Does CNN Explain Tuning Properties of Macaque Face-Processing System?

Rajani Raman (rajani.raman@atr.jp)

Haruo Hosoya (hosoya@atr.jp)
Cognitive Mechanisms Laboratories, ATR International
Kyoto, Japan 619-0288

Abstract

Recent computational studies have emphasized quantitative similarity between convolutional neural networks (CNNs) and the visual ventral stream up to the primate inferotemporal (IT) cortex. However, whether such similarity holds for the face-selective areas, a subsystem of IT, is not clear. To address this question, we extensively investigate CNNs in terms of known tuning properties of the face-processing network in macaque IT. Specifically, we first trained an AlexNet-type CNN model with natural face images. Then, we conducted simulation of four physiological experiments (Freiwald, Tsao, & Livingstone, 2009; Freiwald & Tsao, 2010; Ohayon, Freiwald, & Tsao, 2012; Chang & Tsao, 2017) to make a correspondence between the model layers and the macaque face patches. As a result, we found that higher model layers explained well properties of anterior patches, while no layer had properties close to middle patches; this observation was consistent across model variation. Our results indicate that, although the near-goal representation of face-classifying CNNs has some similarity with the primate face processing system, the intermediate computational process might be rather different, thus calling for a more comprehensive model for better understanding of this system.

Keywords: Deep learning; inferotemporal cortex.

Introduction

Goal-driven deep convolution neural networks (CNNs) have exhibited a remarkable similarity to ventral visual areas in terms of stimulus-response relationship despite that the network itself was not directly optimized to fit neural data (Yamins & DiCarlo, 2016). For example, CNNs optimized for image classification were highly predictive of neural responses not only in the inferotemporal cortex (IT) but also in the intermediate visual areas (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). A natural question arises here: if CNNs explain overall responses in IT, then do they also explain responses in a subsystem of IT?

Among various subsystems of IT, the most well-studied is the macaque face-processing system. This subsystem forms a network consisting of multiple face-selective patches with anatomically tight inter-connections (Moeller, Freiwald, & Tsao, 2008). The network putatively has a functional hierarchy from the middle to the anterior patches with a progressive increase of selectivity to facial identities and invariance in viewing angles (Freiwald & Tsao, 2010). For each patch, a number of tuning properties to specific facial features have been reported in a clear and detailed manner (Chang & Tsao, 2017; Freiwald et al., 2009; Freiwald & Tsao, 2010; Ohayon et al., 2012). Given these experimental facts, the macaque face processing system emerges as an ideal testbed to examine our question regarding the generality of CNNs as a model of higher visual processing.

Thus, in this study, we have thoroughly investigated whether CNNs explain previously reported tuning properties to facial features. For this, we incorporated four major physiological experiments that had been conducted on the middle lateral (ML), anterior lateral (AL), and anterior medial (AM) patches: (1) size invariance and view-identity tuning (Freiwald & Tsao, 2010), (2) shape-appearance tuning (Chang & Tsao, 2017), (3) facial geometry tuning (Freiwald et al., 2009), and (4) contrast polarity tuning (Ohayon et al., 2012). By conducting simulation of these experiments on CNNs with varied architectures and training conditions and by matching the results with the documented experimental data, we have attempted to make a correspondence between the model layers and the face-processing patches.

Results

We started with a representative CNN model, which had architecture similar to AlexNet but was trained on face images from VGG-Face dataset (with data augmentation for size variation) for classification. We refer to this network as 'AlexNet-Face'. We first identified a population of face-selective units in each ReLU layer; we call such unit simply 'unit' from here on. We then conducted each experiment on the face-selective population to see how each model layer corresponded to a face-processing patch. Here, we present the results from the ReLU1, ReLU3, ReLU5 (last convolutional layer), and ReLU7 (second fully-connected layer), since the remaining layers gave more or less interpolated results of the presented layers.

Size Invariance

To investigate the size invariance property, we presented a set of face and object images of various sizes to the model. For each layer, we calculated the average population responses, $\bar{R}_{face}$ and $\bar{R}_{object}$, for face and object images, respectively, for each image size. We quantified the degree of invariance by size-invariance (**SI**) index, which is defined as the minimal fraction of image sizes at which the average population response to faces is sufficiently larger than to objects ($\bar{R}_{face} > 1.4\bar{R}_{object}$). Thus, a lower SI-index indicates a stronger size invariance.

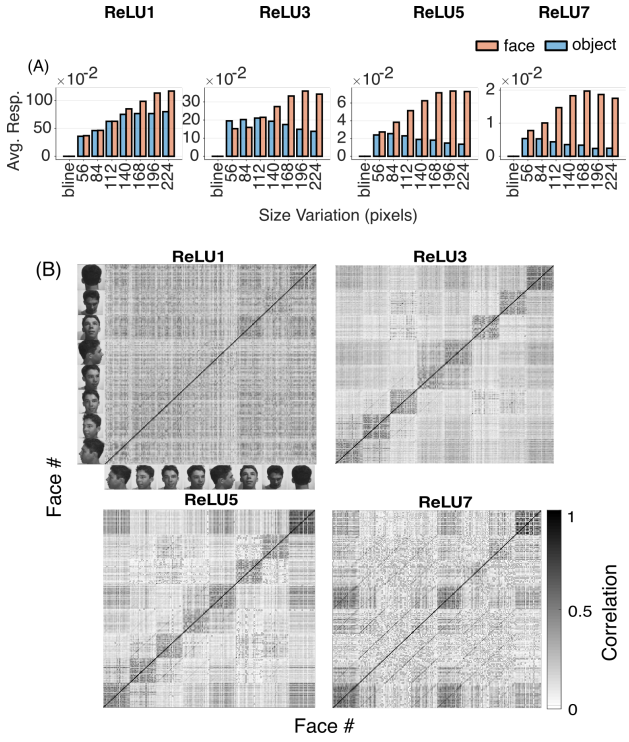

Figure 1: A) The average population response to face or object images of a varied size (x-axis). B) The population response similarity matrix. The face numbers are grouped according to the head orientation.

Not surprisingly, size invariance in the network increased along with its depth (Figure 1A). In particular, the top layer (ReLU7) had the strongest size invariance (SI = 1/4), where the average population response to faces was always larger than the one to objects for all sizes. In comparison to the macaque, the top layer quantitatively came closest, of all layers, to the face patches (ML/AL/AM), which all give SI $\approx 1/8$ (Freiwald & Tsao, 2010).

View-Identity Tuning

To probe the view-identity tuning in each layer, we used the same set of face images as in (Freiwald & Tsao, 2010), which consisted of 25 individuals and 8 head orientations. For each layer, we calculated the correlation between its population responses to each pair of face images. Then, we constructed a population response similarity matrix (**RSM**) of such correlations (Figure 1B), where the face numbers are grouped according to the head orientation (forming 25×25 sub-matrices).

The intermediate layers (ReLU3–5) had RSMs with darker blocks located along the main diagonal, indicating selectivity to a specific head orientation, similarly to ML (Freiwald & Tsao, 2010). The top layer (ReLU7) exhibited a mirror-symmetric pattern in head orientation, as well as para-diagonal lines indicating selectivity to facial identity with some degree of view invariance, which shows a combined property of AL and AM (Freiwald & Tsao, 2010). Thus, the intermediate-to-top layers gradually shifted from view-specific to (partially) view invariant and identity-selective, showing a reasonable similarity with the putative functional hierarchy in the macaque face patches (Freiwald & Tsao, 2010).

Shape-Appearance Tuning

To investigate coding of facial shapes and appearances in the model, we followed the experimental procedure in (Chang & Tsao, 2017). First of all, we constructed a 50-dimensional space of frontal face images, where the first 25 dimensions (shape dimensions) were the first 25 principal components (PC) of landmark coordinates annotated on pre-defined facial features, while the last 25 dimensions (appearance dimensions) were the first 25 PCs of the normalized face images that were obtained by morphing the original images to match the landmarks to their mean coordinates. Then, for the subsequent analyses, we randomly generated a set of face stimuli from the face space.

Shape-Appearance Preference (**SAP**). To examine whether and how much each unit preferred shape or appearance, we first measured the responses of each unit to those images and estimated a 50-dimensional vector of spike-triggered average (STA). Then, we computed the shape preference index, $(S-A)/(S+A)$, where S is the vector length of the 25 shape dimensions and A is the vector length of the 25 appearance dimensions of the STA. We found that the intermediate-to-higher layers (ReLU5–7) had most of the units with negative shape-preference indices (Figure 2A), indicating appearance preference similar to AM (Chang & Tsao, 2017). However, the lower layers (ReLU1–3) mixed both shape-preferring and appearance-preferring units, thus dissimilar to either AM or ML. In addition to these, we found that, throughout all layers, the units tended to have a ramp-like tuning along the STA and a flat tuning along axes orthogonal to the STA (data not shown), similarly to both ML and AM (Chang & Tsao, 2017).

Decoding Performances (**DP**). To reveal how much information each layer contained on the face space, we first decoded the 50-dimensional feature values from the population responses by linear regression. We then quantified the performance of the decoding by the average nearest-neighbor classification accuracy in the feature space for a chosen number of face classes (Chang & Tsao, 2017). Figures 2B shows the decoding accuracies (50-d) as a function of the number of face classes (2–40), along with the cases of decoding only the appearance or shape dimensions. The overall results for the top layer (ReLU7) were comparable to AM (Chang & Tsao, 2017). On the other hand, no layer showed lower accuracies than the top layer or the shape decoding more accurate than the appearance decoding, unlike ML (Chang & Tsao, 2017).

View Tolerance (**VT**). We constructed another face space for profile face images, similarly to frontal faces, and established a correspondence between the two face spaces via linear regression (Chang & Tsao, 2017). We then estimated the STA of each unit in the profile face space and calculated the

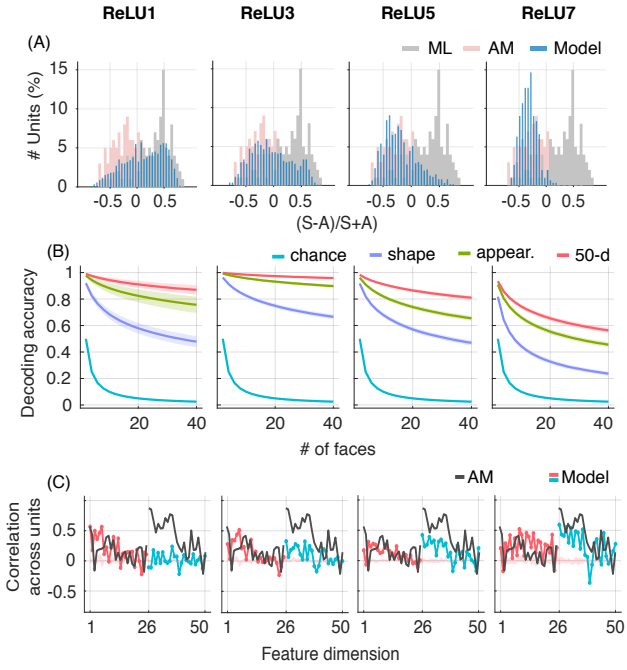

Figure 2: A) The distribution of shape-preference indices. B) Decoding accuracy as a function of the number of face classes. C) Dimension-wise correlations between the frontal and the profile STAs across units.

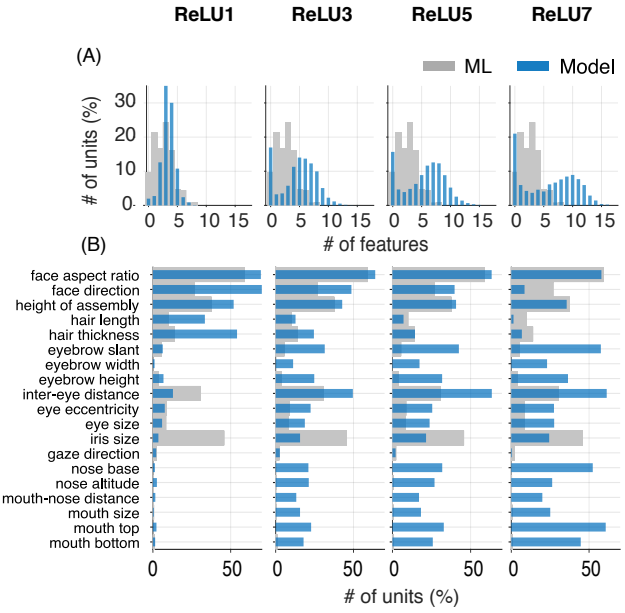

Figure 3: A) The distribution of the number of features that each unit is tuned to. B) The distribution of the number of units tuned to each feature.

correlation between the frontal and the profile STAs for each feature dimension across all units in each layer (Figure 2C). The correlations of the first half of the appearance dimensions increased in the intermediate-to-top layers, indicating a gradual progression of view tolerance along with its depth; in particular, the correlations in the top layer (ReLU7) came closest, of all layers, to AM (Chang & Tsao, 2017). Note that this result is compatible with Figure 1B.

Facial Geometry Tuning

We simulated the experiment on ML in (Freiwald et al., 2009) to investigate coding of cartoon face space. As in the experiment, we used cartoon face images parametrized by 19 different facial features, each ranging from -5 to $+5$ with ± 5 corresponding to the extreme features and 0 corresponding to the mean features. From the responses to randomly generated cartoon face images, we estimated a tuning curve of each unit for each feature, together with its statistical significance using the same criterion as in the experimental study.

As shown in Figure 3A, the distribution of tuned features per unit (**FPU**), i.e., how many features each unit is significantly tuned to, was reasonably similar to ML (~ 3 features on average) in ReLU1 (~ 4 features on average). Also, the distribution of tuned units per feature (**UPF**), i.e., how many units are tuned to each feature, was similar to ML in the lowest layer (Figure 3B); most of the units were tuned to geometrically larger features giving skewed shape to the distribution. However, both distributions gradually diverged from ML in higher layers, accommodating more units tuned to larger numbers of features and more remaining features reportedly absent in ML (Freiwald et al., 2009). In addition to these, we observed that, in all layers, the units had tuning curves mostly ramp-shaped with peaks and troughs at the extreme values (data not shown), similarly to ML (Freiwald et al., 2009).

Contrast Polarity Tuning

We also simulated the experiment on ML in (Ohayon et al., 2012) using mosaic-like cartoon face stimuli. Each stimulus consisted of 11 distinct parts, where each face part was assigned a unique intensity value varying from dark to light. For a set of randomly generated mosaic-like cartoon faces, we analyzed the responses of each unit to identify the preference in contrast polarity (**CP**) between each pair of parts (55 part-pairs in total). In all layers, most of the units had preferences for contrast polarities mainly related to the forehead (Figure 4), the largest geometrical area in the mosaic face. This result is inconsistent with ML (Ohayon et al., 2012), where most neurons were tuned to contrast polarities related to the eyes or the nose and the polarities were consistent across the neurons.

Consistency Across Model Variations

To further test whether the results shown so far carry over to other CNN models, we repeated all the experiments on several other networks: VGG-Face network (VGG-16 architecture trained on VGG-Face images), AlexNet (trained on ImageNet), Oxford102 network (AlexNet architecture trained on Oxford102 flower images), and four other CNNs with depth

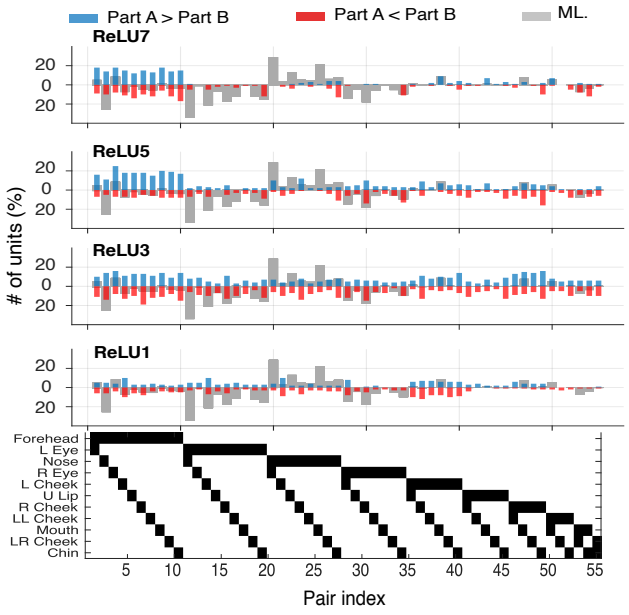

Figure 4: The distribution of contrast polarity preferences on each facial part-pair. Each plot shows positive polarities ($A > B$) in the upper half and negative polarities ($A < B$) in the lower half. The parts in each pair are indicated in the bottom matrix.

variations (trained on VGG-Face images). We found that the results were, indeed, largely consistent across all these networks (data not shown).

Conclusion

In this study, we investigated whether a CNN can be a model of the macaque face-processing network. We attempted to make a correspondence between the model layers and the face patch areas by comparing their key tuning properties. We found that higher model layers tended to explain reasonably well properties of the anterior patch (AM) (Table 1A), while none of the layers simultaneously captured those of the middle patch (ML) (Table 1B). Taken together, despite the prevailing view linking CNNs and IT, our results indicate that the intermediate computational process in the macaque face processing system might be rather different and therefore requires a more refined model to clarify the underlying computational principle. Feedback processing, which CNNs crucially lack, may be one direction to investigate (Hosoya & Hyvrinen, 2017).

Acknowledgments

This work was funded by NEDO (P15009), NICT (1940201), and Grants-in-aid (18H05021, 18K11517, 19H04999), Japan.

Table 1: Layer-wise correspondence with (A) AM or (B) ML. Check mark (✓) indicates consistency with experimental data on size-invariance (SI), population response similarity matrix (RSM) (Freiwald & Tsao, 2010); shape-appearance preference (SAP), decoding performance (DP), view tolerance (VT) (Chang & Tsao, 2017); features per unit (FPU), units per feature (UPF) (Freiwald et al., 2009); contrast polarity tuning (CP) (Ohayon et al., 2012).

(A) AM

	SI	RSM	SAP	DP	VT
ReLU7	✓	✓	✓	✓	✓
ReLU5	✓		✓		
ReLU3					
ReLU1					

(B) ML

	SI	RSM	SAP	DP	FPU	UPF	CP
ReLU7	✓						
ReLU5	✓	✓					
ReLU3		✓			✓	✓	
ReLU1					✓	✓	

References

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*, 1013–1028.e14.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, *330*, 845–851.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*, nn.2363.

Hosoya, H., & Hyvrinen, A. (2017). A mixture of sparse coding models explaining properties of face neurons related to holistic and parts-based processing. *PLOS Computational Biology*, *13*, e1005667.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*, e1003915.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, *320*(5881), 1355–1359.

Ohayon, S., Freiwald, W. A., & Tsao, D. Y. (2012). What Makes a Cell Face Selective? The Importance of Contrast. *Neuron*, *74*, 567–581.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624.