Mooney Face Image Processing in a Deep Convolutional Neural Network Compared to Humans

Astrid Zeman (astrid.zeman@unimelb.edu.au)
Melbourne Centre for Data Science, University of Melbourne,
700 Swanston St, Carlton, VIC 3053 Australia

Tim Leers (tim.leers@kuleuven.be)
Brain & Cognition, KU Leuven
Tiensestraat 102, Leuven 3000 Belgium

Hans Op de Beeck (hans.opdebeeck@kuleuven.be)
Brain & Cognition, KU Leuven
Tiensestraat 102, Leuven 3000 Belgium

Abstract:

Deep Convolutional Neural Networks (CNNs) are criticised for their reliance on local shape features and texture. We test whether CNNs process global shape information in the absence of local shape cues and texture by testing their performance on Mooney stimuli, which are face images thresholded to binary values. More specifically, we assess whether CNNs detect faces in these abstract stimuli, and whether they exhibit the face inversion effect (FIE), where upright stimuli are classified positively at a higher rate compared to inverted. We found that a standard face recognition network (DeepFace) performs perceptual completion and exhibits the FIE, which is present over all levels of specificity. By matching the false positive rate of DeepFace to humans, we found that the network performed closer to the human average for both conditions (62.70% for upright and 42.26% for inverted). Rank order correlation between DeepFace and humans across individual stimuli shows a significant correlation in upright and inverted conditions, indicating a relationship in image difficulty between observers and the model. We conclude that despite the texture and local shape bias of CNNs, which makes their performance distinct from humans, they are still able to process face images holistically.

Keywords: faces; convolutional neural networks; texture bias; Mooney stimuli

Introduction

Despite their success in object recognition, deep convolutional neural networks (CNNs) reveal a shortcoming in processing global over local shape information (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos, et al., 2019; Brendel & Bethge, 2019; Xu, et al., 2018). Baker *et al.* (2018) found that the classification performance of CNNs was resilient to disruptions in global shape, but susceptible to local contour changes, an effect opposite to that of human observers. Geirhos *et al.* (2019) assessed network classification using a cue conflict paradigm, where the texture of one object was superimposed over another object, and CNNs demonstrated a preference for object texture over shape, in contrast to humans. Xu *et al.* (2018) found that CNNs trained to identify faces are more sensitive to changes in texture compared to shape. These studies provide mounting evidence for an emphasis on local processing measures in CNNs, making their performance distinct from humans.

To interrogate the level of local versus global processing in CNNs, we assess their performance using a large battery of Mooney images, which are low-information images of faces that invoke face perception despite their high ambiguity (Mooney, 1957). Mooney images contain only isoluminant black and white areas and therefore have no texture and reduced local shape information. Global shape information processing is necessary for perceiving faces in these two-tone images, in which an observer must be able to integrate incomplete segments into a perceptual whole, known as “perceptual closure” (Wertheimer, 1923). Recently, CNNs have been found to exhibit perceptual closure when trained on natural images (Kim, Reif, Wattenberg, & Bengio, 2019). CNNs that are randomly initialised or trained on random data do not exhibit closure, and closure ability is correlated with higher layers, peaking in the layer prior to prediction read-out (Kim, Reif, Wattenberg, & Bengio, 2019). Given that CNNs are capable of closure, and are able to detect faces (Le, et al., 2012), it is logical that they are able to detect faces within Mooney images, which has indeed been established in a set of Mooney images artificially generated by a CNN (Ke, Yu, & Whitney, 2017).

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0

Method

Stimuli

We use the Mooney face online dataset containing 96 scrambled, 504 upright and 504 inverted images from Schwiedrzik, Melloni, & Schurger (2018).

CNN

DeepFace is a serial network with 6 parameterised layers (4 convolutional followed by 2 fully-connected) and 3 max pooling layers that follow every convolutional layer (Chen, 2016). DeepFace performs face detection, alignment and recognition, and is trained on the CASIA-WebFace database. We assess positive performance of DeepFace using binary recognition.

Results

We first tested DeepFace in its ability to classify whether an image contains a face for each of the three conditions (upright, inverted and scrambled), shown in Figure 1. DeepFace was able to perform perceptual completion, showing upright performance as higher than scrambled, with a difference of 28.92%. DeepFace also exhibited an FIE, with upright performance being higher than inverted (23.41% difference). The network showed results qualitatively consistent with human performance – with both upright and inverted accuracy levels being above the false positive (scrambled) rate.

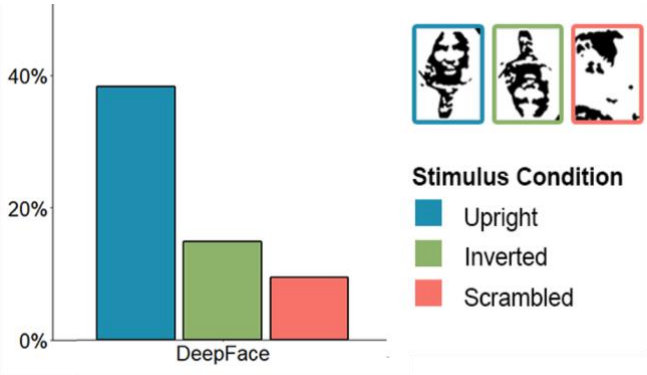


Figure 1: Face detection in DeepFace under three stimulus conditions (upright, inverted & scrambled)

Even though the CNN showed higher performance for upright (38.29%) versus inverted (14.88%) images, this was still below the average human response of 85.73% and 57.25% for these same images, however the false positive rate was also lower for the CNN (9.38%) compared to humans (27.36%). We therefore plotted an ROC curve to better determine the upright versus inverted performance when the false positive rate matched humans. ROC curves for upright (blue) and inverted (green) conditions are shown in Figure 2. For DeepFace, when the false positive rate was 27.08%, upright performance was 62.70% and inverted was 42.26%. Note that upright performance is always higher than inverted, demonstrating an FIE over all false positive levels. The size of the inversion effect remains roughly consistent across different levels of specificity.

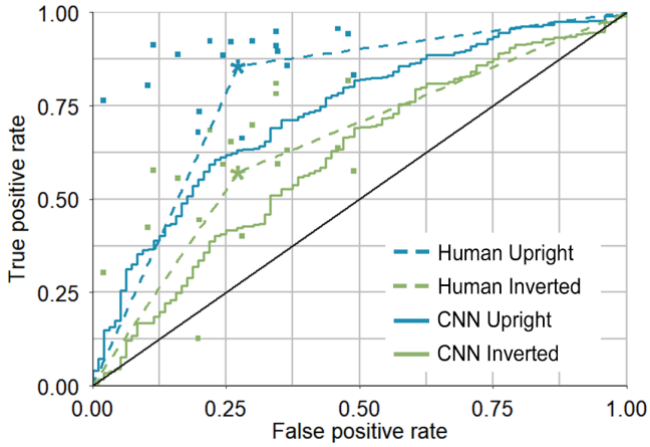


Figure 2: ROC for DeepFace (solid lines), plotted against human performance (dashed lines) for upright (blue) versus inverted (green) images. Asterisks are average human performance, squares are individuals.

Conclusions

In this paper we examined whether a standard CNN used for face detection is able to detect faces in ambiguous Mooney images. We first analysed network performance using the default discrimination levels and found that the false positive rate (using scrambled images) was lower than in human observers. We then examined face detection levels under differing levels of specificity, to determine levels of performance when the false positive rate was matched to humans. From these experiments, we present three main findings: 1. Deep networks are able to detect faces in Mooney stimuli, demonstrating robust detection in these low information images; 2. CNNs exhibit a face inversion effect with Mooney faces, which is evident across all degrees of specificity.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. PLoS Computational Biology, 14(12), e1006613.

Brendel, W., & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. International Conference on Learning Representations (ICLR).

Chen, R. (2016, 9 11). RiweiChen/DeepFace. Retrieved from GitHub: https://github.com/RiweiChen/DeepFace

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. International Conference on Learning Representations (ICLR).

Ke, T.-W., Yu, S. X., & Whitney, D. (2017). Mooney face classification and prediction by learning across tone. IEEE International Conference on Image Processing (ICIP).

Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do Neural Networks Show Gestalt Phenomena? An Exploration of the Law of Closure. arXiv, 1903.01069.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., … Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. ICML.

Mooney, C. M. (1957). Age in the development of closure ability in children. Canadian Journal of Psychology/Revue canadienne de psychologie, 11(4), 219–226

Schwiedrzik, C. M., Melloni, L., & Schurger, A. (2018). Mooney face stimuli for visual perception research. PLoS ONE, 13(7), e0200106.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. Psychologische, 4, 301–350

Xu, T., Zhan, J., Garrod, O. G., Torr, P. H., Zhu, S.-C., Ince, R. A., & Schyns, P. G. (2018). Deeper Interpretability of Deep Networks. arXiv:1811.07807