Humans cannot decipher adversarial images: Revisiting Zhou and Firestone (2019)

Marin Dujmović (marin.dujmovic@bristol.ac.uk)
Gaurav Malhotra (gaurav.malhotra@bristol.ac.uk)
Jeffrey S. Bowers (j.bowers@bristol.ac.uk)

School of Psychological Science, University of Bristol, 12a Priory Road
Bristol, United Kingdom

Abstract:

In recent years, deep convolutional neural networks (DCNNs) have shown extraordinary success in object recognition tasks. However, they can also be fooled by adversarial images (stimuli designed to fool networks) that do not appear to fool humans. This has been taken as evidence that these models work quite differently than the human visual system. However, Zhou and Firestone (2019) carried out a study where they presented adversarial images which fool DCNNs to humans and found that, in many cases, humans chose the same label for these images as DCNNs. They take these findings to support the claim that human and machine vision is more similar than commonly claimed. Here we report two experiments that show that the level of agreement between human and DCNN classification is driven by how the experimenter chooses the adversarial images and how they choose the labels given to humans for classification. Based on how one chooses these variables, humans can show a span of agreement levels with DCNNs; from well below to well above levels expected by chance. Overall, our results do not support a view of large systematic overlap between human and computer vision.

Keywords: deep neural networks; adversarial examples; computer vision; object recognition

Introduction

Deep convolutional neural networks (DCNNs) boast human and even super-human levels of performance in some object recognition tasks. At the same time, these models are vulnerable to adversarial images that are confidently misclassified. Figure 1 shows examples of adversarial images which cause networks to misclassify with a high level of confidence. In addition to being important for security reasons (e.g. in systems like self-driving cars or online security), these types of stimuli are important because they can provide insight about the representations learned during training and about how networks function. Adversarial examples like the one in Figure 1 have also been taken by many to indicate that DCNNs rely on very different features to perform classification compared to human vision. Recently however, researchers have turned this argument on its head and argued that adversarial images may reveal important similarities between DCNNs and human vision (Elsayed et al., 2018; Zhou & Firestone, 2019).

Here we focus on the Zhou & Firestone (2019) study that provides evidence that humans can often decipher DCNN classifications of adversarial images which are thought to be unrecognizable. They take these findings as evidence of important similarities between DCNNs and human vision. We show that their findings are misleading and depend on several important decisions about how the experiments were conducted and results analyzed. To give one example, in Experiment 3a, Zhou and Firestone (2019) instructed to complete 48 trials where they were instructed to choose one of 48 labels to classify each image. The reported result shows that 88% of the participants agree with DCNN classification at a higher than chance level. Although this sounds impressive, it is important to note that if a participant agreed with DCNN classification on more than one out of the 48 trials he or she counted towards the 88% (2/48 is above chance), even though they could have disagreed with the DCNN classification on 46 out of 48 trials. When we re-analyzed their data, we observed the average agreement between participants and DCNNs was under 5/48 images (10.12%). Similarly, they report that for 79% of the images more participants agreed with DCNN classification than would be expected by chance. Again, chance would be approximately 4 out of 200 participants, so images for which 5 or more participants agreed with DCNN classification counted towards the 79%. We were interested in finding out how the choice of adversarial images and the choice of response labels (categories) affects these results. In order to do this, we carried out two experiments.

Experiment 1

Methods

Following Zhou & Firestone (2019), we used stimuli from Nguyen, Yosinski, and Clune (2015) to create three experimental conditions (examples of images can be seen in Figure 1). Condition 1 was composed of adversarial images designed to be classified as ImageNet categories that were generated by an

"indirect encoding" method that ensures that images have regular features (e.g. edges) that often repeat; we call this condition "regular". Condition 2 was composed of images designed to be classified as ImageNet categories but with a different encoding method, resulting in noise-like images; we call this condition "noisy". Finally, condition 3 was composed of images designed to be classified as MNIST digits (handwritten numbers from 0 to 9); thus, the condition is labeled as "MNIST". In all cases, the models identified these images 100% of the time with over 99% confidence.

We chose 10 classes in conditions 1 and 2 (from 1000 ImageNet classes) such that these classes shared no obvious features. This should make it easier for humans to correctly decipher DCNN classification than if the alternative classification choices were categories which overlapped in visual features with the DCNN target class (e.g. computer keyboard and remote control). All the chosen classes and the images used in condition 1 were among the 48 used in Experiment 3a of Zhou and Firestone (2019). We included all 10 digit classes for condition 3. Each condition was a block of trials in the experiment with the order of images within each block randomized for each participant as well as block order.

The study was designed in Gorilla and conducted online through the Prolific platform. In each trial participants (N = 200) were shown a single image and ten labels beneath it. Their task was to choose the label they thought best represented what was on the image. Each image was presented for a maximum of ten seconds after which it was removed and only the labels remained. Participants could make their choice at any time during stimulus presentation. After making their choice, participants indicated a level of confidence on a 0-100% scale. Confidence levels were introduced as an additional metric which provides context for raw agreement levels. DCNN classification is characterized with high confidence but we expected low levels of confidence in this behavioral study.

Results

Two participants were removed from the analysis as their response times were below 500 ms. Average levels of human-DCNN agreement and confidence ratings can be seen in Figure 2. Two one-way repeated measures ANOVAs were conducted to determine the effect of experimental condition on agreement and confidence levels. There was a significant effect of experimental condition on both agreement ($F(2, 394) = 135.84$, $p < .01$, $\eta_p^2 = .41$) and confidence ($F(2, 394) = 24.90$, $p < .01$, $\eta_p^2 = .11$). For agreement levels all pair-wise comparisons (Tukey HSD) were significant. As can be seen in Figure 2, participants showed highest levels of agreement with DCNNs in the regular, and least in the MNIST condition. Confidence ratings were higher in the regular and noisy conditions when compared to the MNIST condition, but a long way below DCNN confidence of 99.6% in all three conditions.

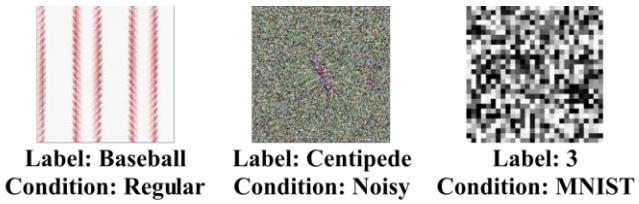

Figure 1: Examples of images in the three experimental conditions from Experiment 1.

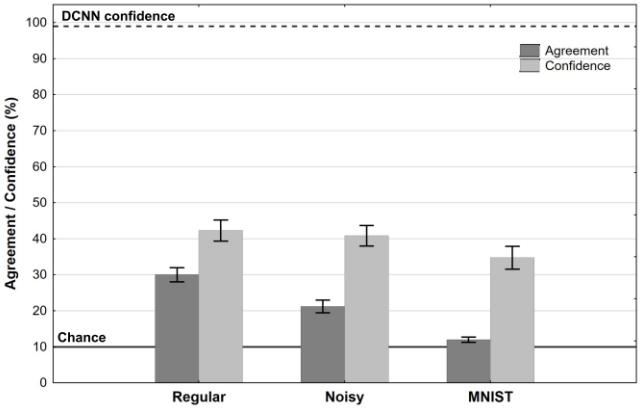

Figure 2: Agreement (average percentage of images on which a participant's choices agree with the DCNN) and confidence as a function of experimental condition in Experiment 1 (spreads show 95% CL).

We also found large variability within each condition reflected in a large range and between-image inconsistencies (Figure 3). In the regular condition participants agreed with DCNN classification at above chance levels for 70% of the images. However, agreement level ranged from 5.05% to 58.59% based on the image shown to the participant. Similar variability can be seen for the noisy (2.52% to 53%) as well as MNIST (4.55% to 35.86%) conditions. The lowest level of variability can be seen in the MNIST condition since for most of the stimuli agreement levels hover around chance.

Experiment 2

The between-condition (Figure 2) as well as between-item (Figure 3) variability in Experiment 1 suggests that the average agreement between humans and DCNNs provides only a coarse outline of a more complicated picture of the factors influencing agreement levels. If humans could indeed "decipher" how the DCNN is performing classification, we would have observed a smaller variance in agreement. We suspected that the relatively large agreement on some images could be

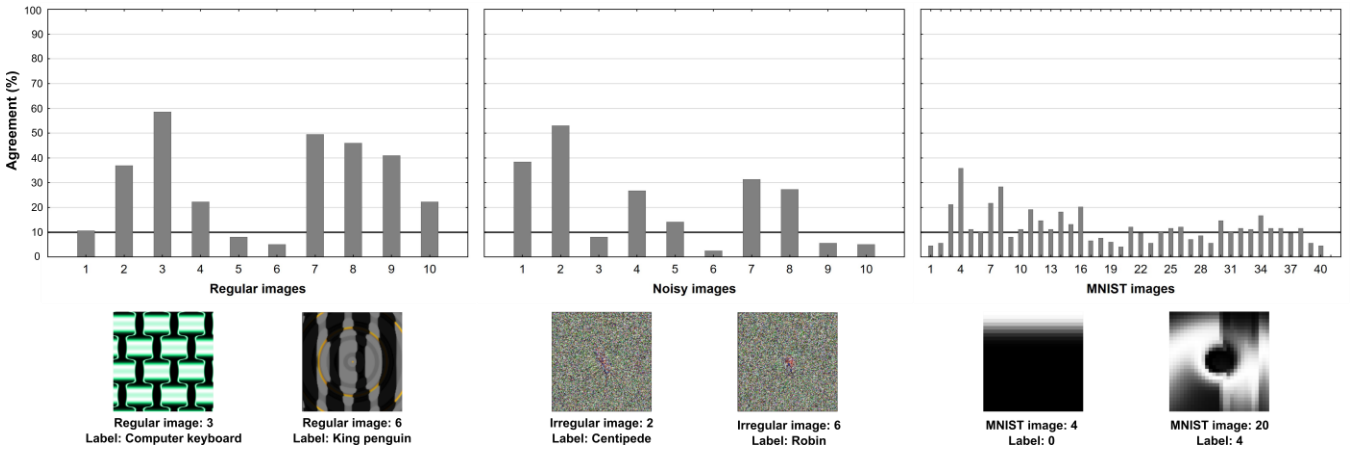

Figure 3: Item agreement levels (percentage of participants that agree with DCNN classification for a particular image) and two images with the highest and lowest agreement (bottom) from each condition.

largely due to the adversarial images containing some key features (red stripes in the regular image in Figure 1) of the target class (baseball) rather than the ability of humans to "decipher" how DCNNs were making these choices. If this was the case, then the average agreement could be exaggerated due to the choice of adversarial images and response alternatives.

Consider the choice of response alternatives first. Note that the ideal experiment should contain all 1000 response alternatives that were given to the DCNN. However, that is infeasible. Therefore, human studies limit the number of response alternatives. But limiting the number of choices may be biasing the results (exaggerating the agreement) by eliminating competing choices that would have been present for the model, especially since we specifically picked these alternatives to minimize overlapping features. If that was the case, then introducing other competing categories should reduce the level of agreement. Alternatively, if participants really can decipher how DCNNs are making the decision, introducing these alternatives should not affect agreement as the DCNN makes these choices with a confidence of 99.6% even in the presence of all 1000 response alternatives. We tested this hypothesis in Experiment 2A.

Now consider the choice of adversarial images. If the participant can anticipate how the DCNN performs classification, they should be able to do this for a wide array of adversarial images that fool a DCNN (again with a confidence of 99.6%). Instead the large variability indicates that participants may only able to do it for a subset of images. To test whether this was systematic and dependent on overt features present within the adversarial image, we hypothesized that we should be able to increase or decrease agreement between human and DCNN classification by choosing images from within the same class that did / did not contain overt features of the target class. We tested this hypothesis in Experiment 2B.

Method

In Experiment 2A we tested 100 participants using the same images as in the regular condition of Experiment 1 but changed the response categories shown to participants. In addition to the label assigned to an image by DCNNs, we showed three labels for categories which, subjectively, have features retained in the image (resemble the image). In total, each image was accompanied by four labels and the labels were different for each image. The experiment was conducted online and only categorization responses were requested of participants.

In Experiment 2B we tested 201 participants using regular images like the ones in Experiment 1. However, we now replaced some of these images with alternative images obtained using the same evolutionary algorithm proposed by Nguyen et al. (2015) and used to generate the images in Experiment 1. This experiment consisted of two conditions: a best-case condition, which contained adversarial images that, subjectively, seemed to contain overt features from objects in the target class and a *worst-case* condition which did not seem to contain such features.

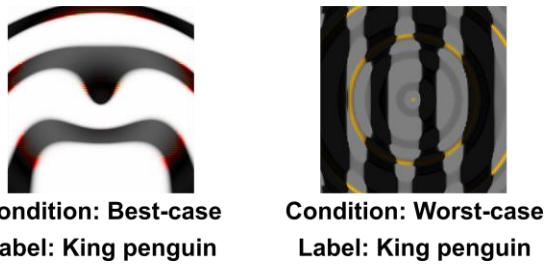

Figure 4: Example of two same-category images from the two experimental conditions in Experiment 2B.

Examples of images from each condition are shown in Figure 4. It is important to emphasize that DCNNs were equally confident (>99%) in classifying images from both conditions. The target categories remained the same as in Experiment 1. Participants in both experimental groups were shown the same category labels and were instructed to choose the label that best represents these images.

Results

As we had predicted, in Experiment 2A (N = 100) the average agreement level drops significantly to near chance levels when the response categories were made more competitive. We observed an agreement of 28.5% (SD = 11.67) with chance being at 25% (Figure 5). The agreement is statistically above chance levels ($t(99) = 3.00$, $p < .01$) but nowhere near the level observed in Experiment 1 (Figure 2).

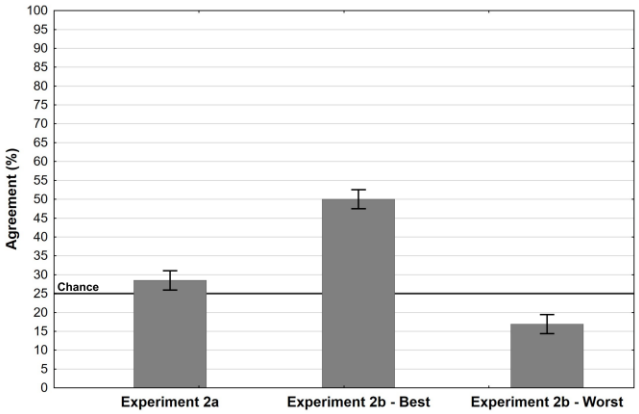

Figure 5: Average levels of agreement in Experiment 2 (spreads show 95% CL).

Results from Experiment 2B (N=101 in best-case and N=100 in worst case) showed that there was a significant difference in the level of agreement with DCNN classification between the two conditions ($t(199) = 17.41$, $p < .01$). The level of agreement in the best-case group was far above the one in the worst-case group (Figure 5). The best-case group was significantly above chance ($t(100) = 16.08$, $p < .01$) while the worst-case was significantly below chance ($t(99) = 7.44$, $p < .01$).

Discussion

Consistent with our re-analysis of the results from Zhou & Firestone (2019), we observed relatively low participant-DCNN agreement in classification of adversarial images. In contrast to DCNNs, participants also showed low levels of confidence in their choices. While participants agreed with neural networks at a higher than chance level in all three experimental conditions of Experiment 1, the most striking finding was how variable the results were with a substantial percentage of responses below chance. When we manipulated response choices and the nature of the generated adversarial images in Experiments 2 we could systematically increase or decrease agreement between the participants and the DCNNs, such that it was easy to construct conditions in which the model was highly (>99%) confident that a given adversarial image belonged to one category whereas humans systematically categorized the image as a member of another category. Our results suggest that that the (limited) agreement between human and DCNN classification occurs when there is more overlap in the visual features of the adversarial image and target category compared to the adversarial image and foil categories. This is very different than claiming that humans can reliably decipher how DCNNs perform classification.

In sum, we take our findings to highlight the differences between the human visual system and DCNNs. These differences are also observed across a range of additional adversarial conditions, including cases in which models confidently misclassify images when the color, orientation, context in which the image is presented are changed, and indeed, when a single pixel in an image is changed (for a review see Akhtar & Mian, 2018). Researchers claiming that the DCNNs provide the best theory yet of human vision need to explain why these models behave so differently than human observers.

Acknowledgments

This research was supported by the European Research Council Grant *Generalization in Mind and Machine,* ID number 741134.

References

Akhtar, N., & Mian, A.S. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access, 6*, 14410-14430.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems* (pp. 3910-3920).

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 427-436).

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature communications*, *10*(1), 1334.