Comparing Brain-Score and ImageNet performance with responses to the scintillating grid illusion

Martin Kent Kraus (mkkraus@email.cz)

Lucy Verkerk (lucy.verkerk@ru.nl)

Sander Keemink (sander.keemink@donders.ru.nl)
Department of Machine Learning and Neural Computing, Donders Institute for Brain, Cognition and Behaviour,
Radboud University, Nijmegen, The Netherlands

Abstract

Perceptual illusions are widely used to study brain processing, and are essential for elucidating underlying function. Successful brain models should then also be able to reproduce these illusions. Some of the most successful models for vision are several variants of Deep Neural Networks (DNNs). These models can classify images with human-level accuracy, and many behavioral and activation measurements correlate well with humans and animals. For several networks it was also shown that they can reproduce some human illusions. However, this was typically done for a limited number of networks. In addition, it remains unclear whether the presence of illusions is linked to either how accurate or brain-like the DNNs are. Here, we consider the scintillating grid illusion, to which two DNNs have been shown to respond as if they are impacted by the illusion. We develop a measure for measuring Illusion Strength based on model activation correlations, which takes into account the difference in Illusion Strength between illusion and control images. We then compare the Illusion Strength to both model performance (top-1 ImageNet), and how well the model explains brain activity (Brain-score). We show that the illusion was measurable in a wide variety of networks (41 out of 51). However, we do not find a strong correlation between Illusion Strength and Brain-Score, nor performance. Some models have strong illusion scores but not Brain-Score, or vice-versa, but no model does both well. Finally, this differs strongly between model types, particularly between convolutional and transformer-based architectures, with transformers having low illusion scores. Overall, our work shows that Illusion Strength measures an important metric, which is important to consider for assessing brain models, and that some models could still be missing out on some processing important for brain functioning.

Keywords: Perceptual illusions; Brain-Score; scintillating grid illusion; representational similarity;

Introduction

The presence of visual illusions could indicate specific statistical assumptions about the world made by our brains (Gregory, 1980; Tyler, 2022; Palmer & Rock, 1994). A good model of the brain should then not only be able to predict brain responses accurately, but also share the assumptions which lead to the presence of illusions. Many illusions have previously been studied in DNNs (Watanabe et al., 2018; Ward, 2019; Ngo et al., 2023; Zhang & Yoshida, 2024). However, most are focused on a few specific models, making broader comparisons across models difficult. This in turn makes it more difficult to determine the mechanisms and causes of the illusion in DNN's. In this study, we focus on the Scintillating Grid illusion (as was done in Sun & Dekel (2021)), where black circles appear at the center of the white disks in our peripheral vision as our gaze shifts. We use a model and illusion agnostic measure to determine the presence of illusion-like responses. We compared 51 networks from various architectural families, and determined whether the presence of illusions was associated with better model performance or increased brain similarity.

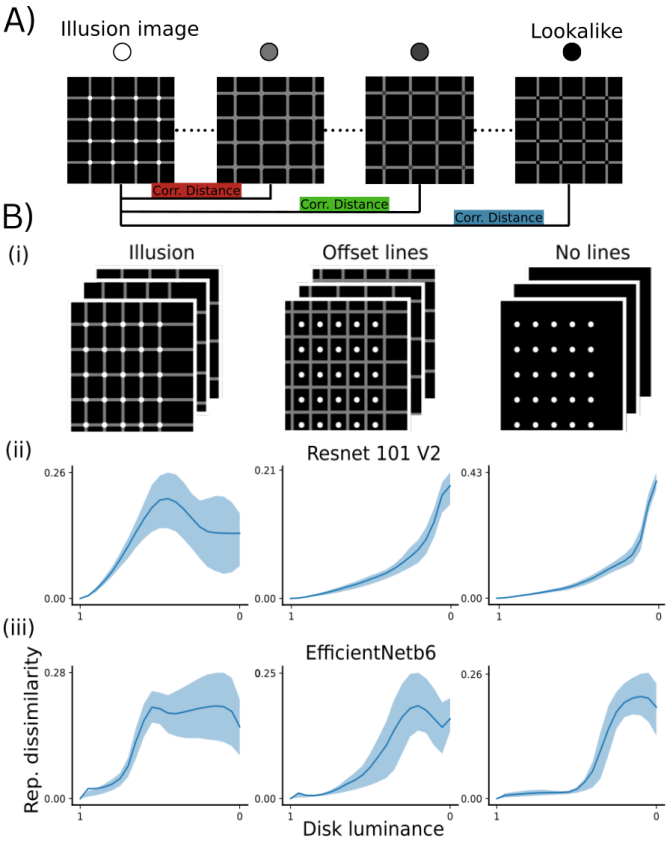

Figure 1: Methodology of measuring illusion responses A) Images are created by transitioning the color of the circles from white to black. Each image is then presented to the network, and the Pearson correlation distance between the white disk image and all other images is computed. B) Example average responses across different image sets. (i) From left to right, example illusion and control images. Controls are needed to check against black and white circles simply being represented similarly by the model. (ii) Average responses for Resnet 101's penultimate layer across the images. The mean (curve) and inter-quartile (areas) are shown. Here, the network has a non-monotonic response to the illusion images, but a fully monotonic response to the controls, indicating an illusory response. (iii) Same as (ii), but for EfficientNetB6. In this case there might be a weak illusory response to the illusion images, but also to the control images, as all curves are non-monotonic.

Methods

Generating images We retrieved both the control and illusion images from Sun & Dekel (2021). These images varied in the size of the grid, the size of the disks and offset from the center. The control images included the 'No lines control'

where the lines of the grid are absent, and the 'Offset lines control' where the disks are moved from the intersections of the grid to the center of the four surrounding lines. All white "illusion" and black "lookalike" dot images were blended together in steps of 5%, as can be seen in Figure 1

Measuring Illusion strength The internal representations of each network were retrieved from the penultimate layer of each network. We take the representational dissimilarity between the white dot image and compare it via correlation distance to all other images (Fig. 1 A). Across images this results in a set of curves showing the change of dissimilarity across luminances (Fig. 1 B). Afterwards, we compute the deviation magnitude as the total change from the highest point in the curve with all the following points as the presence of this non-monotonicity indicates a higher similarity between darker disks and white disks than those in-between (Fig. 1 B), such that

$$R_i = \sum_{l > l_{\text{switch}}} (C_i(q_l) - C_i(q_{l_{\text{switch}}})), \quad (1)$$

where $C_i(q_l)$ is the correlation distance for image type i (e.g. control or illusion) and the l'th image with luminance q_l, $q_{l_{\text{switch}}}$ is the switching point at which images start to become less similar compared to the highest point in the curve (Fig. 1B). Finally, we measure the Illusion strength for a given model as the average deviation magnitude on the illusion images minus the average deviation magnitudes on the control images ($I = \bar{R}^{\text{illusion}} - \sum_n^N \bar{R}_n^{\text{control}}$). Bonferonni corrected one sided Mann-Whitney U tests were performed to test significance.

Models studied Our study was conducted on 51 models. These were chosen from among those listed on Brain-Score.org (Brain-Score, 2018). The families of models chosen can be seen in Fig. 2. For each model, we retrieved a Brain-Score (Schrimpf et al., 2018) which is an amalgamation of metrics evaluating the models behavioral and engineering similarity to the brain. In addition, the ImageNet top-1 scores were taken to indicate the accuracy of the model.

Results

When comparing the behaviour of models to the illusions and controls, 41 out of 51 networks had statistically significant illusion-like responses. The significant models included VGG 19 and Resnet 101, which were also studied by Sun and Dekel (Sun & Dekel, 2021), by which we replicate their results. We found a weak negative correlation between Brain Score and Illusion strength ($r = -0.15$) (Fig. 2A), and no relationship between ImageNet top-1 performance and Illusion strength ($r = -0.06$) (Fig. 2C). Interestingly, the ImageNet top-performing model classes (CViTs and ResNeXts) both have very low Illusion strength. Models in the VGG, ResNet, SqueezeNet and CViT families all showed robust illusion-like responses. Of the 10 non-convolutional Transformer models studied, 7 did not have strong illusion-like responses.

In addition, we investigated how different networks represented the Scintillating Grid across all disk luminances using Representational Dissimilarity Matrices (RDMs) (Fig. 2 C). We found that despite most networks showing illusion like-responses, the internal representations varied significantly, showing that different representations are also possible even if not consistent with the illusory percept expected from human psychophysics. While VGG19 shows the type of RDM pattern one might expect the illusion to produce (with a cross-like pattern indicating cross-similarity between images with respectively lighter and darker disks), further research is needed to determine which of the RDM patterns truly corresponds to human perception.

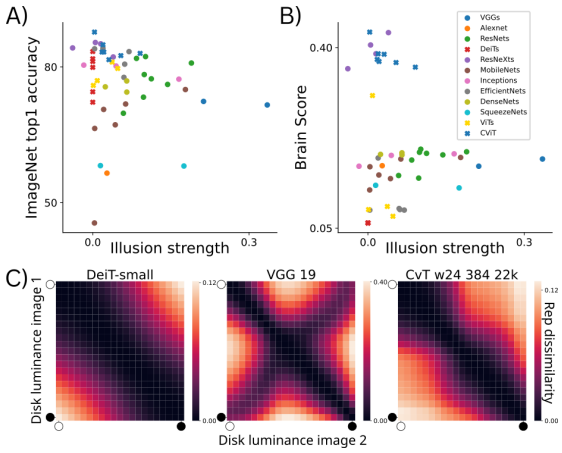

Figure 2: Comparing illusion strength with model performance and brain predictivity, and exploring network representations. (A) Comparing illusion strength to ImageNet performance. Each color corresponds to a different model class (as indicated in the legend), and crosses correspond to transformer networks. (B) Comparing Illusion strength to Brain-Score. (C) Representational Dissimilarity Matrices (RDMs) of representative models on all illusion images. Each row and column represent a different disk luminance. The brighter the color, the more dissimilar the representation. The top row corresponds to the curves drawn in Fig. 1 B, as it compares images with white disks to all other luminances.

Discussion

We conducted the largest to date study of illusions in DNN's, and showed that the vast majority of networks show illusion like responses. We found a significant difference between transformer and convolution based architectures. In models with positive illusion strength, different patterns of representations were found, indicating that models may perceive the Scintillating grid illusion in significantly different ways. Our method opens the way for further studies into other illusions, and for determining which DNNs perceive the world the most similarly to us. We improved Sun & Dekel (2021) method by making explicit use of the control sets. The strength of the illusion does not correlate with either Brain-Score nor performance, indicating that Brain-Score may be incomplete.

References

Brain-Score. (2018). *Brain-Score Vision.* https://web.archive.org/web/20241210140031/https://www.brain-score.org/vision/.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *290*(1038), 181–197.

Ngo, J., Sankaranarayanan, S., & Isola, P. (2023). Is clip fooled by optical illusions? *ICLR 2023*.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic bulletin & review*, *1*(1), 29–55. (Publisher: Springer)

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007. (Publisher: Cold Spring Harbor Laboratory)

Sun, E., & Dekel, R. (2021). ImageNet-trained deep neural networks exhibit illusion-like response to the Scintillating grid. *Journal of Vision*, *21*(11), 15–15. (Publisher: The Association for Research in Vision and Ophthalmology)

Tyler, C. W. (2022). The nature of illusions: A new synthesis based on verifiability. *Frontiers in Human Neuroscience*, *16*, 875829.

Ward, E. J. (2019). Exploring perceptual illusions in deep neural networks. *BioRxiv*, 687905. (Publisher: Cold Spring Harbor Laboratory)

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, *9*, 345. (Publisher: Frontiers Media SA)

Zhang, H., & Yoshida, S. (2024). Exploring Deep Neural Networks in Simulating Human Vision through Five Optical Illusions. *Applied Sciences*, *14*(8), 3429. Retrieved 2024-10-31, from https://www.mdpi.com/2076-3417/14/8/3429 doi: 10.3390/app14083429