational Dissimilarity Matrices or RDMs, where 0 indicates no dissimilarity (pairs of stimuli are within the same category or shape type) and 1 indicates full dissimilarity (pairs of stimuli are in different categories or shape types). See Figure 2 for a visual representation. In Set A, object images are numbered sequentially within each cluster of gerbils, insects, tools and vegetables. In Set B, object images are numbered in order from top to bottom, left to right in Figure 1.

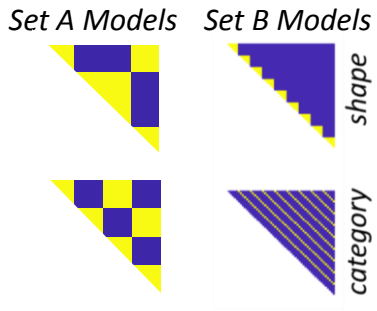

Figure 2: RDMs of shape (top row) and category (bottom row) models for Set A (left) and B (right).

Results

We measure the RSA between shape and category models for every layer of each DNN. Each plot also contains a significance threshold, which is twice the standard deviation of the correlation between the DSM of each DNN layer and 10,000 randomized category conceptual models. Values above the significance threshold fall within $p < 0.05$.

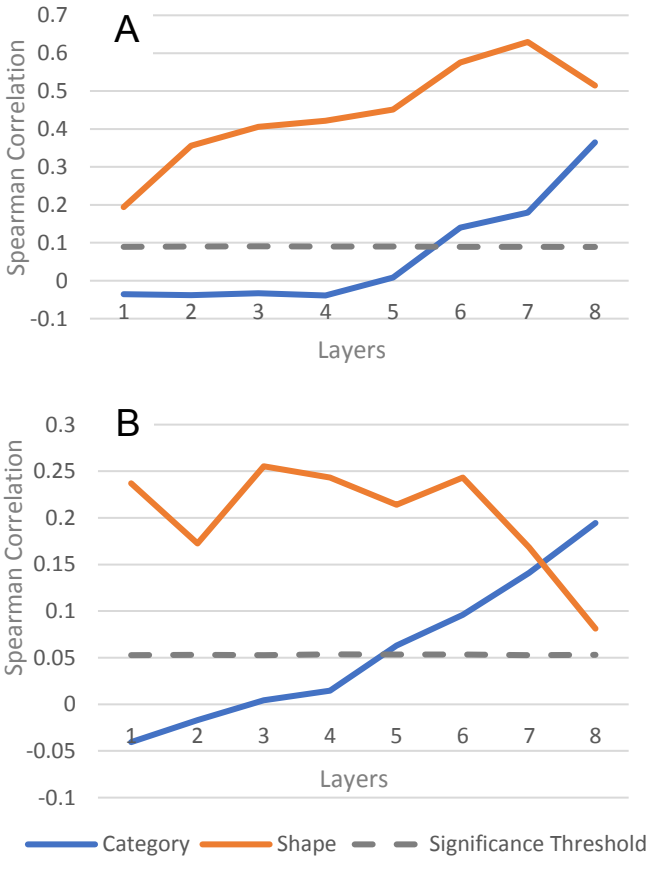

Figure 3: CaffeNet correlations with shape (orange) and category (blue) for Set A (top) and B (bottom).

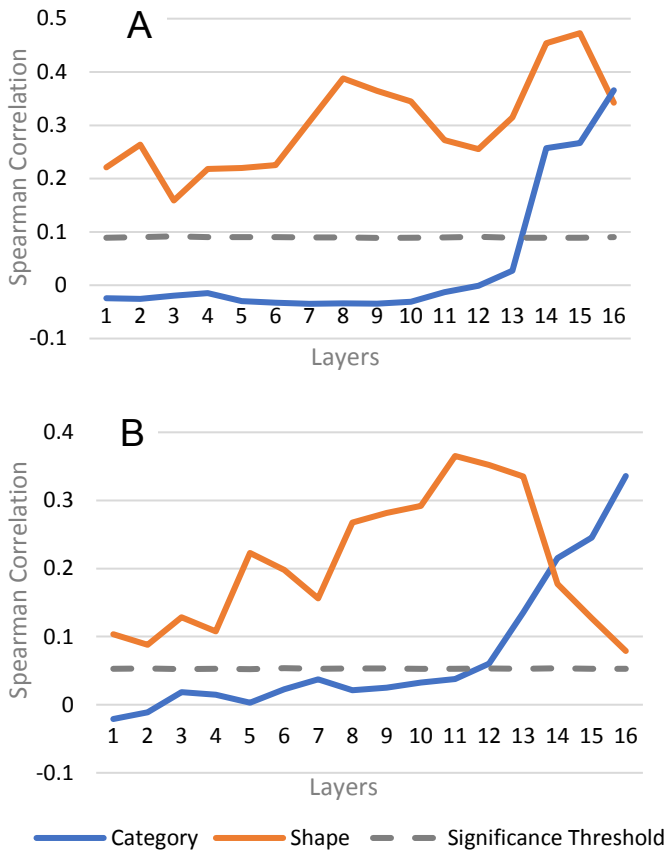

Figure 4: VGG-16 correlations with shape (orange) and category (blue) for Set A (top) and B (bottom).

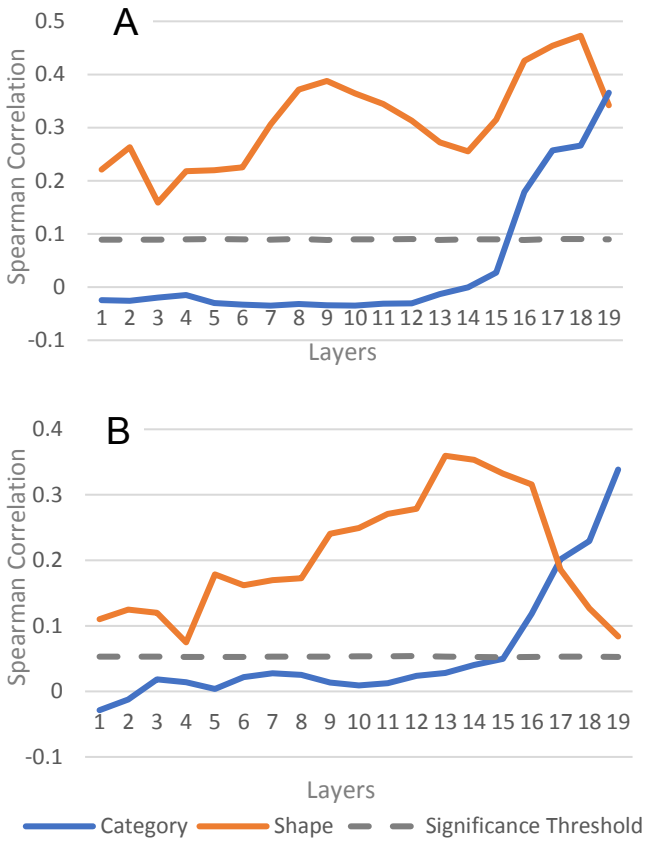

Figure 5: VGG-19 correlations with shape (orange) and category (blue) for Set A (top) and B (bottom).

In CaffeNet (Fig 3), category information remains low in the first few layers before increasing to above significance in later layers, reaching a maximum at the final fully-connected layer for both datasets. In Set B, category information overtakes shape, which does not occur in Set A. VGG-16 (Fig 4) and VGG-19 (Fig 5) illustrate similar trends in performance. In both networks, category overtakes shape for Set B at the penultimate layer, which does not occur in Set A, where category meets the level of shape correlation at the final layer. In VGG-16, category information does not reach significance until layer 14 for Set A, and layer 13 for Set B. In VGG-19, category correlations reach significance at layer 16 for both Sets A and B.

In GoogLeNet (Fig 6), shape exceeds category correlations for all layers in Set A, and most layers in Set B. For both datasets, category information peaks at the final layer, which exceeds shape correlations for Set B but not A. Viewing layer by layer transformations, there are 3 occurrences across both datasets where there is a dip in shape and a rise in category information of greater than 5%, occurring at layers 21, 39 and 51. These layers all perform pool projection – suggesting that this type of operation boosts category information while simultaneously reducing shape.

Conclusions

We extend upon work demonstrating that DNNs are able to encode category-orthogonal properties of objects, providing evidence that these artificial networks learn category semantics independently from low-level visual shape information. We test four leading DNN architectures using two stimulus sets that are carefully designed to orthogonalize shape from category information. Across all DNN architectures tested, shape information peaks prior to category. Category information reaches a maximum at the final layer of all DNNs. In single column architectures (CaffeNet, VGG-16 and VGG-19), category information rises above significance in the final three layers of these deep networks. In a deep parallelized architecture (GoogLeNet), there are rises and falls above and below significance for category information as the layers are traversed, with the largest correlation in category information occurring at the final layer. Our results demonstrate that DNNs are able to significantly represent category semantics in the final layers across multiple DNNs, suggesting they are able to bridge the semantic gap by distinguishing objects beyond rudimentary shape properties.

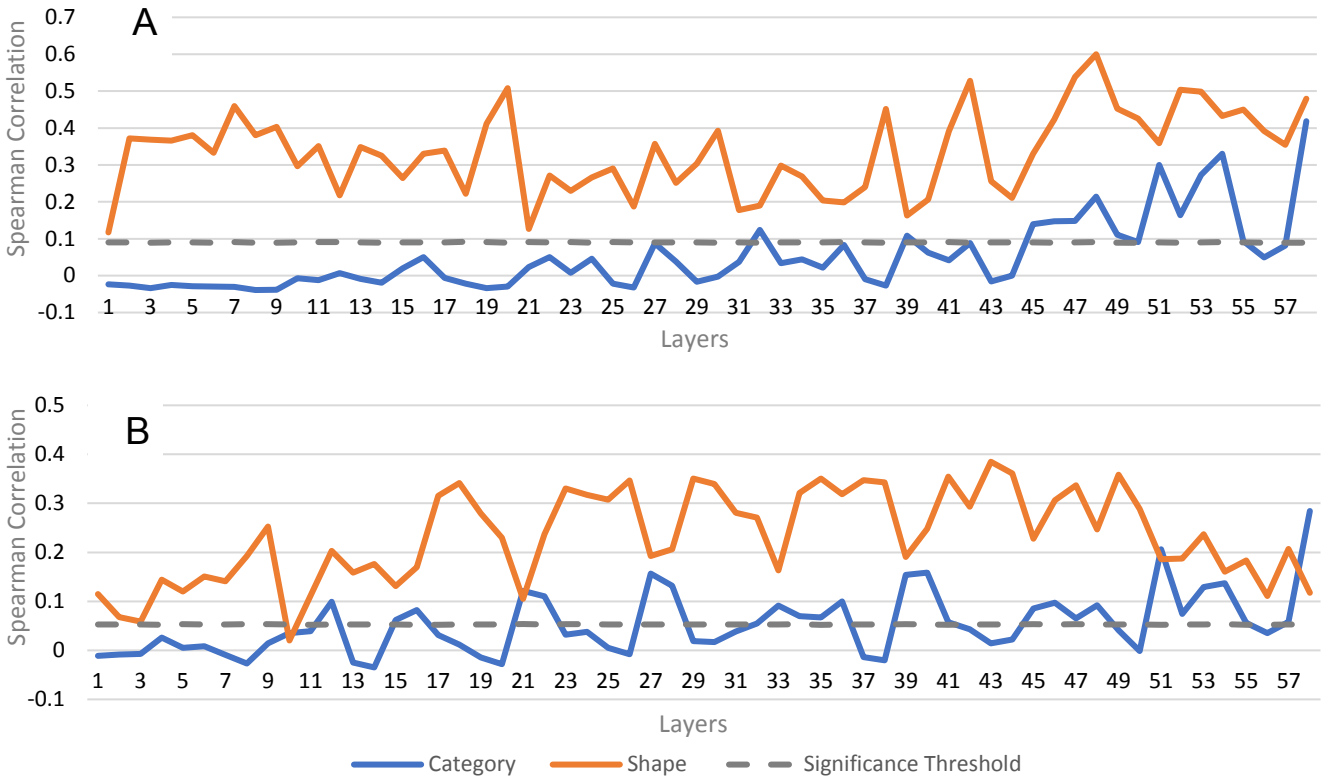

Figure 6: GoogLeNet correlations with shape (orange) and category (blue) for Set A (top) and B (bottom).

Acknowledgments

A.Z. is funded by grant C14/16/031 of the KU Leuven Research Council. S.B. is funded by FWO (Fonds Wetenschappelijk Onderzoek) postdoctoral fellowship 516 (12S1317N). H. O. dB. is funded by the European Research Council (ERC- 2011-StG-284101), a federal research 514 action (IUAP-P7/11), the KU Leuven Research Council (C14/16/031), and Hercules grant ZW11_10 to 515. J.B.R. is funded by FWO[PEGASUS][2] Marie-Sklodowska-Curie Fellowship 12T9217N.

References

Bracci, S., & Op de Beeck, H. (2016). Dissociations and Associations between Shape and Category. The Journal of Neuroscience, 36(2), 432-444.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep Neural Networks In Computational Neuroscience. bioRxiv. doi:https://doi.org/10.1101/133504

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2(4). doi: 10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (NIPS 2012) (pp. 1097-1105). Lake Tahoe: Curran Associates, Inc.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. PLoS Computational Biology, 12(4), e1004896. doi:10.1371/journal.pcbi.1004896

Markowska-Kaczmar, U., & Kwaśnicka, H. (2018). Deep Learning - A New Era in Bridging the Semantic Gap. In H. Kwaśnicka, & L. Jain (Bridging the Semantic Gap in Image and Video Analysis (pp. 123-159). Cham: Springer. doi:https://doi.org/10.1007/978-3-319-73891-8_7

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR, (arXiv:1409.1556). https://arxiv.org/abs/1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 1-9). Boston, MA.