Using Psychophysical Methods to Understand Mechanisms of Face Identification in a Deep Neural Network

Tian Xu
University of Glasgow
Glasgow, United Kingdom
Tian.Xu@glasgow.ac.uk

Oliver Garrod
University of Glasgow
Glasgow, United Kingdom
Oliver.Garrod@glasgow.ac.uk

Steven H Scholte
University of Amsterdam
Amsterdam, Netherlands
H.S.Scholte@uva.nl

Robin Ince
University of Glasgow
Glasgow, United Kingdom
Robin.Ince@glasgow.ac.uk

Philippe G Schyns
University of Glasgow
Glasgow, United Kingdom
Philippe.Schyns@glasgow.ac.uk

Abstract

Deep Convolutional Neural Networks (CNNs) have been one of the most influential recent developments in computer vision, particularly for categorization [20]. The promise of CNNs is at least two-fold. First, they represent the best engineering solution to successfully tackle the foundational task of visual categorization with a performance level that even exceeds that of humans [19, 27]. Second, for computational neuroscience, CNNs provide a testable modelling platform for visual categorizations inspired by the multi-layered organization of visual cortex [7]. Here, we used a 3D generative model to control the variance of information learned to identify 2,000 face identities in one CNN architecture (10-layer ResNet [9]). We generated 25M face images to train the network by randomly sampling intrinsic (i.e. face morphology, gender, age, expression and ethnicity) and extrinsic factors of face variance (i.e. 3D pose, illumination, scale and 2D translation). At testing, the network performed with 99% generalization accuracy for face identity across variations of intrinsic and extrinsic factors. State-of-the-art information mapping techniques from psychophysics (i.e. Representational Similarity Analysis [18] and Bubbles [8]) revealed respectively the network layer at which factors of variance are resolved and the face features that are used for identity. By explicitly controlling the generative factors of face information, we provide an alternative framework based on human psychophysics to understand information processing in CNNs.

1. Introduction

CNNs offer examples of complex, nonlinear projections of high-dimensional input images that can potentially serve as intuition pumps for developing and testing complex models of human visual categorization, using behavioral and brain measures ([3, 4, 17, 36]). However, though CNNs are unquestionably powerful enough to develop embryonic artificial intelligence, their role as intuition pumps for understanding information processing in the brain first requires understanding how they do what they do, so CNNs can serve as actual models. Otherwise, all that is offered is a silicon black box to understand the inner workings of a wet one [16].

A good starting point to such understanding is to uncover the visual information that CNNs process across layers. Following nonlinear learning, the lower convolution layers to the mid and higher-level layers represent features of increasing complexity and receptive field size. Multi-layered deconvolution (deconvnet) [40] can be used to identify these features. Thus, suitably constrained CNNs (by architecture, sub-functions, time and so forth) could in principle learn the mid-to-high-level features that flexibly analyze visual categories in a task-dependent manner, as humans do.

As with the brain, a better understanding of the information processed in CNNs is key to further the understanding of the mechanisms of that processing. Our main contribution is to tackle the challenge of understanding visual information processing in CNNs by adopting a psychophysical approach that emphasizes a better control of stimulus information. Specifically, we used face identification, a circumscribed but important categorization task that has been extensively studied in humans. Unique to our approach, we generated 25M 2D images of faces using an in-house 3D

Figure 1. Extrinsic (a and b) and Intrinsic (c) Factors of Variance of the Generative Model of 3D Faces

generative model of face identity, where we explicitly controlled the high variance of the resulting 2D images along extrinsic and intrinsic factors (see Figure 1). This psychophysical approach provides a unique platform to rigorously test the behavior of CNNs (here, ResNet [9]) in response to changes in each factor of image variance. Then, we used state-of-the-art information mapping techniques from psychophysics to understand information processing within ResNet. Our approach, when combined with psychophysical methods, develops a deeper understanding of the information processing mechanisms that produce similarities in the behavioral performances between deep and human brain networks. These comparisons can depart from analyses of superficial similarities of performance to reveal the deeper similarities of the information processing mechanisms that produce these similar performances.

2. Related Work

Face categorization has been a longstanding topic of human and machine vision research. In human vision, the challenge is to understand where, when and how brain mechanisms realize face identification under the main conditions of variance represented in Figure 1 (plus translation and scaling). The Bubbles technique [8] is a psychophysical technique that can reveal the use of information in recognition tasks from behavioral and brain measures [32, 15]. Specifically, it is possible to represent which brain region(s) represent which specific feature(s) (e.g. of a face) within a given time window post stimulus. Representational Similarity Analysis (RSA, [18]) is another method that proceeds by comparing different sources of responses (e.g. human behavior, computational models and brain activity) to the same stimulus categories.

In computer vision, the focus has recently been on deep learning to increase categorization performance. With a large volume of data, deep learning methods (e.g. DeepFace [34], FaceNet [29], face++ [43]) can perform above human levels. However, as with the human brain, the challenge remains to understand the information processing mechanisms underlying high performance levels.

Much research is focused on understanding the information processing mechanisms of CNNs. Zeiler and Fergus [40] used deconvolutional networks to backtrack the network computations and identify the image patches responsible for patterns of activation. Simonyan et al. [31] presented a visualization technique based on gradient ascent which generates a synthetic image that can maximally activate a unit in a deep network. Zhou et al. [42] proposed Class Activation Maps (CAM) that can highlight the image regions that the network uses for discrimination. As CNNs are inspired by human vision, to compare CNNs to the human brain we can use psychophysical methods that have already been successfully applied to understand the relationship between variations of stimulus dimensions and variations of brain responses. The approach is different to that typical of CNNs because researchers try to isolate the main factors of stimulus variance and to precisely measure responses to these variations.This is difficult to achieve with large datasets of unconstrained 2D images.

3. Dataset: Generative Model of 3D Faces

Several datasets (e.g. Labeled Faces in the Wild database (LFW) [12], Youtube Faces DB [37]) have been used to benchmark face recognition performance in CNNs. However, these databases cannot be used to analyze the effect of each relevant factor of face variance on the performance of the network. This arises because face images vary in the wild according to intrinsic and extrinsic factors as discussed

earlier that unconstrained 2D pictures do not control. Here, we used a Generative Model of 3D Faces (GMF) [39, 41] to generate realistic and well controlled but variable face images to understand how CNNs perceive them.

Specifically, we randomly generated 2000 identities using intrinsic factors of 500 random face morphologies $\times$ 2 genders $\times$ 2 ethnicities. We then varied each identity using additional factors of age (25, 45 and 65 years) and emotions (i.e. "happy", "surprise", "fear", "disgust", "anger", "sad" and "neutral"). For each identity, we also varied extrinsic rendering factors of rotation and illumination (both range from -30° to +30° by increments of 15°) along the X and Y axes, producing a total of 25M 256 $\times$ 256 pixels RGB images. Note that the network architecture learned images with the added extrinsic factors of random face scaling and 2D translation in its data augmentation mode (see Figure 1 and Training Regimes below). Hence the variance of input data was high, but controlled for the main dimensions of face variance.

4. CNN: 10-layer ResNet

A 10-layer ResNet architecture was trained to learn the face identities. We chose ResNet (Residual Network [9]) because it is a state-of-the-art architecture that has achieved highest classification performance on various datasets (e.g. ImageNet [19], COCO [21]). The architecture is composed of several similar residual blocks (i.e. two layers with a shortcut connection, see Figure 2). This facilitates optimization in comparison to direct optimization of a network without shortcut connections. Furthermore, direct connections within ResNet do not add extra parameters nor computational complexity, which keeps the complexity of the network low. Note that we use the simple 10-layer ResNet (ResNet-10) which keeps the network complexity relatively low, simplifying our detailed layer by layer analysis of information processing. ResNet also provides facilities to easily stack more layers (e.g. ResNet-50) for tasks of higher complexity, however, since performance on our task here was already saturated with 10 layers, this was not necessary.

A general training regime was used as benchmark of the overall face identification performance. For training the model, we randomly selected 60% of the generated face images, for a total of 15,750,000 images. The remaining 40% were used as testing images, for a total of 10,500,000 images. At training, we applied data augmentation to increase the data complexity and alleviate the problem of model overfitting. The trained images could be randomly scaled (between 1$\times$ and 2$\times$), translated in the 2D plane (between 0 and 0.3 of the total image width and height).

Across the 10,500,000 testing images, the network correctly generalized the face identity represented in the images with a remarkable accuracy of 99% and performance

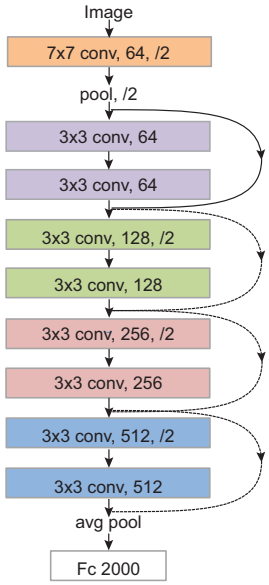

Figure 2. 10-layer ResNet Architecture

saturated across the variations of all intrinsic and extrinsic factors. Now the question is how to interpret the underlying process of the deep neural network.

5. Psychophysical Testing of CNNs

To understand the mechanism of the remarkable performance of this network, we applied several standard psychophysical methods which have been proved to be effective in understanding human visual system.

5.1. Generalization Test

In this section, we used another two training regimes to better understand the generalization performance of ResNet. These two training regimes varied image similarity between the training and the test sets by using similar vs. dissimilar parameters of the Generative Model of 3D Faces.

5.1.1 Training Regimes and Testing

For the first training regime, the 2,000 face identities are split into two disjoint groups, and different parameters were used for each group as shown in Table 1. These parameter values were chosen to minimize the similarity between the image sets. At testing, we swapped specific parameter values across identity groups to understand whether ResNet could generalize across these differences. One testing example (i.e. swapping X Axis Rotation) is highlighted in Table 1. We tested whether a given identity from group 1 trained on X-axis rotations 15° and 30° could generalize to rotations -30°, -15° and 0°, trained on group 2 of identities. Similarly, we experimented with generalization of identities across changes of each of the listed factors of variance

Minimize Similarity		Identity	Emotion	X Axis Rotation Angle	Y Axis Rotation Angle	X Axis Illumination Angle	Y Axis Illumination Angle
Training Data	Group 1	1-1000	Happy Surprise Fear	15 30	-30 -15 0	-30 -15 0	15 30
	Group 2	1001-2000	Disgust Anger Sad	-30 -15 0	15 30	15 30	-30 -15 0
Testing Example: Swapping X Rotation	Group 1	1-1000	Happy Surprise Fear	-30 -15 0	-30 -15 0	-30 -15 0	15 30
	Group 2	1001-2000	Disgust Anger Sad	15 30	15 30	15 30	-30 -15 0

Table 1. Illustration of Training and Testing in the Regime that Minimized Similarity.

Maximize Similarity		Identity	Emotion	X Axis Rotation Angle	Y Axis Rotation Angle	X Axis Illumination Angle	Y Axis Illumination Angle
Training Data	Group 1	1-1000	Happy Surprise Fear	-15 15	-30 0 30	-30 0 30	-15 15
	Group 2	1001-2000	Disgust Anger Sad	-30 0 30	-15 15	-15 15	-30 0 30
Testing Example: Swapping X Rotation	Group 1	1-1000	Happy Surprise Fear	-30 0 30	-30 0 30	-30 0 30	-15 15
	Group 2	1001-2000	Disgust Anger Sad	-15 15	-15 15	-15 15	-30 0 30

Table 2. Illustration of Training and Testing in the Regime that Maximized Similarity

(i.e. Emotion, X Axis Rotation, Y Axis Rotation, X Axis Illumination, Y Axis Illumination).

For the second training regime, we again split the 2,000 face identities into two groups. However, here we maximized similarities between their parameters of both groups, as demonstrated in Table 2 and then trained the network with these parameters. Similar to the first regime, at testing stage, we tested generalization of identities across each of the listed factors of variance (i.e. Emotion, X Axis Rotation, Y Axis Rotation, X Axis Illumination, Y Axis Illumination), and one example (swapping X Axis Rotation) is shown in Table 2.

Furthermore, for both the dissimilar and similar training regimes, at testing we also swapped combinations of parameters (i.e. Rotation on both X and Y axis, Illumination on both X and Y axis, both Rotation and Illumination on both X and Y axis), to further understand the the generalization performance when more drastic changes have been applied. Training and testing face image examples in Tables 1 and 2 are provided in the supplemental material[1].

[1]This is available at https://sites.google.com/site/skytianxu/SuppMatCVPRW.pdf

5.1.2 Results

The overall testing performance is decreased compared to the benchmark (Section 4) for both training regimes. To understand this degradation, we turn to the comparison of ResNet trained on dissimilar vs. similar images with the swapping of parameters as described above and in Tables 1 and 2. As can be seen in Figure 3(a), following dissimilar training, generalization of face identity remains high under changes of emotion, but deteriorates as expected from human performance [10, 35, 25, 1] when generalizing across new rotations and illuminations (i.e. not explicitly learned for this identity), particularly so when these are combined. As such generalization requires drastic extrapolations in the networks modelled face space, the performance remains remarkable.

As expected, following similar training, generalization of face identity remains high across all factors except for

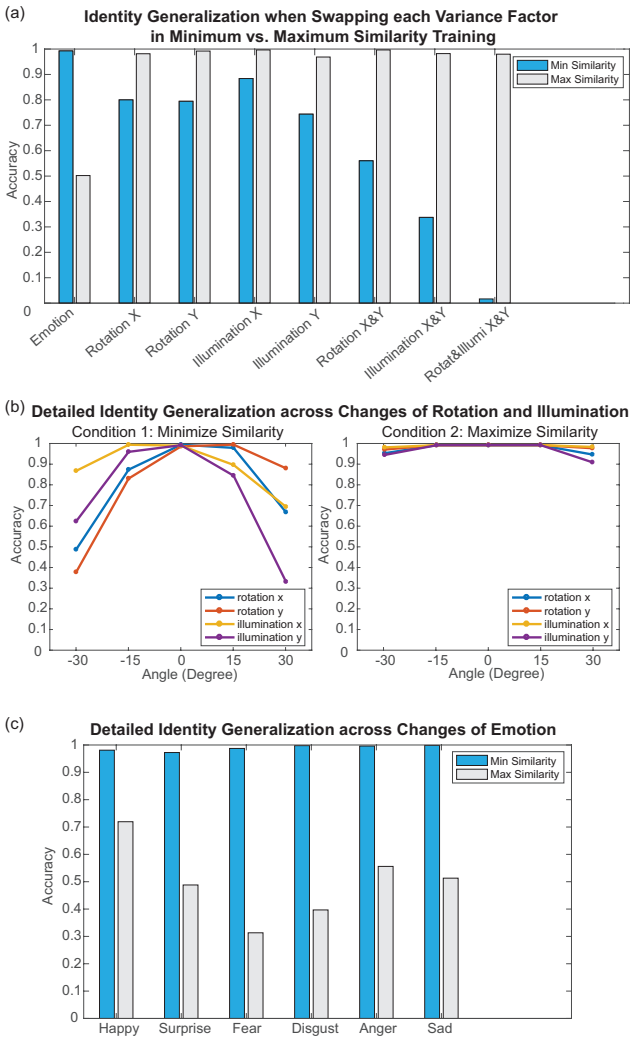

Figure 3. Identity Generalization Performance

emotion (see Figure 3(a)). Figure 3(b) details the identity generalization gradients across changes of rotation and illumination and Figure 3(c) provides the expanded data for generalization of identity under changes of emotion.

In sum, Resnet-10 handled remarkably well generalization of face identity under conditions of high variance. Psychophysical testing of generalization under conditions of dissimilar and similar training revealed the expected drop-off in performance. When the network was trained on data which covered the range of the different test parameters, for example generalization by interpolation and extrapolation to -30°, 0° and +30° of rotation when the identity was trained on -15° and +15° (see Table 2), performance was degraded only slightly, primarily when extrapolating across emotions. When the network had to extrapolate recognition of identity into a more different set of parameters for example generalization to -30°, -15° and 0° of rotation (right side of face showing) when this identity was trained on 15° and 30° (left side of face showing, see table 1), performance was much more heavily degraded. This pattern of performance degradation is qualitatively similar to that observed in human subjects [10, 35, 1]. Such similarity of performance raises the question of whether the network learns representations that are similar to those of humans, or whether it resolves the task by different means. We now follow standard psychophysical methods to understand the use of stimulus information that can explain behavior.

5.2. Spatial Frequency Alterations

The human visual system is known to process stimuli by decomposing their contents locally and simultaneously across a number of different scales. In fact, such Gabor decomposition is an integral part of the most successful CNNs, including ResNet. Hence, it is legitimate to submit the network to a test of its generalization of face identification across a number of spatial frequencies, as is commonly done in human recognition studies (see [24, 33] for reviews).

To this end, we used a Laplacian Pyramid [2] to recursively decompose the frontal view (i.e. 0° of rotation) of each face identity into five Spatial Frequency (SF) bands of one octave each as shown in Figure 4. We produced test stimuli according to separate regimes: (1) Blurring, by peeling off one SF band, from the highest, (2) Deletion, by removing only one of five SF bands and (3) Single, by using only one SF band. We submitted ResNet trained according to general training, as explained above, to the unseen frontal view test images filtered as explained. Figure 4 presents the results.

The first row of Figure 4 shows that generalization accuracy with the full face is perfect (i.e. 100%), making this view an ideal test bed for further investigations of the SF information underlying this performance level. Performance

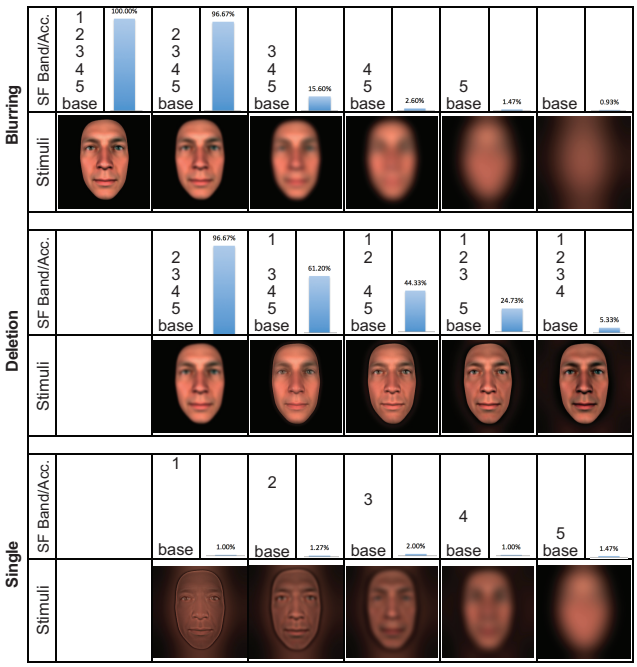

Figure 4. Spatial Frequency Tests (Blurring, Deletion and Single)

degraded nonlinearly in relation to increasing blurring, from 97% when the first SF band was removed, to 16% when the first two bands and more were removed. This is qualitatively similar to human performance, for whom face identity information is represented between 8 and 16 cycles per face width measured at eye level [24], where number of cycles is with reference to the Fast Fourier Transform of the face image. When deleting a single SF band (second row of Figure 4), the performance degraded linearly with deletion of lower SF bands. Finally, the third row demonstrates the importance of SF combinations because performance is dramatically reduced when only one SF band is present.

The interpretation of the ResNet SF generalization to SF alterations is more complicated than human generalization performance (e.g. row 2 and 3 are not observed in humans). ResNet uses information from all SF bands (whereas humans use a narrower mid SF band). Use of SF information progressively increases in ResNet from HSF to LSF (i.e. from band 1 to 5), a performance characteristic not observed in humans [24, 33].

5.3. Bubbles

To fine-tune the Identity SF information used in ResNet, we used Bubbles [8], a method applied successfully to model the stimulus information represented in human behavioral and brain data [14, 26, 15, 28] and also in CNNs [23]. Bubbles samples (with Gaussian apertures) the information represented in each SF band. The stimulus is a recomposed image, revealing only a subset of the information sampled on this trial. Here, we tested ResNet with a

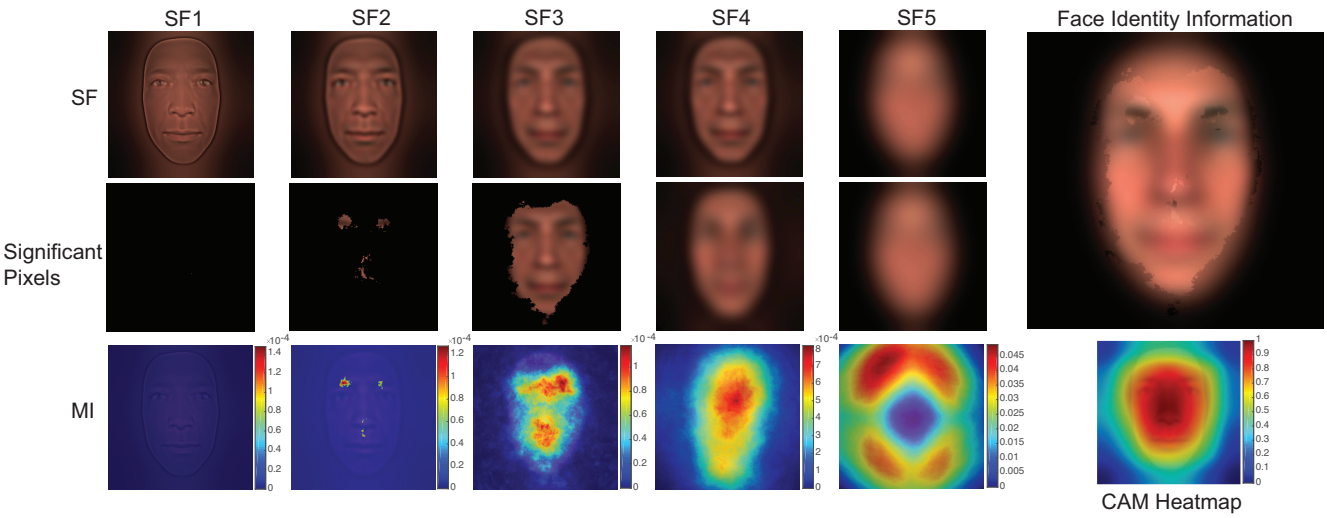

Figure 5. Significant Face Identify Information Across 5 Spatial Frequency Bands using the Bubble Technique

different number of bubbles (between 200 and 2,000, with increments of 200) to change the sampling density and performed 12,000 testing trials with each density applied to the full-face identity images of the testing set of the ResNet trained with the general regime above.

For each pixel at each SF band, we applied Mutual Information (MI) between pixel visibility (i.e. visible vs. not visible, as determined by Gaussian sampling) and network response (i.e. correct vs. incorrect identification of the face from the sampled image) [13]. A permutation test provided corrected statistical significance (FWER $p < 0.05$).

Figure 5 shows the SF decomposition of one image on the first row, the raw MI coefficients for each pixel of each SF band and the statistically significant pixels revealing the features present in each band. The rightmost face reveals the face information that the network uses to identify full faces. As expected from the spatial frequency alteration studies in Section 5.2, Bubbles reveals use of specific features in SF bands 2 to 5. In SF band 2, these features can be identified as representing details of the eyebrows and the nose. SF band 3 represents the centre of the face whereas SF bands 4 and 5 appear to represent coarse information about the full face. These results of Bubbles for face identity partially overlap to those reported in [30].

Here, we also compared the results of Bubbles with a related network visualization technique Class Activation Maps (CAM) [42]. The idea is to extract the output of last convolutional layer of ResNet in response to a face identity and compute the weighted sum of the feature maps resulting in a mapping resolution of 7×7. Following upsampling and averaging across identities, we can compare the face region that are most important to face identity to our result. The CAM Heatmap in Figure 5 shows that the important regions are similar to those revealed by Bubbles. However, unlike Bubbles, this method is not designed to breakdown the features of interest by SFs (cf. the specific eyebrows and nose features revealed in SF band 2).

5.4. Representational Similarity Analysis

Complex multi-layer networks learn intermediate representations across the layers of their architecture. Here, we used Representational Similarity Analysis (RSA) [18] at each layer of the architecture to understand the information processing that is achieved across the layers considered. The core of RSA is simply the Representational Dissimilarity Matrix (RDM), which is well-known in standard pattern classification to underlie analyses such as Multi-Dimensional Scaling [6].

Here, we applied RSA to the activation of each layer of the ResNet in response to 10,000 face images randomly selected from the testing dataset. It is important to emphasize that these images varied across all extrinsic and intrinsic factors of variance as explained earlier. For each layer, we therefore computed the 10,000 $\times$ 10,000 pairwise RDMs (i.e. $1-$correlations) across all pairs of the layer activations of the 10,000 face images. Note that here we use Pearson correlation to compute RDMs. Then we applied Pearson correlation again to compare the sorted RDM (according to the feature category) to the corresponding categorical model. The categorization model represents the ideal Pearson correlations that would result if the layers underneath the layer of interest directly represented the factor of variance considered.

Figure 6 shows the outcome of this analysis and two detailed examples. The identity (ID) is achieved in the 10th layer of ResNet, ethnicity peaks in 6th to 7th layer, and angle of rotation and illumination on the X axis (Anglex and Anglelx in Figure 6(a)) peaks from the 1st to the 4th layers,

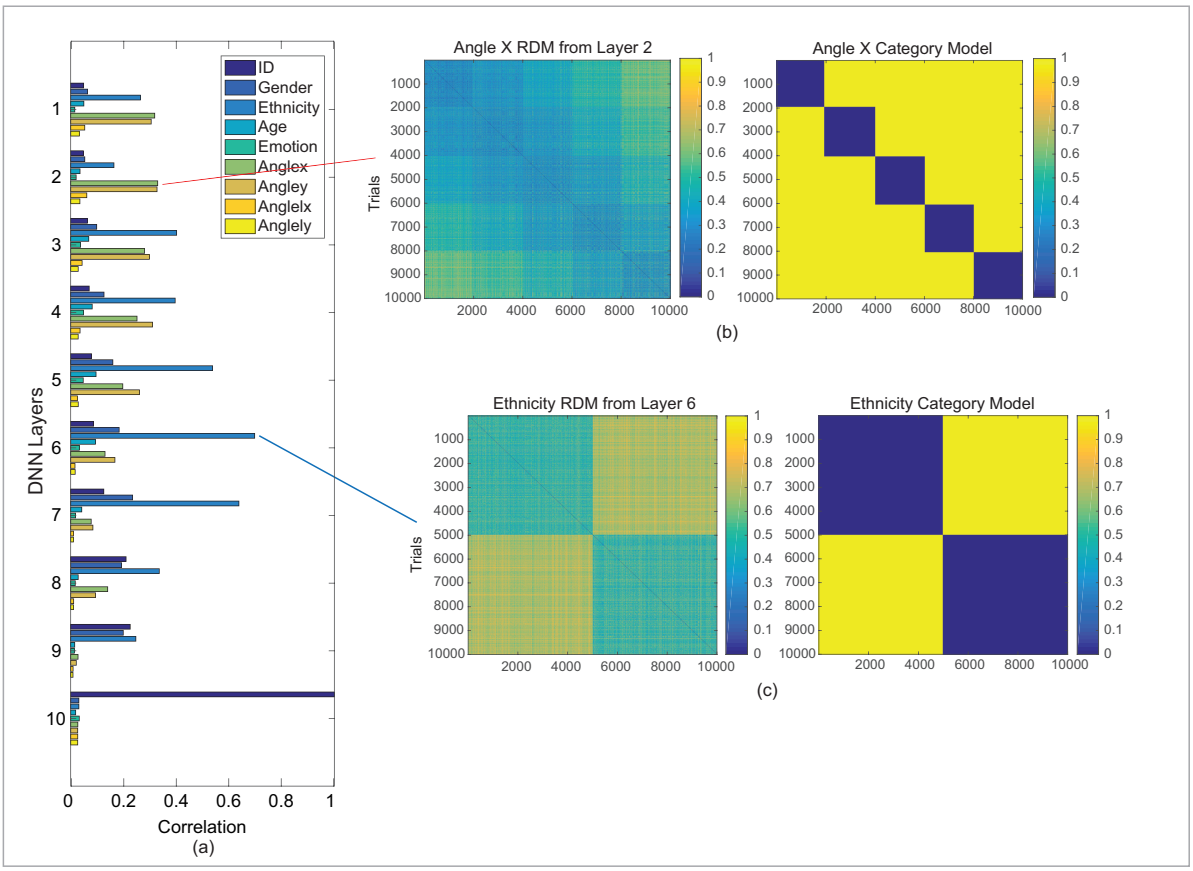

Figure 6. Correlation between RDMs and Category Models for each Factor of Variance across all 10 Layers of ResNet. (Note that Anglex and Angley refer to rotations in X and Y, respectively. AngleIx and AngleIy refer to changes in illumination with X and Y directions.)

etc. Even though the network was trained for identity rather than other specific features, it can still learn these parametric features implicitly. Generally, the network utilizes its first several layers to learn different implicit features of the face, and then develops more invariant representations that can reliably discriminate identity across the parametric variance dimensions in the deeper layers.

Thus, by tightly controlling the categorical sources of variance of an image set used to train a DNN, we derived an understanding of the implicit categorizations that the network could achieve at each layer.

6. Conclusion and Discussion

We used a 3D generative system of face identity with intrinsic (morphology, gender, emotion, age and emotions) and extrinsic (pose, illumination and data augmentation of scaling and translation) factors of variance to generate 25M 2D face images. We submitted a 10-layer ResNet architecture to three training regimes (general, minimizing similarities and maximizing similarities) on a 15,750,000 subset of images. Testing with the remaining subset, we found remarkable (i.e. 99%) generalization performances in psychophysical testing, with deterioration when the images were most dissimilar between training and testing. To start understanding the representations that the system uses to achieve such a high level of performance and compare it with humans, we used the psychophysical techniques of spatial frequency alterations, Bubbles and RSA. Spatial frequency alterations revealed that ResNet uses information across all SF bands, with an increasing reliance on LSF information. In contrast, human face identification relies primarily on mid-band SFs. We then applied Bubbles to understand which specific information ResNet uses in each SF band in full face pictures. We found usage of information as expected between SF bands 2 and 5, with specific details of the eyebrows and the nose in SF band 2. These Bubbles results with ResNet were related to those of Schyns et al. [30] in humans and broadly similar to Class Activation Maps. Finally, Representational Similarity Analysis revealed that whereas identity is resolved only at layer 10, ethnicity is resolved in layer 7 and pose and illumination on the X axis peak between the 1st and 4th layers.

In sum, our psychophysical testing revealed that the Deep Network learned to generalize face identity, consid-

ered to be the most difficult task with faces in a learning situation of high (but controlled) face variance. Performance only catastrophically broke down when inputs were single frequency bands. However, we do not know at this stage whether the network weights could be applied to quickly learn 2D faces from pictures taken in the wild and generalize to new exemplars. This will be the object of further studies though as Burton at al. showed human performance in this task is weak [38].

We now briefly consider the general question of whether our ResNet can be used as a model of human face identification. This is a broad question that we can address at two levels. First, we can evaluate whether ResNet provides a good functional model of human performance — i.e. a performance-to-performance mapping in the context of stimulus-response relationships. Though our data suggest interesting performance-to-performance mapping, we would really need to consider the vast literature on face identification and exhaustively test each known effect — e.g. the varied generalization patterns from single learned views of new identities [10], the poor generalization of the same identities across face pictures taken in the wild [38] and so forth. Such a tally would better characterize the performance palette and identify performance areas requiring improvement of the functional model. This leads the second evaluation of models, in terms of whether ResNet provides a good mechanistic model of human performance – a mechanism-to-mechanism mapping. Here, we are faced with thorny difficulties that deserve extensive discussions beyond the scope of this paper. CNNs are universal function approximators [5, 11], where each nonlinear unit of the network can serve as a building block for the approximation of a complex cognitive function (e.g. associating all images of birds with the category "bird"), which is realized across the units of several layers. In deep, multi-layered architectures, nonlinear building blocks are reused for different functions (e.g. associating all images of other categories with their corresponding labels). To use CNNs as intuition pumps, it is necessary to understand the functions they apply to visual information across nonlinear layers of their hierarchy, before testing the functions as models of information transformation in the brain. And we face a similar situation with the brain, so a mechanism-to-mechanism mapping is not currently achievable and even their use as intuition pumps would require more precise mathematical (or empirical) characterizations of their nonlinear projections. The approach used here to focus on few categories and control the stimulus information with a generative model (rather than use large databases of 2D stimuli) is akin to that used in neuroscience (and psychophysics) to understand the factors that modulate the activity of the brain (and behavior). We similarly used Bubbles to summarize the information processed in the system to achieve performance (cf. Figure 5). Our future work will focus on applying Bubbles and related techniques (e.g. [22]) to the layers and units of DeepNets to better understand how they process performance information and reduce information not directly useful for the task.

Acknowledgements

PGS is funded by the Wellcome Trust (107802/Z/15/Z) and the Multidisciplinary University Research Initiative (MURI) / Engineering and Physical Sciences Research Council (EP/N019261/1).

References

[1] H. H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, 89(1):60–64, 1992. 4, 5

[2] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision*, pages 671–679. Elsevier, 1987. 5

[3] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014. 1

[4] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016. 1

[5] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989. 8

[6] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification and scene analysis 2nd ed. *ed: Wiley Interscience*, 1995. 6

[7] K. Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988. 1

[8] F. Gosselin and P. G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17):2261–2271, 2001. 1, 2, 5

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 3

[10] H. Hill, P. G. Schyns, and S. Akamatsu. Information and viewpoint dependence in face recognition. *Cognition*, 62(2):201–222, 1997. 4, 5, 8

[11] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991. 8

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 2

[13] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human brain mapping*, 38(3):1541–1573, 2017. 6

[14] R. A. Ince, K. Jaworska, J. Gross, S. Panzeri, N. J. Van Rijsbergen, G. A. Rousselet, and P. G. Schyns. The deceptively simple n170 reflects network information processing mechanisms involving visual feature coding and transfer across hemispheres. *Cerebral Cortex*, 26(11):4123–4135, 2016. 5

[15] R. A. A. Ince, N. J. Van Rijsbergen, G. Thut, G. A. Rousselet, J. Gross, S. Panzeri, and P. G. Schyns. Tracing the flow of perceptual features in an algorithmic brain network. *Scientific reports*, 5:17681, 2015. 2, 5

[16] K. N. Kay. Principles for models of neural information processing. *NeuroImage*, 2017. 1

[17] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014. 1

[18] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008. 1, 2, 6

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012. 1, 3

[20] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015. 1

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[22] D. Linsley, S. Eberhardt, T. Sharma, P. Gupta, and T. Serre. What are the visual features underlying human versus machine vision? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2706–2714, 2017. 8

[23] K. Matzen and N. Snavely. Bubblenet: Foveated imaging for visual discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1931–1939, 2015. 5

[24] D. J. Morrison and P. G. Schyns. Usage of spatial scales for the categorization of faces, objects, and scenes. *Psychonomic Bulletin & Review*, 8(3):454–469, 2001. 5

[25] D. Perrett, P. Smith, D. Potter, A. Mistlin, A. Head, A. Milner, and M. Jeeves. Visual cells in the temporal cortex sensitive to face view and gaze direction. In *Proc. R. Soc. Lond. B*, volume 223, pages 293–317. The Royal Society, 1985. 4

[26] I. D. Popivanov, P. G. Schyns, and R. Vogels. Stimulus features coded by single neurons of a macaque body category selective patch. *Proceedings of the National Academy of Sciences*, 113(17):E2450–E2459, 2016. 5

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and Others. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[28] U. Rutishauser, O. Tudusciuc, S. Wang, A. N. Mamelak, I. B. Ross, and R. Adolphs. Single-neuron correlates of atypical face processing in autism. *Neuron*, 80(4):887–899, 2013. 5

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[30] P. G. Schyns, L. Bonnar, and F. Gosselin. Show me the features! Understanding recognition from the use of visual information. *Psychological science*, 13(5):402–409, 2002. 6, 7

[31] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[32] M. L. Smith, P. Fries, F. Gosselin, R. Goebel, and P. G. Schyns. Inverse mapping the neuronal substrates of face categorizations. *Cerebral Cortex*, 19(10):2428–2438, 2009. 2

[33] P. T. Sowden and P. G. Schyns. Channel surfing in the visual brain. *Trends in cognitive sciences*, 10(12):538–545, 2006. 5

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 2

[35] N. F. Troje and H. H. Bülthoff. Face recognition under varying poses: The role of texture and shape. *Vision research*, 36(12):1761–1771, 1996. 4, 5

[36] R. VanRullen. Perception Science in the Age of Deep Neural Networks. *Frontiers in Psychology*, 8, 2017. 1

[37] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011. 2

[38] A. W. Young and A. M. Burton. Are We Face Experts? *Trends in Cognitive Sciences*, 22(2):100–110, mar 2018. 8

[39] H. Yu, O. G. B. Garrod, and P. G. Schyns. Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3):152–162, 2012. 3

[40] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833. Springer International Publishing, Cham, 2014. 1, 2

[41] J. Zhan, O. B. Garrod, N. J. van Rijsbergen, and P. G. Schyns. Efficient information contents flow down from memory to predict the identity of faces. *BioRxiv*, page 125591, 2017. 3

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016. 2, 6

[43] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015. 2