A human-like view-invariant representation of faces in deep neural networks trained with faces but not with objects

Naphtali Abudarham (naphtool@gmail.com)
School of Psychological Sciences, Tel Aviv University
Tel Aviv, 6997801 Israel

Galit Yovel (gality@tauex.tau.ac.il)
School of Psychological Sciences & Sagol School of Neuroscience,
Tel Aviv University
Tel Aviv, 6997801 Israel

Abstract:

Face recognition depends on the generation of a view-invariant representation. Faces are known to engage specialized mechanisms and it is therefore of interest to reveal to what extent is specific experience with faces necessary for the development of this representation. This question is hard to study in humans, but can be studied with Deep Convolutional Neural Networks (DNNs) trained with faces or with objects. To examine whether a face-trained and an object-trained networks generate a human-like, view-invariant representation, we first examined the similarity between the representations of faces across different head views. We then examined whether the networks use the same view-invariant facial features that are used by humans for face recognition. Our findings show that a human-like view-invariant representation of faces emerges at higher layers of a face-trained DNN, but not the object-trained DNN. The representations of faces were similar at lower layers of the face-trained and object-trained networks. These findings may resemble the face and object pathways in the human brain that are similar in low-level areas and diverge at higher levels of the visual cortex. They further imply that invariant face recognition depends on experience with faces, during which the system learns to extract face-specific, invariant features.

Keywords: face-recognition; deep convolutional neural networks; DNN; view-invariance; face-representation; object-recognition;

Introduction

Humans can recognize faces across variations in appearance, such as changes in pose, illumination and expression. This ability to generate a view-invariant representation of faces allows humans to discriminate between identities of different people, as well as to generalize across images of the same person, under large variations in appearances. Faces are known to engage specialized mechanisms that are not used for non-face objects. It is therefore of interest to assess to what extent human-like face representations depend on specific experience with faces or may emerge also from general experience with various object categories.

For obvious reasons, this question is hard to study in humans, as it requires testing individuals who had experience with non-face objects with no exposure to human faces. Therefore, we turned to test this question in Deep Convolutional Neural Networks (DNNs). DNNs have a brain-inspired hierarchical architecture and have exhibited human-level performance in face-recognition (e.,g, Taigman, Yang, Ranzato, & Wolf, 2014), in particular for faces with large variations in appearance. They can therefore serve as good models for human, view-invariant face recognition.

To study the role of experience using the DNN model, we examined the face-representations generated by DNNs trained to either recognize faces, or objects. In addition, we compared the representation of both networks to the representations generated by humans who perform similarity rating tasks on the same set of face stimuli. This approach enables us to examine whether a DNN that was trained to recognize face identities or to recognize object categories, develops a human-like view-invariant representation of faces.

Experiment 1: A view-invariant face representation in humans and DNNs

Methods:

Stimuli: To quantify view-invariance we used images of 15 identities from the color FERET face-image dataset (Phillips, Wechsler, Huang, & Rauss, 1998). For each identity we took 4 images: 1 – a frontal- image, hereby referred to as the "reference" image, 2 – a second frontal image, different from the "reference" image, hereby referred to as the "frontal" image, 3 – a quarter-left image, and 4 – a half-left image (see Figure 1 - top for example images). All face images were of adult Caucasian males, had adequate lighting, and had no glasses, hats or facial hair. The images were cropped just

below the chin to leave only the face, including the hair and ears.

Face-trained and Object-trained DNNs: To compare between the face representations generated by a face-trained DNN and an object-trained DNN we used two pre-trained state-of-the-art DNNs: OpenFace (Amos, Ludwiczuk, & Satyanarayanan, 2016), which was trained on the CASIA and FaceScrub datasets, containing approximately 500,000 images of approximately 10,000 identities, and the pre-trained Inception-V3 object-recognition network from the pytorch "model-zoo", which was trained on the ImageNet dataset. We used these two networks because they both have the same Inception-V3 backbone. The training sets of the face and object-trained DNNs contain many variable images for each label, to enable the DNNs to generate a view-invariant representation for faces or objects, respectively.

Extracting representations from DNNs: During training, DNNs are optimized to assign the correct labels to the images in the training sets, these labels being either identities, in the case of faces, or object categories, in the case of objects. Images are processed by DNNs through a hierarchy of computational layers, each layer performing some multiplication/convolution of the previous layer output, followed by some non-linear function and optional pooling. The final stage of processing is a series of fully connected layers, ending with an output layer which has a size equal to the number of labels in the training set. We ignored the labels and the output layer, and examined the representation of the last layer of the network. The last layer, before the output layer, is in fact the final feature-extraction performed by the network. Similarly, we examined the representation of any layer within the network to study how the representation evolves from lower to higher level layers.

Quantifying view-invariance of face-representations in DNNs: To quantify the view-invariance of the face-representations generated by the face-trained and the object-trained DNNs, we measured the L2 distances between the representations generated by the DNNs for face images in different views. To this end, we took the 15 reference images (see Stimuli section above), ran a forward pass on them through the DNN, and used the last layer of the DNN to obtain 15 feature-vector representations, one for each image. Then we measured the L2 distances between all possible 105 pairs of the 15 images, obtaining a vector of 105 distances. This was the reference distance-vector. We then repeated this process for each of the 3 other image types we had – the frontal, quarter-left and half-left images, thereby creating a total of 4 distance vectors. The distances in all vectors were in matching order of identities for all types of images. To quantify to what degree is the face-representation generated by the DNN view-invariant or view-dependent, we measured the Pearson correlation between the reference distance-vector and each of the 3 other distance vectors. The correlation between the reference distance-vector and the frontal distance-vector was taken as baseline since both vectors included distances between representations of frontal images of the same people. A view-invariant representation is reflected in similar correlations of all face views with the reference frontal view. A view-dependent representation is reflected in higher correlation with a frontal pair and lower correlations as the angle view increases relative to the frontal faces. These correlations between feature-vector distances were computed for the face-trained and object-trained DNNs.

Quantifying view-invariance of face-representation in humans: To compare the view invariance of face-representations in DNNs to that of humans, we used the same approach described in the previous paragraph, only now, instead of using distance-vectors, we collected image similarity ranking from human subjects. Each subject rated similarity between the 105 possible image pairs of one of the 4 image types on a scale of 1 (very different) to 6 (very similar) (as control we also added the 15 identical pairs, but they were later removed from analysis). In total we had 40 participants rating similarities, an average of 10 participants per image type. The final similarity vector was calculated by taking the mean similarity rating for each image pair across the participants. Next, we calculated the correlations between the similarity vectors, as explained above.

Results and Discussion

Figure 1 shows the correlations between distance-vectors based on representations generated by the last layer of the face-trained DNN (1B), humans (1C) and object-trained DNN (1D). The pattern of view-invariance of the face-trained DNN is similar to human results, showing invariance across frontal and quarter-left views, with lower generalization for half-left views ($p < .05$).

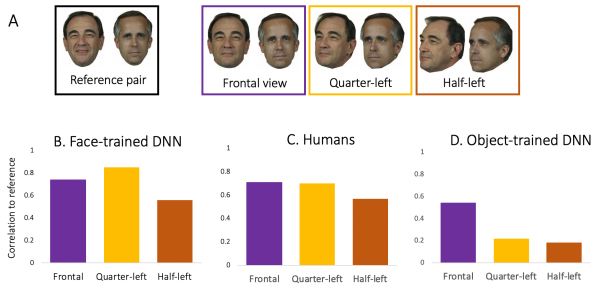


Figure 1: A human-like view-invariant face representation in face-trained but not object-trained DNNs

In contrast, the object-trained DNN shows overall lower correlations indicating worse generalization across different images of the same identities, and view-specific representations, indicated by much lower correlations for the quarter-left and half-left faces ($p < .005$).

In addition, we computed distance-vectors based on representations from all the layers of the networks, including the raw images, before any processing by the networks. We

found that the initial layers are similar in both the face-trained and the object-trained DNNs, indicating that these layers perform similar low-level processing in both networks. As we progress up the network we found that view-invariance emerges at higher layers of the face-trained, but not object-trained DNN (Figure 2).

The results of Experiment 1 indicate that a human-like view-invariant face representation emerges at the higher layers of the face-trained but not the object-trained DNN. These results indicate that during training to recognize faces, the face-trained DNN has learned to extract high-level facial features that allow it to generalize across head views. This view-invariant representation did not emerge in the object-trained DNN, which was trained to categorize objects, and has learned to extract different types of features that enable it to categorize different objects across different views. These features may also be view-invariant, but they are not adequate to generalize across faces in different views.

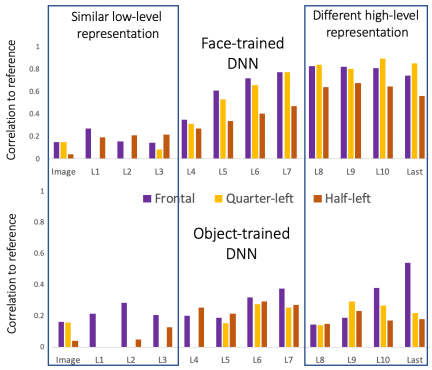


Figure 2: A human-like view-invariant face representation in higher layers of face-trained but not object-trained DNNs. Similar representations in low-level layers.

Experiment 2: View-invariant features in humans and DNNs

Given that the face-trained but not the object-trained DNN generated a human-like view invariant representation, in Experiment 2 we asked whether the face-trained network uses view-invariant features that are used by humans. In a series of previous studies (Abudarham & Yovel, 2016, 2018; Abudarham, Shkiller, & Yovel, 2019), we discovered a subset of view-invariant facial features that are used by humans to recognize faces (i.e., critical features). Here we tested whether the face-trained DNN, but not the object-trained DNN, uses the same set of view-invariant facial-features as humans use to recognize faces. To that end we measured the L2 distances between face-representations of 2 types of face-pairs: 1 - an original face vs. the same face in which we replaced the same critical features that humans use for face-recognition, and 2 – an original face vs. the same face in which we replaced non-critical features. If the face-trained DNN uses the same critical features as humans, then the distances between faces that differ in critical features will be larger than for faces that differ in non-critical features. We examined these distances in a previous study in a face-trained DNN and in humans (Abudarham & Yovel, 2016; Abudarham et al., 2019) and here we compared that to results of the object-trained DNN.

Methods

Stimuli: 25 faces were used to generate image pairs. For each of the 25 faces we had an original image, an image in which we replaced critical features (modified from the original image), and an image in which we replaced non-critical features (also modified from the same original image). In addition, we had a different not-modified image of that person, which we used as a reference image. Thus, we created 4 image pairs: 1 - Same pair – the reference image vs. the original image, 2 - Different pair – the reference image vs. a reference image of a different identity, 3 - Critical features pair – the reference image vs. the original image with different critical features, and 4 – Non-critical feature pair – the reference image vs. the original image with different non-critical features (See Figure 3A for example image pairs).

Measuring image similarity: We used the same face-trained and object-trained DNNs as in Exp. 1, and the same method for generating face representations from the face images. Image similarity was L2 distances between image representations.

Results and Discussion

Normalized similarity scores were calculated by dividing all the distances by the maximum distance across all pairs. (see Figure 3B-D). In the face-trained DNN (Fig. 3B), the similarity scores for Same identity pairs are very low compared with scores for Different identity pairs, as expected from a DNN trained to recognize faces. Similarity for the Critical feature change pairs are not significantly different from scores for Different pairs. This indicates that when changing Critical features in images, the same features that are used by humans to recognize faces, the face-trained DNN perceives these faces as different identities, meaning that the face-trained DNN is sensitive to the same critical features as humans. The similarity scores for non-critical feature changes were significantly smaller than for the critical feature change pairs ($p<0.001$), indicating that similar to humans, these changes did not cause a change in identity for the face-trained DNN. Figure 3C shows the human similarity scores on the same stimuli (Abudarham et al., 2019). The pattern of results in the face-trained DNN is similar to human results. Finally, we ran the same experiment using representations generated by the object trained-DNN (Fig. 3D). Here there are no significant differences in similarity scores for critical and non-critical features changes, indicating that the object-trained DNN is not sensitive to the critical features used by humans to recognize faces. ANOVA with DNN type (Face-trained, Object-Trained) and Face Type (Same, Low-PS,

High-PS, Different) revealed a significant interaction between these two factors ($p < .005$).

Fig. 4 shows the differences between distances for critical vs non-critical pairs across the layers of the face-trained and object-trained DNN. We see that the distances between original and critical-feature changes get larger than non-critical feature changes only in high layers of the face-trained DNN. This is similar to Exp. 1, indicating that the sensitivity

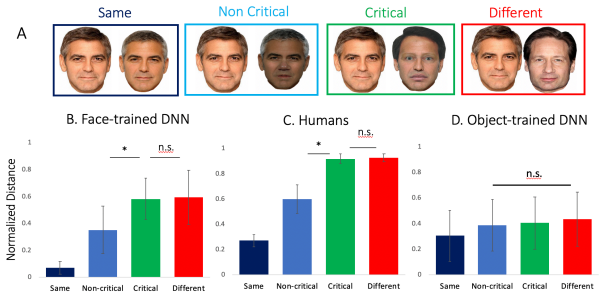
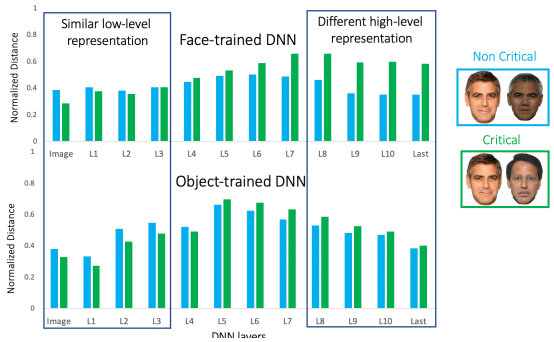


Figure 3: A Face-trained, but not object-trained DNN is sensitive to the same invariant features used by humans



to the critical features emerges at the higher layers of the face-trained but not the object-trained DNN (Figure 4).

Figure 4: Sensitivity to human-like view-invariant features (green>cyan) emerges at higher layers of the face but bot object trained network. Similar representation in lower-level layers.

Results of Experiment 2 indicate that the face-trained but not the object-trained DNN uses the same critical features as humans use for face recognition. As we found in previous studies on humans, these features are invariant to pose changes. Therefore, we suggest that these same features are used by the face-trained DNN to generate a view-invariant face representation. In addition, the sensitivity to these features emerges only at the higher layers of the face-trained network, suggesting that the network learns to extract these high-level features at later stages of processing, after training with large variability faces.

Conclusions

We found that the face-trained, but not the object-trained DNN, generates a human-like view-invariant face representation. We found that this representation emerges at higher layers of the face-trained network, while lower layers of both networks are similar. These finding may resemble the neural pathways for processing faces and objects, that are similar in the low-level visual areas, and diverge to dedicated modules in higher levels of the visual system. We also found that this view-invariant face representation relies on the same set of view-invariant facial features that humans use for recognizing faces. We suggest that humans and face-trained DNNs learn to use the same invariant features based on the experience with faces in variable views, and that this view invariant representation cannot be learned from experience with non-face objects.

References

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. Journal of Vision, 16(3). https://doi.org/10.1167/16.3.40

Abudarham, N., & Yovel, G. (2018). Same critical features are used for identification of familiarized and unfamiliar faces. Vision Research. https://doi.org/10.1016/j.visres.2018.01.002

Abudarham, Naphtali, Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. Cognition. https://doi.org/10.1016/j.cognition.2018.09.002

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). OpenFace: A general-purpose face recognition library with mobile applications. Retrieved from http://cmusatyalab.github.io/openface/

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 16(5), 295–306.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 1701–1708). IEEE.