Visual Expertise and the Familiar Face Advantage

Nicholas M. Blauch (blauch@cmu.edu)
Center for the Neural Basis of Cognition, Carnegie Mellon University
Pittsburgh, PA 15213 United States

Marlene Behrmann (behrmann@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 United States

David C. Plaut (plaut@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 United States

Abstract

Human expertise for recognizing unfamiliar faces has recently been called into question, highlighting a deficit when compared to familiar face recognition. We present simulations of a fixed-architecture deep convolutional neural network (DCNN) with different training regimens, highlighting the extent to which learning to recognize many "familiar" faces allows for robust, but incomplete, generalization to new "unfamiliar" faces as compared to performance after familiarization. With some training, verification performance for previously unfamiliar faces improves modestly, but the performance difference between unfamiliar and familiar faces is much smaller than the performance boost from pre-training on faces as compared to objects in the ImageNet 1000-way image classification database. We also assess the generalization performance of our networks to other fine-grained visual tasks such as bird species and car model verification. We find that expert face recognition does not improve generalization to birds or cars compared to a network trained on a subset of ImageNet with all vehicles and birds removed. We conclude that the specific learned statistics within a domain of visual expertise determine its generalization to other domains, in contrast with domain-general accounts which highlight level of processing over domain-specific statistics.

Introduction

Faces are perhaps the most important visual stimulus for humans. As such, adult humans develop substantial expertise with faces, allowing effortless recognition of a very large number of known individuals, along with recall of associated identity-specific semantic information. While it is clear that humans are experts at face recognition, many questions remain concerning the specifics of this expertise, such as whether innate mechanisms for face recognition exist, whether this expertise arises through general learning mechanisms common to those recruited when becoming a visual expert in other stimulus classes, and whether this expertise is specific to familiar faces. In particular, this last question has received substantial scholarly interest recently (Young and Burton, 2018)(Rossion, 2018)(Sunday and Gauthier, 2018), and has important societal implications, in particular in security settings (Young and Burton, 2018). The debate can be summarized as follows. It is often claimed without qualification that humans are experts at face recognition. However, a body of research has shown that humans are substantially worse at processing faces of unfamiliar than familiar individuals. In particular, face verification – determining whether two faces are of the same person identity – is worse for unfamiliar than familiar faces. As such, Young and Burton (2018) claimed that human expertise in face recognition is specific to familiar faces. Other work from these researchers has argued that unfamiliar faces are "not faces" and are rather processed as objects (Megreya and Burton, 2006). However, these claims have been received with sharp disagreement. Sunday and Gauthier (2018) argue that humans are experts at unfamiliar face recognition when compared to the appropriate baseline of general object recognition. Further, Rossion (2018) argues that humans are indeed experts at all forms of visual face recognition, and that what differs between familiar and unfamiliar face recognition lies at the level of semantics, claiming that such semantic information allows for further gains in face verification where discrimination at the perceptual level alone is noisy or more difficult.

While we agree with points made by each of these authors, yet a further alternative account exists: human shortcomings in unfamiliar face recognition might be attributed entirely to the nature of the visual task, rather than to a mechanism specifically designed for recognizing familiar faces. If the within-identity variability frequently exceeds the between-identity variability, and some of the within-identity variability is individual-specific, then unfamiliar face recognition will necessarily be worse than familiar face recognition. We sought to understand how much of the variability in natural face images can be learned by a generic face recognition mechanism, and how much of the variability must be learned for each individual exemplar. The claim that face recognition expertise is specific to familiar faces suggests that either the majority of variability is individual-specific, or the human recognition mechanism fails to capture important aspects of generic variability which would improve recognition of unfamiliar faces. To measure the generic and individual-specific variability of faces, and the extent to which different forms of learning can untangle these forms of variability for successful recognition, we adopted a machine learning approach using a standard convolutional neural network

(VGG-16) with multiple controlled training regimens.

We tested unfamiliar and familiar face recognition by holding out a set of identities during the training of the network; unfamiliar face recognition was performed after pre-training, and familiar face recognition was performed throughout and following fine-tuning on images of the previously unfamiliar identities. The generalization of our pre-trained network to unfamiliar identities serves as a metric of how much generic variability may benefit face recognition. The generalization gap between this performance and performance following familiarization provides a metric of how much individual-specific variability is required for successful verification.

Methods

We used the VGG-16 architecture as the deep convolutional neural network in our simulations. For pre-training, we started with either the ImageNet 1000-way classification database, or the VGGFace2 face recognition database, both containing over a million images. From ImageNet, we used a subset of approximately 600 categories for which entry-level labels were available (Ordonez et al., 2013), and then removed all bird and vehicle categories. We refer to the dataset with standard labels as "ImageNet" and the set with entry-level labels as "ImageNet-entry", notably containing the same images. For VGGFace2, we removed several identities which overlapped with other databases such as Labeled Faces in the Wild, leaving 8051 identities in a dataset we call "VGGFace2-full". We then determined a subset of VGGFace2-full which matched our ImageNet subset in total images, however with more categories, which we refer to as "VGGFace2". A portion of each database was held-out of training for validation. Pre-training was performed with stochastic gradient descent over the full network, using an initial learning rate of 0.1 which was allowed to decrease 5 times by a factor of 10 upon plateau in validation set performance, up to 50 epochs of training. Additionally, an untrained network was tested, with randomly initialized weights as preceding pre-training.

To test unfamiliar face recognition, the output layer was preserved with the same number of identity nodes as were required for pre-training (1000 for the random network). For testing familiar face verification, in the first epoch, we appended new identity units to the existing ones. Analyses and network fine-tuning were run utilizing the PyTorch neural network modeling package (Paszke et al., 2017) in the Python programming language.

To test familiar face recognition, we utilized the same images used in testing unfamiliar face recognition, but preceded testing with fine-tuning on a training set of images of the same identities. Here, fine-tuning refers to stochastic gradient descent back-propagated through the fully-connected layers only, with the weights of earlier convolutional layers held fixed. The network was not trained for face verification explicitly, but rather just for face identification on the new set of identities, using a cross-entropy loss. A fixed learning rate of 0.01 and momentum of 0.9 were used to prevent the need for a validation set given limit images.

To perform face-verification, we adopted a threshold-free similarity-based approach which may be applied to any layer of the network, including the input images. First, given a set of feature responses $[\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ over images, the cosine distances between all test-set images were computed as $D_{i,j} = \cos(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and then normalized to a range of 0-1. A range of thresholds $\theta_k \in [0, 1]$ was then used to compute a matrix of same/different judgments $\boldsymbol{Y_k} = \boldsymbol{D} > \theta_k$. The $\boldsymbol{Y}_k$ matrices are then compared to the true same/different matrix to compute true positive and false positive rates t_k and f_k. The vectors $\boldsymbol{t}$ and $\boldsymbol{f}$ then constitute a Receiver-Operating-Characteristic (ROC) curve, and the area under the curve (AUC) was computed with numerical integration. Finally, d' was computed as $d' = \sqrt{2} \cdot \text{invnorm}(\text{AUC})$, where invnorm$(x)$ returns the value where the standard normal CDF equals x. This approach was applied to image pixels as well as the output of each block of VGG16.

Our verification experiments utilized 3 datasets: the deep-funneled images(Huang et al., 2012) of the Labeled Faces in the Wild (LFW) database (Huang et al., 2007), the Caltech-UCSD birds database, and the Stanford Cars database. For LFW, the smallest dataset, identities with at least 18 images were selected and 10 images were held out for the test set. For the other databases, 19 images were held out for testing.

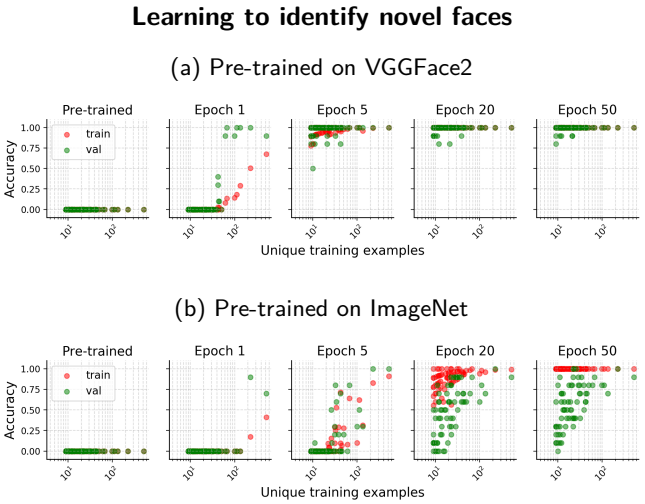


Figure 1: Output accuracy of the deep neural network for training and test images, before and throughout fine-tuning on a new set of faces in Labeled Faces in the Wild (LFW), per identity as a function of the number of training images for that identity.

Results

In Figure 1, the training and testing accuracy are presented at various stages of learning for a network trained on faces in VGGFace2 and one trained on objects and animals in ImageNet. While the face-trained network quickly and robustly learned to categorize both training and testing images for the new identities, the object-trained network learned much more slowly and struggled to generalize its learned knowledge to held-out test images.

Since identification accuracy does not allow us to compare unfamiliar and familiar face recognition, the main approach we discuss next is based on face verification, as described in Methods. The distance-based verification metric is applied to image pixels and each layer in the network, allowing for the possibility that an earlier layer will yield superior performance, and providing a measure of the extent to which performance is based on image- or low-level statistics. Verification ROC curves are shown for face and object trained networks in Figure 2 (b) and (c), reinforcing the benefit of experience with faces in verifying unfamiliar faces. As described in Methods, we used these ROC curves to compute area-under-the-curve which was then converted to d'. In Figure 2d, face verification d' is plotted for each form of pre-training, before and after familiarization fine-tuning on the test set of VGGFace2.

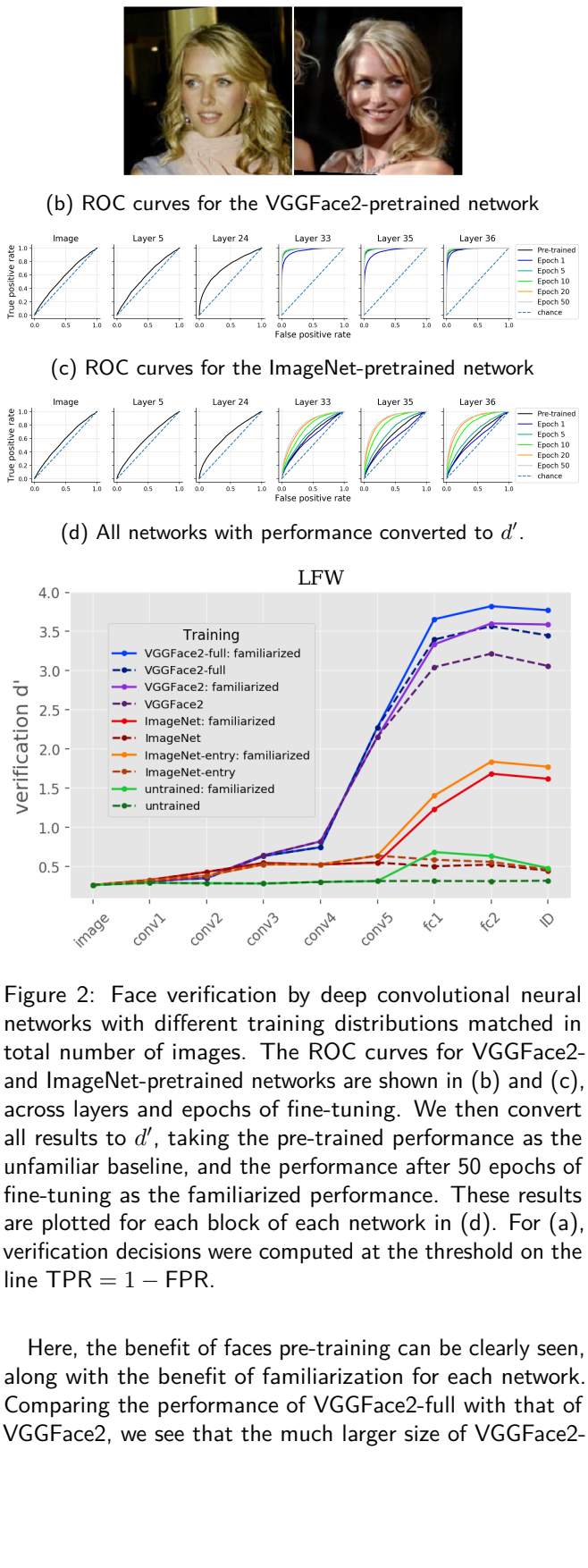


Figure 2: Face verification by deep convolutional neural networks with different training distributions matched in total number of images. The ROC curves for VGGFace2- and ImageNet-pretrained networks are shown in (b) and (c), across layers and epochs of fine-tuning. We then convert all results to d', taking the pre-trained performance as the unfamiliar baseline, and the performance after 50 epochs of fine-tuning as the familiarized performance. These results are plotted for each block of each network in (d). For (a), verification decisions were computed at the threshold on the line $\text{TPR} = 1 - \text{FPR}$.

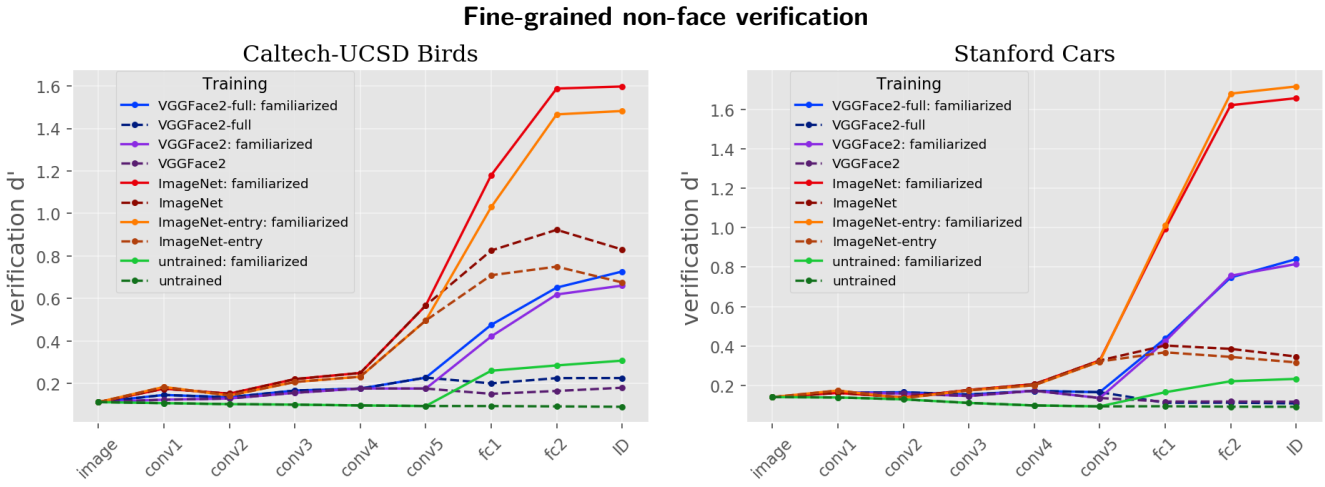
Here, the benefit of faces pre-training can be clearly seen, along with the benefit of familiarization for each network. Comparing the performance of VGGFace2-full with that of VGGFace2, we see that the much larger size of VGGFace2-full lead to moderate gains in verification performance, gains which are slightly larger for unfamiliar than familiar faces. However, these gains were much smaller than the gain from training on faces vs. objects in ImageNet, seen by comparing the results for VGGFace2 with those for ImageNet or ImageNet-entry.

Finally, we assessed verification of two non-face stimulus sets: bird species in the Caltech-UCSD database, and car models in the Stanford Cars database, shown in Figure 3. Importantly, we removed all birds and cars from ImageNet in the selection of our training set, so as to ensure that performance was not based on any domain expertise. Despite this, the ImageNet-trained models still performed better on verification of these fine-grained nonface categories. This result suggests that the image statistics learned for the development of face expertise are specific to faces and generalize only weakly to other categories. Further, the size of the face training distribution had a minimal effect on performance on these non-face categories.

To assess the influence of categorization level, which some research has shown to be important for generalization to new fine-grained recognition tasks (Tong et al., 2008), we also trained a network to recognize the ImageNet categories at a coarser entry level. In contrast with the idea that fine-grained performance for novel stimulus domains is tightly linked to the categorization level of previously learned stimuli, we found a weak and inconsistent effect of trained categorization level on novel fine-grained verification. For birds, a finer grained trained categorization level improved performance moderately. However, for cars, the level of categorization had a very small effect that reversed sign before and after familiarization. Notably, face recognition is finer grained than either of the ImageNet categorization tasks, and yet produced substantially worse performance than either of the ImageNet-trained networks on these non-face verification tasks.

Discussion

The idea that humans are poor at unfamiliar face recognition has gained considerable attention recently, in part due to a host of research demonstrating deficits on challenging face verification tasks for unfamiliar vs. familiar faces. We argue that this deficit is a natural consequence of the extreme difficulty of unfamiliar face verification. In this paper, we sought to elucidate this difficulty, placing human performance in context. Our simulations revealed that a network trained for face recognition was impaired on recognition of unfamiliar faces compared to its performance following familiarization. However, performance on unfamiliar faces in deep layers of the network was very good compared to earlier layers, suggesting that generic face variability learned in pre-training aided performance on unfamiliar faces. We found that increasing the extent of experience with familiar face exemplars modestly improved generalization to new unfamiliar identities, suggesting that continued experience



(a) Results for the Caltech-UCSD birds database

(b) Results for the Stanford Cars database

Figure 3: Verification of birds (a), and cars (b) by deep convolutional neural networks with different training distributions, using the same approach as described in Figure 2.

may allow for further untangling of generic face image variability, which likewise improves recognition even after familiarization. Crucially, by testing networks pre-trained on objects rather than faces, we demonstrated that modest performance on unfamiliar face recognition depends on learned domain knowledge of generic face variability through pre-training on face images.

Further, we assessed generalization of the same DCNN models used in the face verification experiments to two non-face domains of fine-grained visual recognition – bird species and car models. We found here that face pre-training produced substantially worse performance on the novel fine-grained recognition tasks, as compared to training on a subset of ImageNet in which bird and vehicle categories were removed. Further, we found that the level of trained categorization level had a small and inconsistent effect on fine-grained recognition, in contrast with earlier simulations that highlighted an influence of trained categorization level on generalization to new fine-grained tasks (Tong et al., 2008). We believe that these results can be reconciled by recognizing that the networks here were trained on many more stimuli in a deep convolutional architecture capable of extracting deeper semantically-relevant image statistics in the learned domains. As a result, image domain statistics dominated the effect of task, which might have been more salient in a shallower architecture.

The claim that identity-specific invariance lies at the heart of difficulty in unfamiliar face recognition has been explored in computational simulations by Kramer et al. (2018). As in our study, they found improved performance with familiarization of a set of tested identities. However, their simulations yielded very poor performance on unfamiliar faces. In contrast, here, we show that impressive but sub-optimal face verification performance is possible at the outset given prior experience with faces, with the best performance in the penultimate layer of the deep convolutional neural network (DCNN) Our results thus add an important piece of contextual information about the conditions of successful unfamiliar face recognition. In future work, we plan to collect comparable human data to more directly compare the effects of visual familiarity in face recognition by humans and machine learning systems.

References

Huang, G. B., Mattar, M. A., Lee, H., and Learned-Miller, E. (2012). Learning to Align from Scratch. In *NIPS*.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical report, University of Massachusetts, Amherst.

Kramer, R. S., Young, A. W., and Burton, A. M. (2018). Understanding face familiarity. *Cognition*, 172:46–58.

Megreya, A. M. and Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, 34(4):865–876.

Ordonez, V., Deng, J., Choi, Y., Berg, A. C., and Berg, T. L. (2013). From Large Scale Image Categorization to Entry-Level Categories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2768–2775.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Facebook, Z. D., Research, A. I., Lin, Z., Desmaison, A., Antiga, L., Srl, O., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS*, pages 1–4.

Rossion, B. (2018). Humans Are Visual Experts at Unfamiliar Face Recognition. *Trends in Cognitive Sciences*, 22(6):471–472.

Sunday, M. A. and Gauthier, I. (2018). Face Expertise for Unfamiliar Faces : A Commentary on Young and Burton ' s " Are We Face Experts ?". *Journal of Expertise*, x(x):1–7.

Tong, M. H., Joyce, C. A., and Cottrell, G. W. (2008). Why is the fusiform face area recruited for novel categories of expertise ? A neurocomputational investigation. *Brain Research*, (1202):14–24.

Young, A. W. and Burton, A. M. (2018). Are We Face Experts? *Trends in Cognitive Sciences*, 22(2):100–110.