Representation of adversarial images in deep neural networks and the human brain

Chi Zhang[a] (zcboluo@hotmail.com), Xiaohan Duan[a] (d125237229@163.com),
Ruyuan Zhang[b*] (zhan1217@umn.edu), Li Tong[a*] (tttocean@163.com)
[a]National Digital Switching System Engineering and Technological Research Center
Zhengzhou, Henan, 450000 China
[b]Center for Magnetic Resonance Research, Department of Radiology
University of Minnesota, Minneapolis, MN 55455 USA
* Co-senior author

Abstract:

Many studies have demonstrated the prominent similarity between deep neural networks (DNNs) and human vision. However, one recent study (Nguyen et al., 2015) challenged this idea and showed that some artificially generated adversarial images can successfully 'fool' the even most state-of-the-art DNNs but not human vision. Specifically, DNNs can accurately recognize adversarial noise (AN) images but not adversarial interference (AI) images, and vice versa in humans. In this paper, we aim to use functional magnetic resonance imaging (fMRI) to elucidate the neural mechanisms that underlie these dissociable behaviors. We measured neural responses in the human brain towards regular, AN and AI images, and quantify the representational similarity between the three types of images in a DNN and in the human brain respectively. Results demonstrated that the representational similarity in the DNN reflects image similarity more than perceptual similarity. We also found that the DNN misrepresents low- and middle-level visual features compared to human vision. These results offer new insight into the development of both human visual models and deep neural networks in future work.

Keywords: DNN, fMRI, adversarial image

Introduction

Deep neural networks (DNNs) have become a contention focus in recent years due to its remarkable representational power of visual features. Recent research has strived to link DNNs to the human visual system (Yamins and DiCarlo, 2016). For example, recent fMRI and single-unit recording studies have shown that visual features in different layers of DNNs can explain the neural responses along the ventral and the dorsal cortical pathway in the human brain, revealing the consistency between two systems in spatiotemporal visual processing (Güçlü and van Gerven, 2015;Cichy et al., 2016;Hong et al., 2016;Güçlü and van Gerven, 2017;Horikawa and Kamitani, 2017). The close links between DNNs and neuroscience suggest that modern DNNs not only can quickly approach or even surpass behavioral performance of human vision, but also bore strong resemblance to the neural representation in the brain.

On the other hand, current DNNs are in general still considerably worse than human vision in many aspects, revealing some fundamental difference between the two systems. One potent example is the artificially generated adversarial images that can successfully 'fool' the even most state-of-the-art DNNs (Fig. 1A). Adversarial noise (AN) images look like meaningless noise to humans. DNNs, however, classify them into common object categories with surprisingly high confidence (Nguyen et al., 2015). On the contrary, humans can easily recognize the adversarial interference (AI) images generated by adding a small amount of noise to the regular (RE) images. But the same manipulation severely impairs DNNs' ability to recognize AI images (Szegedy et al., 2013). These intriguing effects demonstrate the profound difference between DNNs and human vision as same image inputs produce drastically distinct perceptual outcomes in the two systems.

To further improve DNNs to achieve human-level performance, we should characterize to what extent the visual representation in DNNs approximates to or differs from the representation in the human visual system. Past research used the images stimuli that generate highly similar percepts in both systems and it is therefore not surprising that their representations are also found to be similar. Only by using the stimuli that cause dissociable perceptual outcomes, can we truly elucidate in what aspect the representations of the two systems differ. As such, we performed a fMRI experiment on three human participants and measured the neural response towards the regular and the adversarial images. Our aim is to uncover the neural mechanisms of the dissociable visual behaviors between DNNs and humans when recognizing adversarial images.

Materials and Methods

Stimuli. We used three types of images: regular, AN, and AI images, respectively (Fig. 1A). 40 RE images span 40 representative object categories. 40 AN images were generated based on the RE images using the method described in (Nguyen et al., 2015). Briefly, to generate an AN image that corresponds to a RE image, we used backpropagation method to calculate the gradient of the posterior probability of the corresponding class of the RE image. We then followed the gradient to increase a chosen unit's activation by adjusting the noise image. Optimization

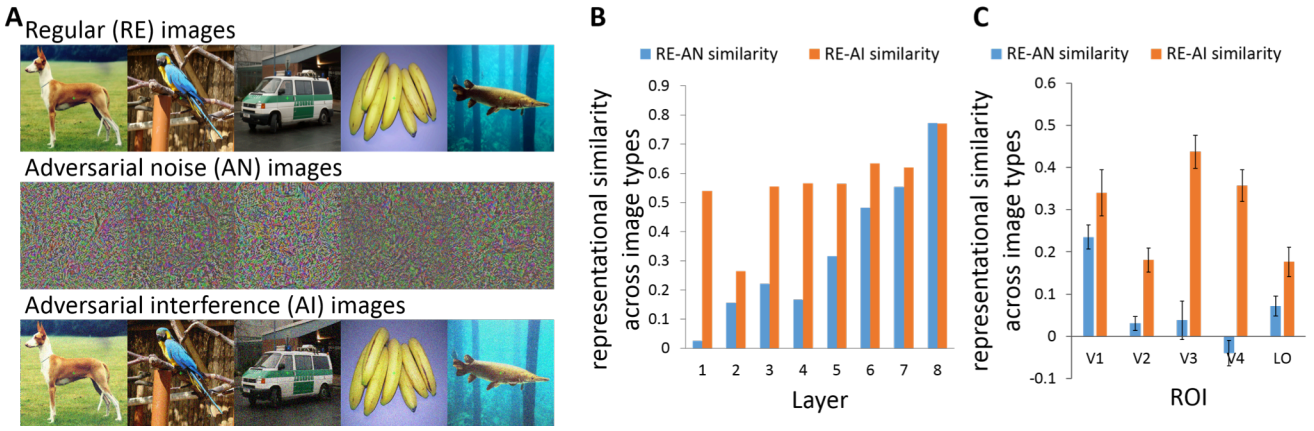

Figure 1. A) Image stimuli. Panel A provides a few example regular (RE, upper row) images, the adversarial noise (AN, middle row) images and the adversarial interference (AI, bottom row) images respectively. Humans can easily recognize AI but not AN images, whereas DNNs can recognize AN images with over 99% confidence but not AI images. B) Representational similarity between three types of images in the DNN. C) Representational similarity between three types of images in the human brain. The similarity values and errorbars are calculated via bootstrapping using all voxels in a region. Errorbars represent the 95% confidence interval of the correlation value. The similarity of responses between RE and AI images are significantly higher than that between RE and AN images in all ROIs (t-test, $P < 0.0001$).

began from the ImageNet mean (plus small Gaussian noise to break symmetry) and continued until the DNN confidence for the target class reaches 99%.

The AI images were generated by the similar optimization approach that began from the corresponding RE images.

In summary, we used a total of 120 images, 40 for each image type.

MRI experiments. The main experiment for each subject included two scanning sessions, with 5 runs in each session. A scanning run contained 120 stimulus trials and 9 blank trials. Within a stimulus trial, a blank and an image (12°× 12°) were alternately presented for 2s. A fixation point (0.2° ×0.2°) was shown at center-of-gaze throughout the entire run. A 20s blank period was added to the beginning and the end of each run, respectively. Subjects were instructed to maintain steady fixation throughout the entire run and press buttons to perform an animal recognition task–whether the object in an image belongs to animals. In addition, five visual regions of interest (i.e., V1-V4, LO) were localized in another retinotopic experiment and another functional localizer experiment.

Representational similarity analysis. We used the Alexnet (Krizhevsky et al., 2012) as our DNN and calculated the representational geometry for the three types of images. Between every pair of images in a same image type, we computed dissimilarities (1 - Spearman's correlation) of the activation of all units in each layer. This yielded a 40 (object categories) × 40 representational dissimilarity matrix (RDM) for each image type and in each DNN layer. We then correlated the RDM of the RE images and the RDM of the AN images, the resulted correlation denoted as "RE-AN similarity". Same calculation was repeated to obtain "RE-AI similarity". These analyses were performed for all DNN layers (Fig1. B).

We used the similar method to calculate representational similarity between image types in the human brain. 300 voxels (100 voxels from each subject) were selected in each ROI based on the correlation of response patterns towards a same picture across trials. Similar "RE-AN similarity" and "RE-AI similarity" values were calculated using the same method described above, except that we used voxels instead of artificial neurons (Fig1. C).

Results

For the DNN, the RE and the AN images are "perceptual" very similar but vice versa with respect to the RE and the AI images, even though the later pair is truly more comparable with respect to image features. If the representation in the DNN mainly reflects the input visual features, the RE-AI similarity will be higher than the RE-AN similarity since the RE and the AI images are more similar in image features. On the contrary, if representation in the DNN follows the "perceptual" outcome rather than image input, the RE-AN similarity should be higher than the RE-AI similarity.

Result showed that the RE-AI similarity values are generally higher than the RE-AN similarity values across different layers (t-test, $P = 0.0084$), indicating that visual representation in the DNN is more consistent with image input rather than perceptual output (i.e., classification labels). The same result occurs in the human visual system, which is not surprising given that RE and AI images look much more similar than RE and AN images to human vision.

We are also interested in how representational similarity changes across different layers in the DNN and regions in the brain. We noticed that the RE-AN similarity increases from low- to higher-level layers (Fig1. B) in the DNN and decreases from low- to higher-level visual areas in the brain (Fig1. C). Presumably humans cannot extract any semantic information from the AN images as such the neural response in the brain should reflect mostly bottom-up processing. The decreasing similarity between the RE and the AN images in the brain suggest that two types of images share more low- rather than high-level features. If we believe this is the ground truth, the opposite trend in the DNN reveals that the DNN misses some fundamental aspects of hierarchical visual representation. Also, given that all brain regions here are believed to process low- or middle-level visual features, this result also implies that the DNN might not satisfactorily account for some basic visual analyses in the human brain, although many computer vision studies claim DNNs are capable of doing so. Taken together, our results highlight the need for achieving not only high performance but also accurate representation in future practice of developing DNNs.

Acknowledgments

This work was supported by the National Key R&D Program of China under grant 2017YFB1002502 and the National Natural Science Foundation of China (No. 61701089, No. 61601518 and No. 61372172).

References

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports* 6**,** 27755.

Güçlü, U., and Van Gerven, M.A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience* 35**,** 10005-10014.

Güçlü, U., and Van Gerven, M.a.J. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage* 145**,** 329-336.

Hong, H., Yamins, D.L., Majaj, N.J., and Dicarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience* 19**,** 613-622.

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications* 8**,** 15037.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). "Imagenet classification with deep convolutional neural networks", in: *Advances in neural information processing systems*), 1097-1105.

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images", in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: IEEE), 427-436.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Yamins, D.L., and Dicarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* 19**,** 356-365.