Relating deep neural network representations to EEG-fMRI spatiotemporal dynamics in a perceptual decision-making task

Tao Tu, Jonathan Koss, Paul Sajda
Columbia University
{tt2531,jk3953,psajda}@columbia.edu

Abstract

The hierarchical architecture of deep convolutional neural networks (CNN) resembles the multi-level processing stages of the human visual system during object recognition. Converging evidence suggests that this hierarchical organization is key to the CNN achieving human-level performance in object categorization [22]. In this paper, we leverage the hierarchical organization of the CNN to investigate the spatiotemporal dynamics of rapid visual processing in the human brain. Specifically we focus on perceptual decisions associated with different levels of visual ambiguity. Using simultaneous EEG-fMRI, we demonstrate the temporal and spatial hierarchical correspondences between the multi-stage processing in CNN and the activity observed in the EEG and fMRI. The hierarchical correspondence suggests a processing pathway during rapid visual decision-making that involves the interplay between sensory regions, the default mode network (DMN) and the frontal-parietal control network (FPCN).

1. Introduction

Understanding the neural correlates of rapid object recognition requires a comprehensive delineation of the neural cascades in time and in space. Specifically, a distributed brain network must temporally coordinate its activity during rapid decision-making. Perceptual decisions activate both sensory processing as well as high-level frontal control both of which can potentially interact with the motor system to generate behavior [9]. Under circumstances when there is inadequate sensory evidence in the stimulus, ambiguity arises in the decision process. The brain must potentially employ more complex processing that is not simply feed-forward, and instead utilize feedback pathways to integrate prior biases for choice [20]. In addition, processing related to directed attention is likely to switch the brain's processing between internal goals and external cues, and this can depend on the amount of sensory evidence received [15]. Thus there are several spatiotemporal processes occurring in a coordinated manner that lead to our ability to decide and act.

Simultaneous EEG and fMRI measurements make it possible to non-invasively observe spatiotemporal dynamics of the human brain while subjects make simple or complex decisions. EEG provides millisecond time-resolved measurements of the brain activity in response to external stimuli or change of brain states. Complementary to EEG, fMRI provides millimeter spatial-resolved measurements of hemodynamic activity across the whole brain. In this paper we leverage these multimodal neuroimaging measurements with recent advances in computational models for visual object recognition, the goal being to gain new insights into the spatiotemporal network-level processing underlying rapid perceptual decision-making.

Deep convolutional neural networks (CNN) have been the state-of-the-art for automated object recognition tasks for several years and are now able to achieve performance comparable to humans on such recognition tasks [14]. These models' structure contain a hierarchy of layers through which input images are fed to produce the resulting classification. Considering the similarities in structural organization and performance between these networks and the human brain, we hypothesize that comparisons between the layer representations of CNN and the spatiotemporal representations of the brain under the same task will shed light on the otherwise opaque workings of the human brain during rapid decision making. Specifically, we capitalize on a computational framework termed representational similarity analysis (RSA) which enables the comparison across measures of modalities by transforming the measurement of each different modality into a common similarity space that represents the activity pattern of the brain in response to the experimental stimuli [13]. Relating the layer activations of CNN to the spatiotemporal dynamics of the brain in response to the same set of experimental stimuli reveals a hierarchical correspondence between the CNN and the brain both in space and time. We find that this hierarchical correspondence further implicates a dynamical attention switch-

ing neural mechanism during decision ambiguity.

2. Related Work

Leveraging recent advances in computer vision, a number of studies have demonstrated the organizational similarity between the feature representation in human visual pathway and convolutional neural networks. Güçlü *et al*. [7, 8] built predictive models from the layer representations of CNN to predict the BOLD responses to natural image and movie stimuli. In line with the work by Eickenberg *et al*. [6], their findings showed that the hierarchical feature representation of the CNN is in congruence with the representation organization in both the dorsal and ventral visual pathways of the human brain. To enable the integration of information from multiple sources of brain measurements, several studies used RSA [13] to compare the representational similarity across modalities (EEG, MEG, fMRI, computational models and behavioral measurements). For example, Cichy *et al*. [3] showed the spatial and temporal hierarchical correspondence between the human brain and CNN in visual object categorization via a comparison across MEG, fMRI and CNN. Later, they investigated the representation in scene recognition between MEG and CNN, specifically relating the temporal dynamics in the MEG to multi-stage scene processing in the CNN [2]. Lastly, Kheradpisheh *et al*. [11] compared the representational similarity between human behavioral measures and different neural network models to study the viewpoint invariance in object recognition.

Most of these works have focused on establishing the hierarchical organization correspondence between the feature representations in CNN and in human visual pathways, however, none investigated the similarity between the CNN and human brain during a perceptual decision making where subjects were actively engaged in a decision process regarding the choice of a potentially ambiguous or noise image of an object category. Previous studies [19, 16, 20] have shown that the mapping from sensation to action involves a coordination of a cascade of neural events. In particular, the process taps into the attentional allocation and executive control functions of the brain and hence will recruit high-level brain regions in addition to sensory processing regions. Therefore, in this work, by comparing the simultaneously recorded EEG and fMRI with the CNN in representational space, we attempt to obtain temporally and spatially resolved brain dynamics during rapid object recognition in which the decision ambiguity was varied and dynamic switching between internal and external attention was potentially observable.

3. Methods

3.1. Stimuli and Experimental Paradigm

The stimulus image set consisted of a set of 30 face, 30 car, and 30 house images. The phase coherence of the images was degraded at a high coherence (50%) level and at a low coherence (35%) level using the weighted mean phase algorithm [5]. The phase coherence modulates the amount of sensory evidence in the stimuli and thus influences the decision ambiguity. Twenty-one subjects participated in the study. Subjects performed an event-related three-choice visual categorization task. On each trial, an image of a face, car, or house was presented for 100 ms. Subjects reported their choice of the image category by pressing one of the three buttons on an MR-compatible button response pad. Each subject participated in four runs of the categorization task. In each run, there were 180 trials (30 per condition; 6 conditions: face high, car high, house high, face low, car low, and house low). Therefore, simultaneous EEG and fMRI data from 720 trials (240 of each category and 360 of each coherence) were acquired for each subject during the entire experiment. We excluded data from three subjects in our analysis because of missing stimulus sequence information. More details in data recording and experiment design can be found in [16, 20]. Figure 1 illustrates the experimental design.

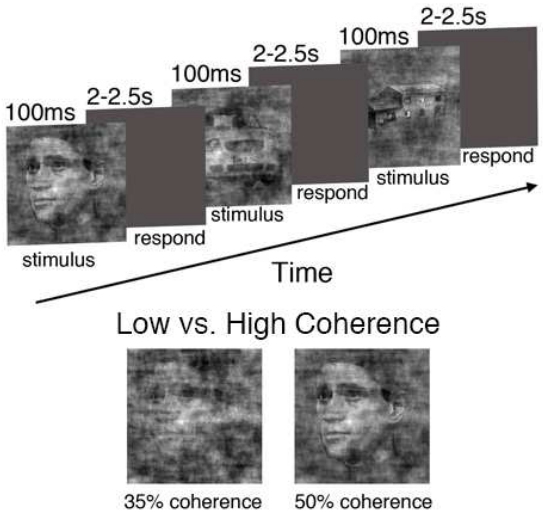

Figure 1. Illustration of the Face vs. Car vs. House visual categorization task. The phase coherence of the stimulus image modulates the decision ambiguity.

3.2. Training Image Set for CNN

The training image dataset consisted of 6,000 images of three object categories. Face images were acquired from the

LFW dataset [10]. House images were acquired from the MIT Places database [23]. Car images were acquired from the Stanford Cars dataset [12]. We selected 2,000 house images, 2,000 car images, and 1,000 face images from the datasets. We also generated 1,000 artificial face images from 100 of the natural face images via a 3D morphable model [1]. The artificial face images were designed to look similar to the face stimulus images presented to human subjects and thus including them in the training set improves the generality of CNN on the stimulus image set. We converted all images to grayscale and resized them to 224×224. Moreover, we augmented the training set by including the degraded images at 35% and 50% phase coherence levels for all categories. This yields a total of 18,000 images in the training set for CNN. The validation set containing 6,300 images and the test set containing 8,000 images were created in the same manner. Figure 2 shows some representative images used for CNN training.

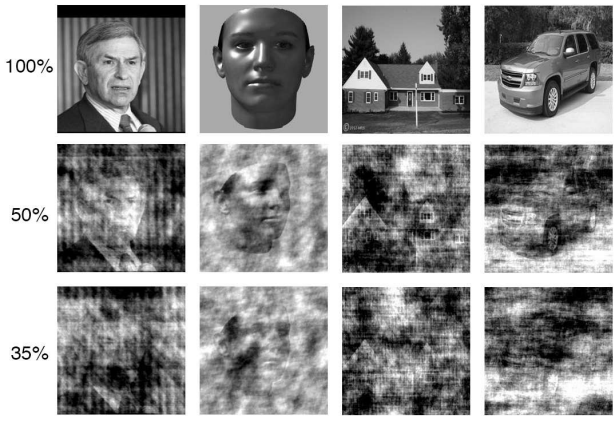

Figure 2. Representative face, car and house images in the training set. The artificial face images were designed to look similar to face images presented to human subjects and included in the training set to make CNN more generalizable to the stimulus image set.

3.3. CNN Architecture and Training

We used a deep convolutional neural network (CNN) architecture as described in [14] (VGG-16). In our design, the CNN consists of the convolutional base of VGG-16 and 2 fully connected (FC) layers as shown in Figure 3. We divided the convolutional portion of this network into five subgroups (layer 1-6). The first two groups consist of two convolutional layers followed by a max pooling layer and the following three groups consist of three convolutional layers followed by a max pooling layer. Each of these conv layers uses the rectified linear unit (ReLU) as its activation function. The first FC layer has 1024 hidden units and the ReLU activation function. Dropout was applied to this layer with the probability of retaining a hidden unit being $p = 0.5$. The final three-unit FC layer has a softmax activation function that outputs the class label for face, car and house. The training process consisted of two stages. In stage 1, we froze the weights in all convolutional layers and only trained the weights in two fully connected layers. We used RMSprop algorithm to minimize the cross-entropy objective function with a mini-batch size of 30, learning rate of 2×10^{-5}, and 10 training epochs. In stage 2, we fine-tuned the weights in all convolutional layers and the fully connected layers using the same optimization scheme but with a much smaller learning rate of 1×10^{-7} for only 2 training epochs to avoid overfitting.

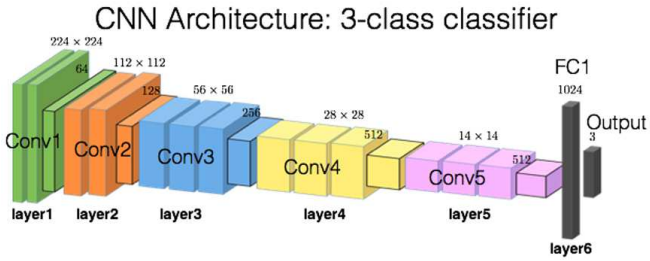

Figure 3. The architecture of the CNN consists of the convolutional base from a pre-trained VGG-16 and two fully connected layers. The last layer outputs a 3-class label.

3.4. Representational Similarity Analysis (RSA)

We used the RSA to characterize the relationship between the temporal (EEG) and spatial (fMRI) representations of the brain and the layer representations of CNN during the rapid decision-making process [4, 3]. Specifically, for each modality, we computed a representational similarity matrix (RSM) with a dimension of 180×180 corresponding to the number of stimulus images. Each entry in the RSM denotes the distance between two stimulus images. The more similar two images are, the less distance they are away from each other. In this study, we used Pearson correlation as a distance measure between each pair of the 180 images. Figure 4 shows the overview of the analysis methods.

EEG RSM. To obtain temporally resolved EEG RSMs, we computed the RSM at different time windows in a sliding window fashion spanning from 0 ms to 1000 ms post-stimulus onset. Each window has a 10 ms width and an overlap of 5 ms with the adjacent ones. For each time window, we first averaged the multi-channel EEG activity across all the time points in the window, yielding a 41×1 feature vector corresponding to each stimulus image. We then averaged the 41×1 feature vectors across 4 trials of the same stimulus as the final EEG feature used for the computation of the distance measure in the RSM. For each pair of images, we calculated the Pearson correlation between

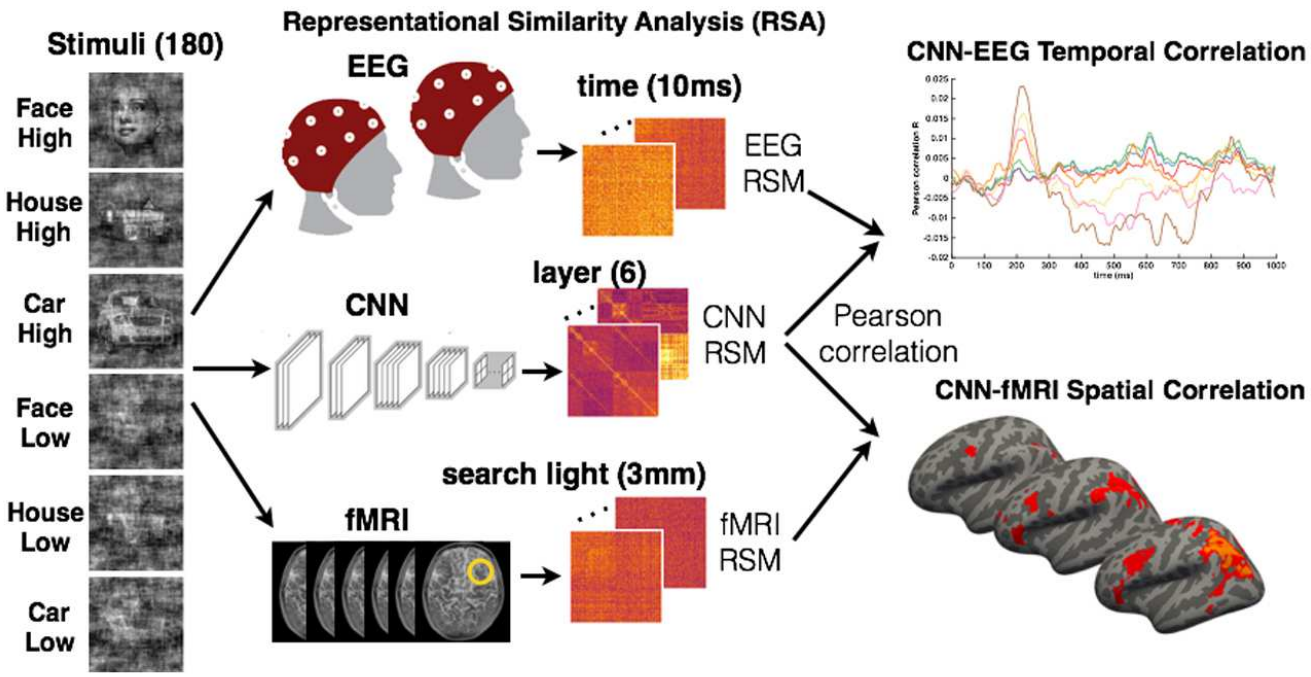

Figure 4. Illustration of RSA between CNN layer activations and EEG-fMRI recordings. Constructing RSMs from CNN, EEG and fMRI enables comparisons across layer, time and space.

them, yielding a 180×180 EEG RSM at each time window.

fMRI RSM. To obtain spatially resolved fMRI RSMs, we used a volume-based searchlight approach implemented in CoSMoMVPA toolbox [17]. In this approach, the whole brain volume is randomly divided into a number of spheres (searchlight). Each searchlight is a 3mm-radius sphere centering on a voxel. Each searchlight contains approximately 40 neighboring voxels, yielding a multidimensional voxel-pattern fMRI feature vector for each stimulus. Since BOLD activity is a lagged hemodynamic response following each stimulus onset, we first performed a temporal deconvolution [21] on the BOLD data to generate a trial-level voxel response for each stimulus, from which we can then construct voxel feature vector for each stimulus. Similar to the computation of EEG RSM, for each subject, we averaged the trial-level voxel responses for each stimulus and computed the fMRI RSM at each searchlight cluster, yielding a spatially localized RSM across the whole brain.

CNN RSM. To obtain the layer-wise RSM, we extracted the filter activation vector in response to each stimulus image at each layer. For each convolutional subgroup, we first concatenated all filter activations (all convolution layers and max-pooling layer), then we performed a dimension reduction using Principal component analysis (PCA) on the concatenated activation vector to retain 95% of the total variance. For the fully connected layer, we only performed a PCA dimension reduction on the filter activation vector. We then computed the layer-wise RSM using the filter activation vectors of each pair of stimulus images. In total, we obtained 6 layer RSMs for the 5 convolutional subgroups and the first fully connected layer. The output layer was excluded in this analysis since it only outputs a binary sequence that indicates a 3-class label.

EEG vs. CNN. To establish the temporal correspondence between the layer representations of the CNN and the EEG, we calculated, for each subject, the Pearson correlation between the layer RSM with each of the EEG RSMs at all time windows, yielding a time course of the correlation for each layer. Subject-wise time course of the correlation between EEG and CNN for each layer was averaged to obtain the group-level time course. The significance of the correlation at each time window was determined using a permutation test where we randomly flipped the sign of the time course for each subject 10,000 times to obtain an empirical null distribution of the group-level correlation at each time window. We then used a cluster-mass correction [18] at $p < 0.05$ to account for multiple comparisons across time windows. The peak latency of the time course for each layer was selected as the time of the maximum correlation score

around 200 ms post-stimulus onset. We also computed R^2 between the peak latency and the layer number as a linear measure of the temporal hierarchy across CNN layers. The significance of the R^2 score was determined using a similar permutation procedure.

fMRI vs. CNN. To establish the spatial correspondence between the layer representations and the fMRI, we calculated the correlation between the layer RSM and each searchlight RSM across the whole brain for each subject. The subject-wise spatial correlation map of each layer was then averaged to obtain the group-level spatial correlation map. The significance of the spatial maps were determined using a similar permutation procedure with a threshold free cluster enhancement (TFCE) cluster correction to account for multiple comparisons in space at $p < 0.05$.

4. Results

4.1. CNN Performance

We evaluated the performance of the CNN on both the test image set and the stimulus image set. Our goal of training the CNN is to not only maximize the performance on the stimulus images but also to achieve the performance most similar to the subjects' behavioral performance. The CNN classification accuracy on the test image set (8,000 images) was 98.83%, which indicates a model that captures robust representations with good generalization. Then we evaluated the performance of CNN on the stimulus image set. For the 90 high coherence images, the accuracy of CNN was 100%. For the 90 low coherence images, the accuracy was 64.44% (93.33% on house images, 76.67% on car images, and 23.33% on face images). Table 1 lists the comparison of the performance between human subjects and the CNN. Although the overall CNN performance on the stimulus images is comparable to the human performance (94.00% at the high coherence, 58.00% at the low coherence), it is worth noting that, at the low coherence level, human subjects achieved their highest accuracy on the face category, suggesting a face perceptual bias, which was previously reported by Tu *et al*. [20] and shown being a result of network integrations in the brain. The CNN, however, achieved the highest accuracy on the house category. This is understandable since there is no ecological reason, as there is for humans, for the CNN to be biased towards faces. Higher accuracy for houses by the CNN is likely explainable by the simple fact that houses have much more linear structure and power in oriented spatial frequencies that are easier to represent and learn in a CNN model.

Human	High coherence	Low coherence
Face	96.14%	61.31%
Car	92.35%	53.10%
House	93.91%	60.25%
CNN	High coherence	Low coherence
Face	100%	23.33%
Car	100%	76.67%
House	100%	93.33%

Table 1. Accuracy of the CNN and human subjects on the stimulus image set

4.2. Temporal Correspondence Between CNN and EEG

For each CNN layer, we computed the Pearson correlation between the layer RSM and the EEG RSM across all time windows. The group-average temporal evolution of the correlation for all layers is shown in Figure 5A. All layers showed a significant correlation with the EEG representation around 200 ms except for layers 1 and 2. The increasing trend of peak latencies ($R^2 = 0.85$, $p = 0.007$) across CNN layers as shown in Figure 5B suggests a hierarchical correspondence between the temporal representation of the brain decision processing and the CNN multi-stage processing.

4.3. Spatial Correspondence Between CNN and fMRI

For each CNN layer, we computed the Pearson correlation between the layer RSM and the search light RSM across the whole brain volume. Figure 5C shows the group-average spatial correlation of each CNN layer and the brain. The low-level layers activate the sensory processing region such as lateral occipital cortex (LOC) [19], while the high-level layers correlate more with the attentional and executive control networks. The emergence of the DMN in mid-level layers (layers 3 and 4) and its interaction with the FPCN in high-level layers (layers 5 and 6) suggests a dynamical switching of the internal and external attentional processes during rapid decision making with perceptual ambiguity [15].

5. Discussion

Perceptual decision-making is believed to involve activation of a distributed brain networks, engaging sensory processing, attention allocation, working memory, decision formation, action generation and decision monitoring. In particular, when sensory evidence is ambiguous, the efficient reallocation of attentional resources between perceptual input processing (external attention) and internal bias and representation (internal attention) is likely important for task performance. We observed a correlation between the cascading layers in the CNN and spatial activations in the brain: at low-level layers, the CNN correlated with sensory processing regions; at the mid-level layers, the DMN emerged and interacted with the FPCN, a regulator of at-

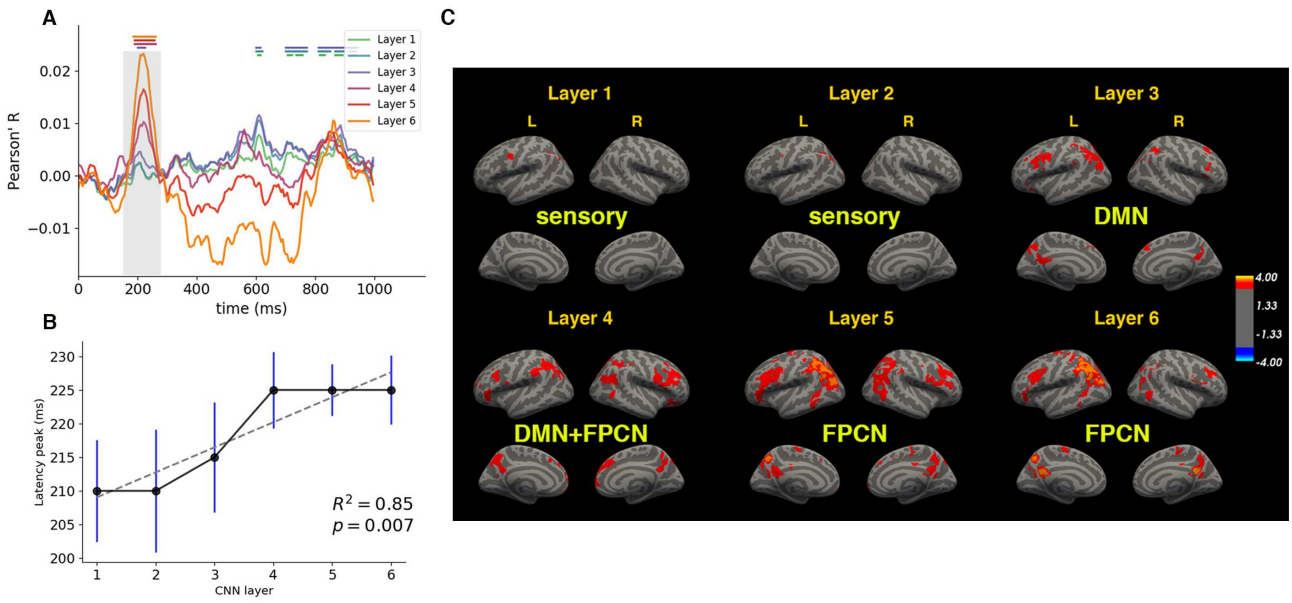

Figure 5. A, Pearson correlation was computed between the layer RSM and EEG RSMs across all time windows. The horizontal lines indicate significant time windows after cluster-correction at $p < 0.05$. B, The peak latency of each layer is around 200 ms. The peak latency increases as layer number increases ($R^2 = 0.85$, $p = 0.007$, permutation test). Error bar denotes standard error determined by a bootstrap technique. C, For each layer, the Pearson correlation was computed between the layer RSM and fMRI searchlight RSMs across the whole brain volume. Significant clusters were determined using a TFCE permutation test at $p < 0.05$, with Bonferroni correction across 12 hemispheres.

tention; at the high-level layers, the FPCN and motor areas were observed, where these regions are typically recruited to complete the decision processing. Interesting is that this cascade suggests a dynamical switching of attention between the external and internal focuses during rapid perceptual decision making with ambiguity. Future work will investigate this by conducting EEG-fMRI experiments which explicitly modulate a subject's internal and external attention during rapid decision making.

It is also worth discussing our methodology for training a CNN to perform the 3-category classification on face, car, house images that mapped to the task subjects did in our experiment. These images, which are from Imagenet and acquired in naturalistic settings, are very different from the stimulus images presented to the human subjects during the experiment (e.g. experiment images have grey background, luminance, contrast and spatial frequency equated, while naturalistic images used for the CNN are color and occur in scenes with context and taken from arbitrary viewpoint). In particular, the face stimuli used in the human experiment differ greatly from natural face images as they were generated from a morphable 3D face model. In order to have the CNN achieve human-level performance for this stimulus set, we included the synthetic face images in its training set. We used a small learning rate, a small number of training epochs and applied dropout on the FC layer to prevent overfitting. We also found that the more we fine-tuned the network on the natural image set, the worse the network performance would generalize to the stimulus images. Therefore, to increase the performance on the stimulus set matching the experiment, we created a training dataset sufficiently similar to the stimuli used in human experiment to train the network. This suggests the representations learned by large CNN models can be related to representations in biological brain networks. However care must be taken to make sure that the CNN models are tuned with a small amount of additional data to capture the particulars of the limited stimulus set and task governing the experiment.

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 3

[2] R. M. Cichy, A. Khosla, D. Pantazis, and A. Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358, 2017. 2

[3] 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 recogni-

tion reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016. 2, 3

[4] R. M. Cichy, D. Pantazis, and A. Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455, 2014. 3

[5] S. Dakin, R. Hess, T. Ledgeway, and R. Achtman. What causes non-monotonic tuning of fmri response to noisy images? *Current Biology*, 12(14):R476–R477, 2002. 2

[6] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017. 2

[7] U. Güçlü and M. A. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. 2

[8] U. Güçlü and M. A. van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017. 2

[9] H. R. Heekeren, S. Marrett, P. A. Bandettini, and L. G. Ungerleider. A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010):859, 2004. 1

[10] 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. 3

[11] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6:32672, 2016. 2

[12] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 554–561. IEEE, 2013. 3

[13] 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

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 3

[15] R. Leech and D. J. Sharp. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32, 2013. 1, 5

[16] J. Muraskin, T. R. Brown, J. M. Walz, T. Tu, B. Conroy, R. I. Goldman, and P. Sajda. A multimodal encoding model applied to imaging decision-related neural cascades in the human brain. *NeuroImage*, 2017. 2

[17] N. N. Oosterhof, A. C. Connolly, and J. V. Haxby. Cosmomvpa: multi-modal multivariate pattern analysis of neuroimaging data in matlab/gnu octave. *Frontiers in neuroinformatics*, 10:27, 2016. 4

[18] C. R. Pernet, N. Chauveau, C. Gaspar, and G. A. Rousselet. Limo eeg: a toolbox for hierarchical linear modeling of electroencephalographic data. *Computational intelligence and neuroscience*, 2011:3, 2011. 4

[19] M. G. Philiastides and P. Sajda. Eeg-informed fmri reveals spatiotemporal characteristics of perceptual decision making. *Journal of Neuroscience*, 27(48):13082–13091, 2007. 2, 5

[20] T. Tu, N. Schneck, J. Muraskin, and P. Sajda. Network configurations in the human brain reflect choice bias during rapid face processing. *Journal of Neuroscience*, 37(50):12226–12237, 2017. 1, 2, 5

[21] B. O. Turner, J. A. Mumford, R. A. Poldrack, and F. G. Ashby. Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage*, 62(3):1429–1438, 2012. 4

[22] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016. 1

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 3