Naturalistic dataset augmentations lead to more human-like recognition of occluded objects in convolutional neural networks.

David D Coggan (david.coggan@vanderbilt.edu)
Department of Psychology, Vanderbilt University
2201 West End Ave, Nashville, TN 37235

Frank Tong (frank.tong@vanderbilt.edu)
Department of Psychology, Vanderbilt University
2201 West End Ave, Nashville, TN 37235

Abstract

Convolutional neural networks (CNNs) are currently the strongest overall predictors of human neural and behavioral responses to object stimuli. However, CNNs are typically much more susceptible than humans to image perturbations such as occlusion. Here, we investigated how augmenting training datasets might lead to more occlusion-robust CNNs that better predict human visual behavior. To address this question, we trained separate instances of the same CNN architecture (CORnet-S) to classify the ImageNet 1k dataset either a) without augmentation, b) with occlusion by artificially generated shapes without texture, c) with occlusion by naturalistic shapes derived from photographs, also without texture, and d) with occlusion by naturalistic shapes with original textures preserved. After training, we used an occluded object stimulus set from a human behavioral study to measure classification accuracy and predictivity of human responses for each model. Compared to the standard dataset, we found that both artificial and natural occlusion-training led to increased accuracy, however, only natural occlusion training led to greater human-likeness, with separate benefits of naturalistic shape and texture. Overall, these findings indicate that human occlusion robustness may be shaped by the specific forms of occlusion that occur in nature.

Keywords: behavior, CNN, robustness, occlusion

Introduction

Over the last decade, advances in computer vision have had a substantial impact on how human object perception and associated cortical regions are investigated and understood. Deep, image-computable architectures such as convolutional neural networks (CNNs) and, more recently, transformer models have scored highly on visual tasks such as object classification, in some cases approaching human-level performance (He, Zhang, Ren, & Sun, 2016; Naseer et al., 2021). Subsequent computational neuroscience studies revealed that the intermediate representations in convolutional models strongly predict neural responses in primate visual cortices to the same images (Güçlü & a. J. van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018), suggesting a high level of correspondence between CNNs and the brain. However, when assessed using images distorted by noise, blur or occlusion, CNNs typically perform much worse than humans (Geirhos et al., 2018; Jang, McCormack, & Tong, 2021), suggesting that they lack the perceptual mechanisms that afford visual robustness in humans.

Object occlusion is a relatively under-investigated image perturbation for which there is a striking human-machine robustness gap. Despite obscuring objects causing a partial or spatially fragmented view of an object, Human observers can often readily perceive and recognize objects despite being partially obscured by other objects. By contrast, covering just 50% of an image with randomly positioned black patches is sufficient to reduce classification accuracy by a CNN to drop from approximately 75% to less than 1% (Naseer et al., 2021). Previous studies have shown that recurrent processing can partially account for the human advantage, as disrupting recurrent processing in humans leads to weaker occlusion robustness, while augmenting feedforward models with recurrent processing improves robustness and correspondence with human neural responses (Tang et al., 2018; Rajaei, Mohsenzadeh, Ebrahimpour, & Khaligh-Razavi, 2019; Svanera, Morgan, Petro, & Muckli, 2021). However, it is less clear how differences in the visual diet of humans and CNNs might also account for the occlusion robustness gap. For instance, CNNs are typically trained on databases constructed by scraping photographs of objects from the internet which, while naturalistic, generally consist of clearer, higher-quality views than those experienced by a human during development. The occlusion robustness gap may therefore be a product of the distribution of the training data.

In the present study, we explored how augmenting training datasets with greater levels and various forms of occlusion may lead to more occlusion-robust CNNs that better predict human visual behavior. We found that, while any form of occlusion-training led to increased robustness, only naturalistic forms of occlusion led to more human-like classification behavior. Moreover, this effect was partially reduced, but survived the removal of the natural occluder image texture during training. Taken together, these results indicate that human occlusion robustness mechanisms emerge from both the quantity and quality of occlusion that occurs in the natural world.

Methods/Results

Human Behavior

We presented 30 human subjects with 752 images from eight different object categories (Figure 1A). Images were taken from ImageNet (Deng et al., 2010) and converted to grayscale. Each image was presented without occlusion or with a unique occluding shape superimposed over the object. Nine different types of occluder were used, eight of which were computer generated and one of which was based on the photographs of objects. Occluders had a uniform texture, either black or white, and revealed between 10% and 80% of the object image. The pairing of object images and occluders was randomized for each subject. Each image was presented to human subjects at 10 degrees visual angle for 100ms before being replaced by a pink Fourier noise pattern. The subject then made an 8 AFC classification response. Classification accuracy as a function of visibility is shown for each occluder type and color in Figure 1B.

Computational Modeling

We trained four CORnet-S architectures (Kubilius et al., 2018) to classify ImageNet 1k with the following augmentations to the dataset: no augmentation; occlusion by uniformly colored computer-generated shapes; occlusion by uniformly colored shapes derived from natural object images; occlusion by natu-

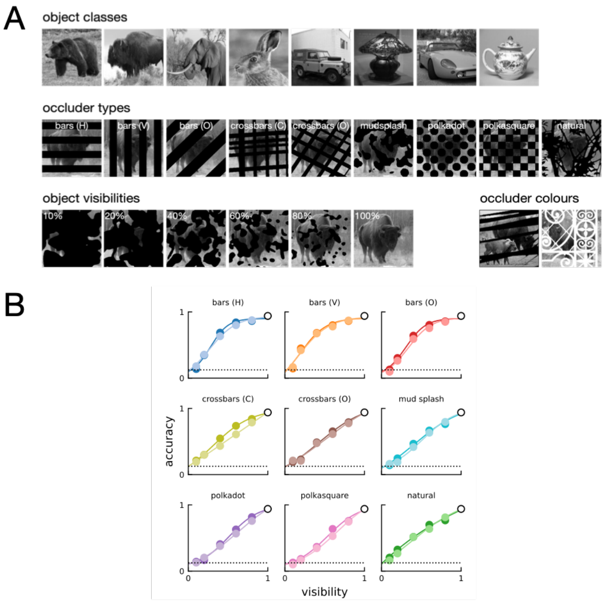

Figure 1: Stimuli and human behavioral results. A: Example stimulus from each condition. B: Group mean classification accuracy for each occluder type, color and visibility. White and black occluders shown in lighter and darker colors, respectively. White circle shows performance for unoccluded images. Colored curves are sigmoid functions fitted the accuracy data. Dotted line shows chance performance.

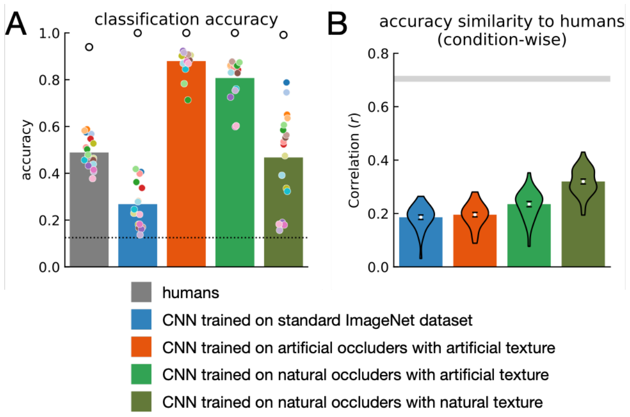

Figure 2: Computational modeling results. A: Mean classification accuracy for all occluded images (bar) and for each occluder type/color (colored dots, mapping shown in Figure 1). B: Human-likeness of each CNN, measured as the correlation between the set of colored points shown in A. Correlation was performed separately for each subject, with violin plot showing the distribution of coefficients. White dot and bar height show mean across subjects; errorbars show SEM across subjects. Noise ceiling is shown as the grey bar.

ral objects with original textures intact. After training, we input images from the behavioral experiment into the models and measured classification accuracy. All occlusion-trained models showed greater accuracy than the model trained on the standard ImageNet dataset (Figure 2A). To directly compare each model with humans, we correlated the set of mean accuracies across the different occlusion conditions (Figure 2B). Compared to the standard dataset, training on artificial occluders did not significantly affect human likeness (t(29)=1.07, corrected p=1). However, training on natural occluder shapes with artificial textures led to higher human likeness than both standard training (t(29)=4.87, corrected p=.0002) and artificial occlusion training (t(29)=7.04, corrected p=.0001). Finally, the combination of natural occluder shapes with natural textures led to higher human-likeness than natural shapes alone (t(29)=9.03, corrected p=.0001).These results suggest that augmented occlusion training can lead to the acquisition of more robust human-like representations, but only if the occluding stimuli resemble the shape and textural properties of natural images.

Discussion

We found that training only on naturalistic forms of occlusion increased led to more human-like performance in CNNs. This indicates that increasing the quantity of occlusion during training does not guarantee more human-like occlusion-robust models. Instead, more qualitative aspects of real-world occlusion appear to influence human occlusion robustness. Specifically, both the shape and texture of naturalistic occluder shapes were independently advantageous in producing human-like patterns of behavior. This finding adds to previous research on the human-machine occlusion robustness gap, which has identified recurrent processing mechanisms as a critical feature that affords more accurate classification and/or more better predictions of human behavioral or neural responses to occluded images (Rajaei et al., 2019; Tang et al., 2018).

Even our most human-like model accounted for only half of the explainable variance in the human data. Given that the CNN used here already contains recurrent connections, this indicates that other factors are needed to account for human occlusion robustness - an exciting target for future modeling.

While different data augmentation led to different test accuracies, these differences are confounded by different degrees of similarity in the forms of occlusion applied during training and testing. Indeed, this factor strongly predicts the pattern of model accuracies. Further testing of these models on an out of distribution occluder dataset is necessaru to establish whether there are reliable differences in robustness across these models.

In summary, we present evidence that the robust perception of occluded objects in humans is shaped by the specific forms of occlusion that appear in the natural world.

Acknowledgments

This research was supported by grants from the National Institutes of Health R01-EY035157 (to F.T.) and P30-EY008126 to the Vanderbilt Vision Research Center (Director Dr. Calkins).

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2010, 3). Imagenet: A large-scale hierarchical image database. In (p. 248-255). Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/cvpr.2009.5206848

Geirhos, R., Schütt, H. H., Temme, C. R. M., Bethge, M., Rauber, J., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *2018-Decem*, 7538-7550.

Güçlü, U., & a. J. van Gerven, M. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*, 10005-10014. Retrieved from `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5023-14.2015` doi: 10.1523/JNEUROSCI.5023-14.2015

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In (p. 770-778). Retrieved from `http://image-net.org/challenges/LSVRC/2015/`

Jang, H., McCormack, D., & Tong, F. (2021, 12). Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS Biology*, *19*. doi: 10.1371/JOURNAL.PBIO.3001418

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*. doi: 10.1371/journal.pcbi.1003915

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & Dicarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, *10.1101*. Retrieved from `https://doi.org/10.1101/408385` doi: 10.1101/408385

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2021). Intriguing properties of vision transformers.. Retrieved from `https://git.io/Js15X.`

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S. M. (2019, 5). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Computational Biology*, *15*. doi: 10.1371/journal.pcbi.1007001

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007. Retrieved from `https://www.biorxiv.org/content/early/2018/09/05/407007` doi: 10.1101/407007

Svanera, M., Morgan, A. T., Petro, L. S., & Muckli, L. (2021). A self-supervised deep neural network for image completion resembles early visual cortex fmri activity patterns for occluded scenes. *Journal of Vision*, *21*, 1-17. doi: 10.1167/jov.21.7.5

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., ... Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 8835-8840. doi: 10.1073/pnas.1719397115