Activations of Deep Convolutional Neural Network are Aligned with Gamma Band Activity of Human Visual Cortex

Ilya Kuzovkin[1,*], Raul Vicente[1,*,†], Mathilde Petton[2,3], Jean-Philippe Lachaux[2,3], Monica Baciu[4,5], Philippe Kahane[6,7], Sylvain Rheims[8,2,9], Juan R. Vidal[4,5] and Jaan Aru[1,10,*,†]

[1] *Computational Neuroscience Lab, Institute of Computer Science, University of Tartu, Estonia*
[2] *INSERM U1028, CNRS UMR5292, Brain Dynamics and Cognition Team, Lyon Neuroscience Research Center, Lyon, France*
[3] *Université Claude Bernard, Lyon, France*
[4] *University Grenoble Alpes, LPNC, F-38040 Grenoble, France*
[5] *CNRS, LPNC UMR 5105, F38040 Grenoble, France*
[6] *Inserm, U1216, F-38000 Grenoble, France*
[7] *Neurology Department, CHU de Grenoble, Hôpital Michallon, F-38000 Grenoble, France*
[8] *Department of Functional Neurology and Epileptology, Hospices Civils de Lyon and Université Lyon, Lyon, France*
[9] *Epilepsy Institute, Lyon, France*
[10] *Department of Penal Law, School of Law, University of Tartu, Estonia*

*Corresponding authors. E-mail: ilya.kuzovkin@gmail.com; jaan.aru@gmail.com; raulvicente@gmail.com
†These authors contributed equally to this work.

Previous work demonstrated a direct correspondence between the hierarchy of the human visual areas and layers of deep convolutional neural networks (DCNN) trained on visual object recognition. We used DCNNs to investigate which frequency bands carry feature transformations of increasing complexity along the ventral visual pathway. By capitalizing on intracranial depth recordings from 100 patients and 11293 electrodes we assessed the alignment between the DCNN and signals at different frequency bands in different time windows. We found that activity in low and high gamma bands was aligned with the increasing complexity of visual feature representations in the DCNN. These findings show that activity in the gamma band is not only a correlate of object recognition, but carries increasingly complex features along the ventral visual pathway. These results demonstrate the potential that modern artificial intelligence algorithms have in advancing our understanding of the brain.

Significance Statement

Recent advances in the field of artificial intelligence have revealed principles about neural processing, in particular about vision. Previous works have demonstrated a direct correspondence between the hierarchy of human visual areas and layers of deep convolutional neural networks (DCNNs), suggesting that DCNN is a good model of visual object recognition in primate brain. Studying intracranial depth recordings allowed us to extend previous works by assessing when and at which frequency bands the activity of the visual system corresponds to the DCNN. Our key finding is that signals in gamma frequencies along the ventral visual pathway are aligned with the layers of DCNN. Gamma frequencies play a major role in transforming visual input to coherent object representations.

Introduction

Biological visual object recognition is mediated by a hierarchy of increasingly complex feature representations along the ventral visual stream (DiCarlo et al., 2012). Intriguingly, these transformations are matched by the hierarchy of transformations learned by deep convolutional neural networks (DCNN) trained on natural images (Güçlü and van Gerven, 2015). It has been shown that DCNN provides the best model out of a wide range of neuroscientific and computer vision models for the neural representation of visual images in high-level visual cortex of monkeys (Yamins et al., 2014) and humans (Khaligh-Razavi and Kriegeskorte, 2014). Other studies have demonstrated with fMRI a direct correspondence between the hierarchy of the human visual areas and layers of the DCNN (Güçlü and van Gerven, 2015; Eickenberg et al., 2016; Seibert et al., 2016; Cichy et al., 2016b). In sum, the increasing feature complexity of the DCNN corresponds to the increasing feature complexity occurring in visual object recognition in the primate brain (Kriegeskorte, 2015; Yamins and DiCarlo, 2016).

However, fMRI based studies only allow one to localize object recognition in space, but biological visual object recognition is also specific in time and frequency. With time-resolved magnetoencephalography (MEG) recordings it has been demonstrated that the correspondence between the DCNN and neural signals peaks in the first 200 ms (Cichy et al., 2016b; Seeliger et al., 2017). Here we test the remaining dimension: that biological visual object recognition is also specific to certain frequencies. In particular, there is a long-standing hypothesis that especially gamma band $(30 - 150$ Hz) signals are crucial for object recognition (Singer and Gray, 1995; Singer, 1999; Fisch et al., 2009; Tallon-Baudry et al., 1997; Tallon-Baudry and Bertrand, 1999; Lachaux et al., 1999; Wyart and Tallon-Baudry, 2008; Lachaux et al., 2005; Vidal et al., 2006; Herrmann et al., 2004; Hipp et al., 2011; Gaillard et al., 2009; Srinivasan et al., 1999; Levy et al., 2015). Hence, if DCNN capture biological object recognition there should be a correspondence between the DCNN layers and gamma signals along the ventral visual pathway.

To empirically evaluate the specific role of gamma frequency in visual object recognition we assessed the alignment between the responses of layers of the DCNN and the neural signals in five distinct frequency bands and three time windows along the areas constituting the ventral visual pathway. Based on the previous findings we expected that: 1) mainly gamma frequencies should be aligned to the DCNN; 2) the correspondence between the DCNN and gamma should be confined to early time windows; 3) the correspondence between gamma and the DCNN layers should be restricted to visual areas. In order to test these predictions we capitalized on direct intracranial depth recordings from 100 patients with epilepsy and a total of 11293 electrodes implanted throughout the cerebral cortex.

Studying the alignment between the DCNN and gamma frequencies would also help to elucidate the role of gamma band signals in object recognition. The classic view is that gamma band activity signals the emergence of coherent object representations (Singer and Gray, 1995; Singer, 1999; Fisch et al., 2009). However, it is possible that gamma frequencies carry feature transformations of increasing complexity instead of reflecting solely the final product of object recognition. Suggestive evidence for this view is provided by studies demonstrating that feedforward activity from lower to higher visual areas is carried by the gamma frequencies along the ventral visual pathway (Van Kerkoerle et al., 2014; Bastos et al., 2015; Michalareas et al., 2016). The existence of quantifiable increase of feature complexity along the layers of DCNN allows one to use the DCNN as a computational model to assess whether signals in the gamma frequency indeed reflect such gradual transformations.

We observed that activity in the gamma range along the ventral pathway is statistically significantly aligned with the activity along the layers of DCNN: gamma $(31 - 150$ Hz) activity in the early visual areas correlates with the activity of early layers of DCNN, while the gamma activity of higher visual areas is better captured by the higher layers of the DCNN. We also found that neural activity in the theta range $(5 - 8$ Hz) throughout the visual hierarchy correlated with higher layers of DCNN.

Materials and Methods

Our methodology involves four major steps described in the following subsections. In "Patients and Recordings" we describe the visual recognition task and data collection. In "Processing of Neural Data" we describe the artifact rejection, extraction of spectral features and the electrode selection processes. "Processing of DCNN Data" shows how we extract activations of artificial neurons of DCNN that occur in responses to the same images as were shown to human subjects. In the final step we map neural activity to the layers of DCNN using representational similarity analysis. See Figure 1 for the illustration of the analysis workflow.

Patients and Recordings

100 patients of either gender with drug-resistant partial epilepsy and candidates for surgery were considered in this study and recruited from Neurological Hospitals in Grenoble and Lyon (France). All patients were stereotactically implanted with multilead EEG depth electrodes (DIXI Medical, Besançon, France). All participants provided written informed consent, and the experimental procedures were approved by local ethical committee of Grenoble hospital (CPP Sud-Est V 09-CHU-12). Recording sites were selected solely according to clinical indications, with no reference to the current experiment. All patients had normal or corrected to normal vision.

Electrode Implantation

Eleven to 15 semi-rigid electrodes were implanted per patient. Each electrode had a diameter of 0.8 mm and was comprised of 10 or 15 contacts of 2 mm length, depending on the target region, 1.5 mm apart. The coordinates of each electrode contact with their stereotactic scheme were used to anatomically localize the contacts using the proportional atlas of Talairach and Tournoux (Talairach and Tournoux, 1993), after a linear scale adjustment to correct size differences between the patients brain and the Talairach model. These locations were further confirmed by overlaying a post-implantation CT scan (showing contact sites) with a pre-implantation structural MRI with VOXIM® (IVS Solutions, Chemnitz, Germany), allowing direct visualization of contact sites relative to brain anatomy.

All patients voluntarily participated in a series of short experiments to identify local functional responses at the recorded sites (Vidal et al., 2010). The results presented here were obtained from a test exploring visual recognition. All data were recorded using approximately 120 implanted depth electrode contacts per patient with a sampling rates of 512 Hz, 1024 Hz or 2048 Hz. For the current analysis all recordings were downsampled to 512 Hz. Data were obtained in a total of 11293 recording sites.

Stimuli and Task

The visual recognition task lasted for about 15 minutes. Patients were instructed to press a button each time a picture of a fruit appeared on screen (visual oddball paradigm). Non-target stimuli consisted of pictures of objects of eight possible categories: houses, faces, animals, scenes, tools, pseudo words, consonant strings, and scrambled images. The target stimuli and last three categories were not included in this analysis. All the included

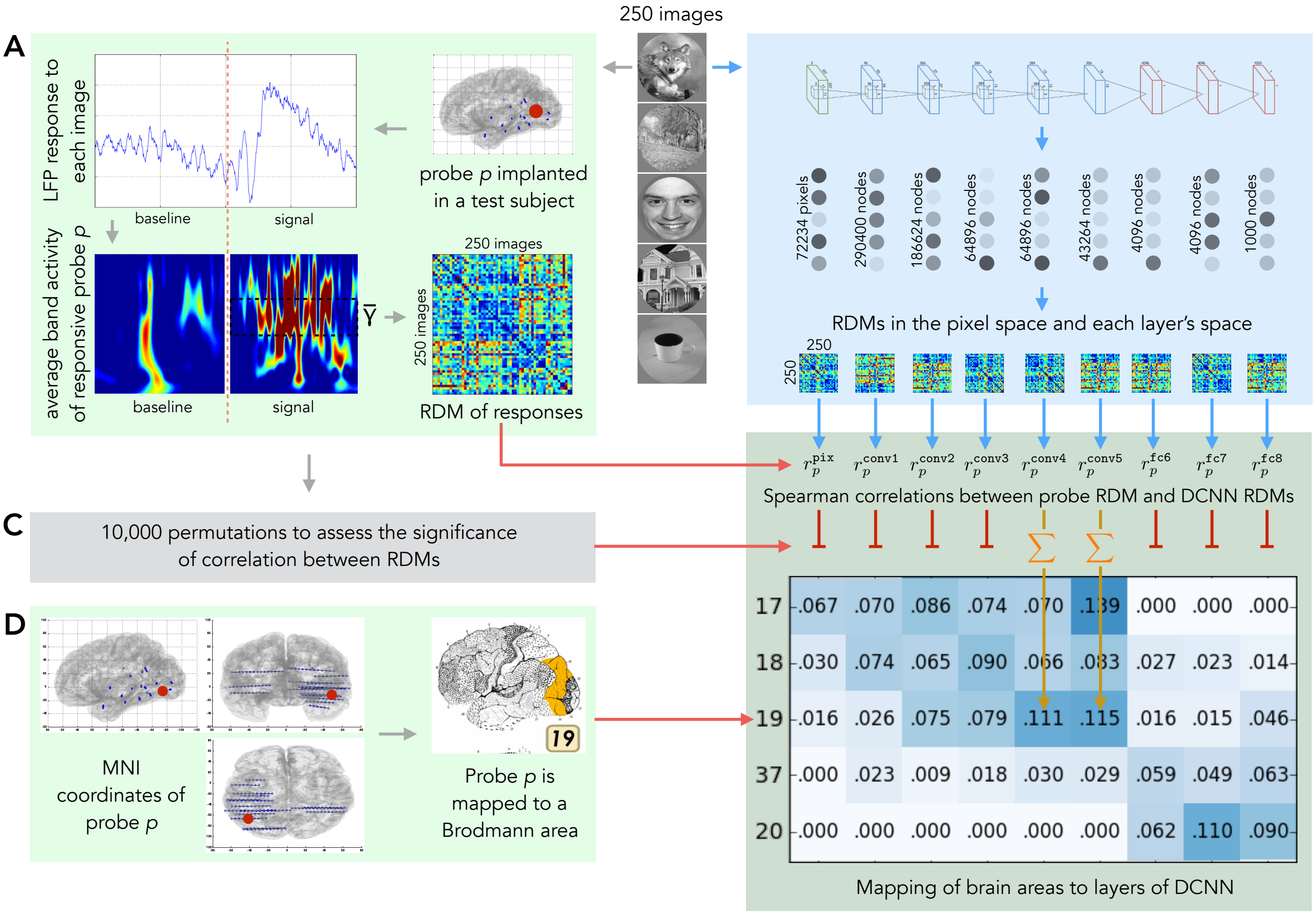
250 images
A
LFP response to each image
baseline
signal
probe p implanted in a test subject
average band activity of responsive probe p
baseline
signal
Ȳ
250 images
250 images
RDM of responses
B
72234 pixels
290400 nodes
186624 nodes
64896 nodes
64896 nodes
43264 nodes
4096 nodes
4096 nodes
1000 nodes
RDMs in the pixel space and each layer's space
250
250
C
10,000 permutations to assess the significance of correlation between RDMs
D
MNI coordinates of probe p
19
Probe p is mapped to a Brodmann area
E
$r_p^{\mathtt{pix}}$ $r_p^{\mathtt{conv1}}$ $r_p^{\mathtt{conv2}}$ $r_p^{\mathtt{conv3}}$ $r_p^{\mathtt{conv4}}$ $r_p^{\mathtt{conv5}}$ $r_p^{\mathtt{fc6}}$ $r_p^{\mathtt{fc7}}$ $r_p^{\mathtt{fc8}}$
Spearman correlations between probe RDM and DCNN RDMs
Σ Σ
17 .067 .070 .086 .074 .070 .139 .000 .000 .000
18 .030 .074 .065 .090 .066 .083 .027 .023 .014
19 .016 .026 .075 .079 .111 .115 .016 .015 .046
37 .000 .023 .009 .018 .030 .029 .059 .049 .063
20 .000 .000 .000 .000 .000 .000 .062 .110 .090
Mapping of brain areas to layers of DCNN

Figure 1 Overview of the analysis pipeline. 250 natural images are presented to human subjects (panel A) and to an artificial vision system (panel B). The activities elicited in these two systems are compared in order to map regions of human visual cortex to layers of deep convolutional neural networks (DCNNs). **A:** LFP response of each of 11293 electrodes to each of the images is converted into the frequency domain. Activity evoked by each image is compared to the activity evoked by every other image and results of this comparison are presented as a representational dissimilarity matrix (RDM). **B:** Each of the images is shown to a pre-trained DCNN and activations of each of the layers are extracted. Each layer's activations form a representation space, in which stimuli (images) can be compared to each other. Results of this comparison are summarized as a RDM for each DCNN layer. **C:** Subject's intracranial responses to stimuli are randomly reshuffled and the analysis depicted in panel A is repeated 10000 times to obtain 10000 random RDMs for each electrode. **D:** Each electrode's MNI coordinates are used to map the electrode to a Brodmann area. The figure also gives an example of electrode implantation locations in one of the subjects (blue circles are the electrodes). **E:** Spearman's rank correlation is computed between the true (non-permuted) RDM of neural responses and RDMs of each layer of DCNN. Also 10000 scores are computed with the random RDM for each electrode-layer pair to assess the significance of the true correlation score. If the score obtained with the true RDM is significant (the value of $p < 0.001$ is estimated by selecting a threshold such that none of the probes would pass it on the permuted data), then the score is added to the mapping matrix. The procedure is repeated for each electrode and the correlation scores are summed and normalized by the number of electrodes per Brodmann area. The resulting mapping matrix shows the alignment between the consecutive areas of the ventral stream and layers of DCNN.

stimuli had the same average luminance. All categories were presented within an oval aperture (illustrated on Figure 1). Stimuli were presented for a duration of 200 ms every $1000 - 1200$ ms in series of 5 pictures interleaved by 3-s pause periods during which patients could freely blink. Patients reported the detection of a target through a right-hand button press and were given feedback of their performance after each report. A 2-s delay was placed after each button press before presenting the follow-up stimulus in order to avoid mixing signals related to motor action with signals from stimulus presentation. Altogether, we measured responses to 250 natural images. Each image was presented only once.

Processing of Neural Data

The final dataset consists of 2823250 local field potential (LFP) recordings – 11293 electrode responses to 250 stimuli.

To remove the artifacts the signals were linearly detrended and the recordings that contained values $\geq 10\sigma_{images}$, where σ_{images} is the standard deviation of responses (in the time window from -500ms to 1000ms) of that particular probe over all stimuli, were excluded from data. All electrodes were re-referenced to a bipolar reference. The signal was segmented in the range from -500 ms to 1000 ms, where 0 marks the moment when the stimulus was shown. The -500 to -100 ms time window served as the baseline. There were three time windows in which the responses were measured: $50 - 250$ ms, $150 - 350$ ms and $250 - 450$ ms.

We analyzed five distinct frequency bands: θ ($5 - 8$ Hz), α ($9 - 14$ Hz), β ($15 - 30$ Hz), γ ($31 - 70$ Hz) and Γ ($71 - 150$ Hz). To quantify signal power modulations across time and frequency we used standard time-frequency (TF) wavelet decomposition (Daubechies, 1990). The signal $s(t)$ is convoluted with a complex Morlet wavelet $w(t, f_0)$, which has Gaussian shape in time (σ_t) and frequency (σ_f) around a central frequency f_0 and defined by $\sigma_f = 1/2\pi\sigma_t$ and a normalization factor. In order to achieve good time and frequency resolution over all frequencies we slowly increased the number of wavelet cycles with frequency ($\frac{f_0}{\sigma_f}$ was set to 6 for high and low gamma, 5 for beta, 4 for alpha and 3 for theta). This method allows obtaining better frequency resolution than by applying a constant cycle length (Delorme and Makeig, 2004). The square norm of the convolution results in a time-varying representation of spectral power, given by: $P(t, f_0) = |w(t, f_0)s(t)|^2$.

Further analysis was done on the electrodes that were responsive to the visual task. We assessed neural responsiveness of an electrode separately for each region of interest – for each frequency band and time window we compared the average post-stimulus band power to the average baseline power with a Wilcoxon signed-rank test for matched-pairs. All p-values from this test were corrected for multiple comparisons across all electrodes with a false discovery rate (FDR) procedure (Genovese et al., 2002). In the current study we deliberately kept only positively responsive electrodes, leaving the electrodes where the post-stimulus band power was significantly weaker than the average baseline power for future work. Table 1 contains the numbers of electrodes that were used in the final analysis in each of 15 regions of interest across the time and frequency domains.

	θ	α	β	γ	Γ
$50 - 250$ ms	1299	709	269	348	504
$150 - 350$ ms	1689	783	260	515	745
$250 - 450$ ms	1687	802	304	555	775

Table 1 Number of positively responsive electrodes in each of the 15 regions of interest in a time-resolved spectrogram.

Each electrode's MNI coordinates were mapped to a corresponding Brodmann brain area (Brodmann, 1909) using Brodmann area atlas contained in MRICron (Rorden, 2007) software.

To summarize, once the neural signal processing pipeline is complete, each electrode's response to each of the stimuli is represented by one number – the average band power in a given time window normalized by the baseline. The process is repeated independently for each time-frequency region of interest.

Processing of DCNN Data

We feed the same images that were shown to the test subjects to a deep convolutional neural network (DCNN) and obtain activations of artificial neurons (nodes) of that network. We use `Caffe` (Jia et al., 2014) implementation of `AlexNet` (Krizhevsky et al., 2012) architecture (see Figure 7) trained on `ImageNet` (Russakovsky et al., 2015) dataset to categorize images into 1000 classes. Although the image categories used in our experiment are not exactly the same as the ones in the `ImageNet` dataset, they are a close match and DCNN is successful in labelling them.

The architecture of the `AlexNet` artificial network can be seen on Figure 7. It consists of 9 layers. The first is the input layer, where one neuron corresponds to one pixel of an image and activation of that neuron on a scale from 0 to 1 reflects the color of that pixel: if a pixel is black, the corresponding node in the network is not activated at all (value is 0), while a white pixel causes the node to be maximally activated (value 1). After the input layer the network has 5 *convolutional layers* referred to as `conv1-5`. A convolutional layer is a collection of filters that are applied to an image. Each filter is a small image that represents a particular visual pattern. A filter is applied to every possible position on an input image and if the underlying patch of an image coincides with the pattern that the filter represents, the filter becomes activated and translates this activation to the artificial neuron in the next layer. That way, nodes of `conv1` tell us where on the input image each particular visual pattern occurred. Hierarchical structure of convolutional layers gives rise to the phenomenon we are investigating in this work – increase of complexity of visual representations in each subsequent layer of the visual hierarchy: in both the biological and artificial systems. Convolutional layers are followed by 3 *fully-connected* layers (`fc6-8`). Each node in a fully-connected layer is, as the name suggests, connected to every node of the previous layer allowing the network to decide which of those connections are to be preserved and which are to be ignored.

For each of the images we store the activations of all nodes of DCNN. As the network has 9 layers we obtain 9 representations of each image: the image itself (referred to as layer 0) in the pixel space and the activation values of each of the layers of DCNN. See pane B of figure 1 for the cardinalities of those feature spaces.

Mapping Neural Activity to Layers of DCNN

Once we extracted the features from both neural and DCNN responses our next goal was to compare the two and use a similarity score to map the brain area where a probe was located to a layer of DCNN. By doing that for every probe in the dataset we obtained cross-subject alignment between visual areas of human brain and layers of DCNN. There are multiple deep neural network architectures trained to classify natural images. Our choice of `AlexNet` does not imply that this particular architecture corresponds best to the hierarchy of visual layers of human brain. It does, however, provide a comparison for hierarchical structure of human visual system and was selected among other architectures due to its relatively small size and thus easier interpretability.

Recent studies comparing the responses of visual cortex with the activity of DCNN have used two types of mapping methods. The first type is based on linear regression models that predict neural responses from DCNN activations (Güçlü and van Gerven, 2015). The second is based on representational similarity analysis (RSA) (Kriegeskorte et al., 2008). We used RSA to compare distances between stimuli in the neural response space and in the DCNN activation space (Cichy et al., 2016a).

Representational Dissimilarity Matrices

We built a representation dissimilarity matrix (RDM) of size *number of stimuli* × *number of stimuli* (in our case 250 × 250) for each of the probes and each of the layers of DCNN. Given a matrix $\mathrm{RDM}^{\text{feature space}}$ a value $\mathrm{RDM}_{ij}^{\text{feature space}}$ in the ith row and jth column of the matrix shows the Euclidean distance between the vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ that represent images i and j respectively in that particular feature space. Note that the preprocessed neural response to an image in a given frequency band and time window is a scalar, and hence correlation distance is not applicable. Also, given that DCNNs are not invariant to the scaling of the activations or weights in any of its layers, we preferred to use closeness in Euclidean distance as a more strict measure of similarity. In our case there are 10 different features spaces in which an image can be represented: the original pixel space, 8 feature spaces for each of the layers of the DCNN and one space where an image is represented by the preprocessed neural response of probe p. For example, to analyze region of interest of high gamma in $50-250$ ms time window we computed 504 RDM matrices on the neural responses – one for each positively responsive electrode in that region of interest (see Table 1), and 9 RDM matrices on the activations of the layers of DCNN. A pair frequency band and a time window, such as "high gamma in 50-250 ms window" is referred to as *region of interest* in this work.

Representational Similarity Analysis

The second step was to compare the $\mathrm{RDM}^{\text{probe } p}$ of each probe p to RDMs of layers of DCNN. We used Spearman's rank correlation as measure of similarity between the matrices:

$$\rho_{\text{layer } l}^{\text{probe } p} = \mathrm{Spearman}(\mathrm{RDM}^{\text{probe } p}, \mathrm{RDM}^{\text{layer } l}). \qquad (1)$$

As a result of comparing $\mathrm{RDM}^{\text{probe } p}$ with every $\mathrm{RDM}^{\text{layer } l}$ we obtain a vector with 9 scores: $(\rho_{\text{pixels}}, \rho_{\text{conv1}}, \ldots, \rho_{\texttt{fc8}})$ that serves as a distributed mapping of probe p to the layers of DCNN (see pane E of Figure 1). The procedure is repeated independently for each probe in each region of interest.

Statistical significance and controls

To assess the statistical significance of the correlations between the RDM matrices we run a permutation test. In particular, we reshuffled the vector of brain responses to images 10000 times, each time obtaining a dataset where the causal relation between the stimulus and the response is destroyed. On each of those datasets we ran the analysis and obtained Spearman's rank correlation scores. To determine score's significance we compared the score obtained on the original (unshuffled) data with the distribution of scores obtained with the surrogate data. If the score obtained on the original data was bigger than value obtained on the surrogate sets with $p < 0.001$ significance we considered the

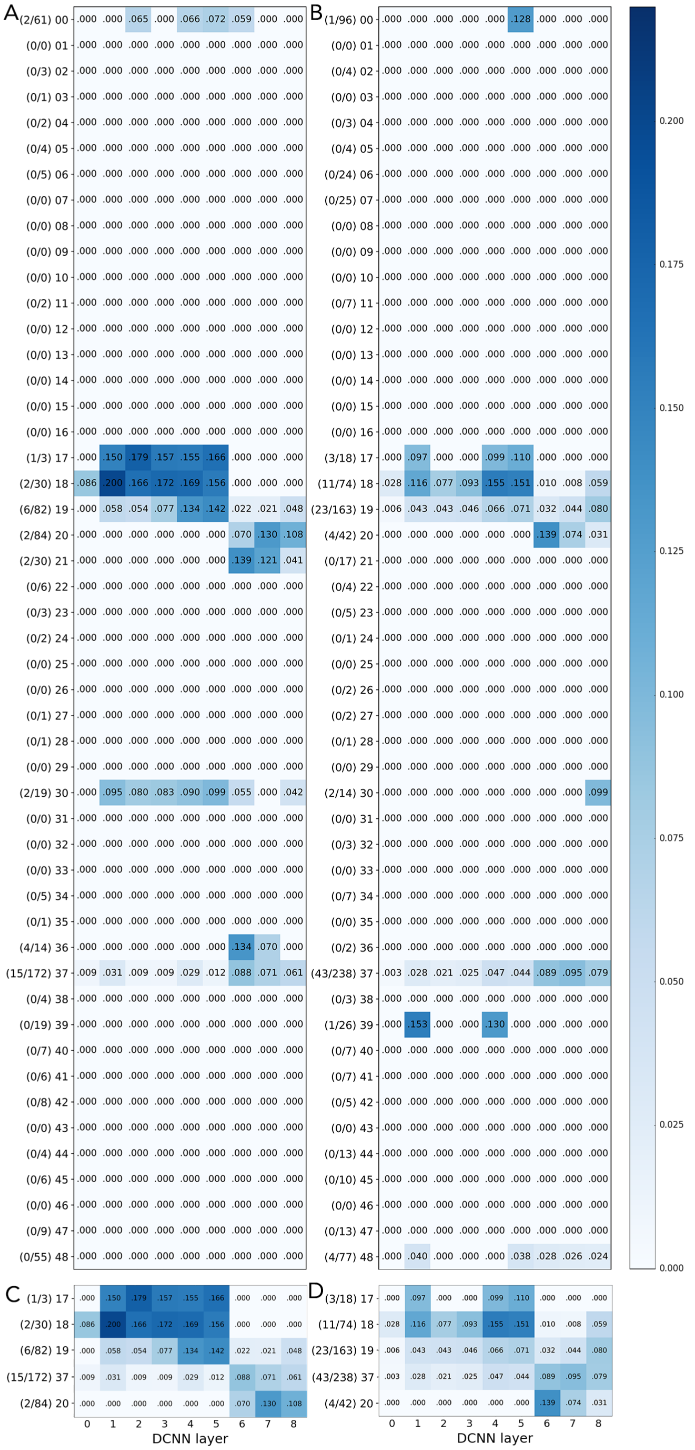

Figure 2 Mapping of the activity in Brodmann areas to DCNN layers. Underlying data comes from the activity in low gamma (31-70 Hz, subfigures A and C) and high gamma (71-150 Hz, subfigures B and D) bands in 150-350 ms time window. C and D are subselection of the areas that constitute ventral stream: 17, 18, 19, 37, 20. There are two important observations to made out of this plot: a) statistically significant neural responses are specific to visual areas b) the alignment between the ventral stream and layer of DCNN is clearly visible. Area 0 contains the regions of the brain not mapped by the atlas. The numbers on the left of each panel show the number of significantly correlating probes in each area out of the total number of responsive probes in that area.

score to be significantly different. The threshold of $p = 0.001$ is estimated by selecting such a threshold that on the surrogate data none of the probes would pass it.

To size the effect caused by training artificial neural network on natural images we performed a control where the whole analysis pipeline depicted on figure 1 is repeated using activations of a network that was not trained – its weights are randomly sampled from a Gaussian distribution $\mathcal{N}(0, 0.01)$.

Quantifying properties of the mapping

To evaluate the results quantitatively we devised a set of measures specific to our analysis. *Volume* is the total sum of significant correlations (see Equation 1) between the probes in a subset of brain areas A and DCNN layers L:

$$V^{\text{areas } A}_{\text{layers } L} = \sum_{a \in A} \sum_{l \in L} \sum_{p \in S^a_l} \rho^{\text{probe } p}_{\text{layer } l}, \tag{2}$$

where A is a subset of brain areas, L is a subset of layers, and S^a_l is the set of all probes in area a that significantly correlate with layer l.

We express *volume of visual activity* as

$$V^{\{17,18,19,37,20\}}_{\text{all layers}}, \tag{3}$$

which shows the total sum of correlation scores between all layers of the network and the Brodmann areas that are located in the ventral stream: $17, 18, 19, 37$, and 20.

Visual specificity of activity is the ratio of volume in visual areas and volume in all areas together, for example visual specificity of all of the activity in the ventral stream that significantly correlates with any of layers of DCNN is

$$S^{\{17,18,19,37,20\}}_{\text{all layers}} = \frac{V^{\{17,18,19,37,20\}}_{\text{all layers}}}{V^{\text{all areas}}_{\text{all layers}}} \tag{4}$$

The measures so far did not take into account hierarchy of the ventral stream nor the hierarchy of DCNN. The following two measures are the most important quantifiers we rely on in presenting our results and they do take hierarchical structure into account.

The *ratio of complex visual features to all visual features* is defined as the total volume mapped to layers `conv5`, `fc6`, `fc7` divided by the total volume mapped to layers `conv1`, `conv2`, `conv3`, `conv5`, `fc6`, `fc7`:

$$C^{\text{areas } A} = \frac{V^{\text{areas } A}_{\texttt{conv5,fc6,fc7}}}{V^{\text{areas } A}_{\texttt{conv1,conv2,conv3,conv5,fc6,fc7}}}. \tag{5}$$

Note that for this measure layers `conv4` and `fc8` are omitted: layer `conv4` is considered to be the transition between the layers with low and high complexity features, while layer `fc8` directly represents class probabilities and does not carry visual representations of the stimuli (if only on very abstract level).

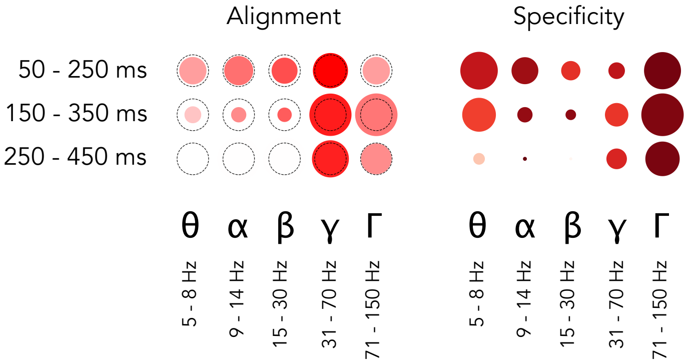

Figure 3 Overall relative statistics of brain responses across frequency bands and time windows. The left panel shows the alignment between visual brain areas and DCNN layers (see Equation 6). The color indicates the correlation value (ρ) while the size of the marker shows the logarithm (so that not significant results are still visible on the plot) of inverse of the statistical significance of the correlation, dotted circle indicates $p = 0.0003(3)$ – the Bonferroni-corrected significance threshold level of 0.005. The right panel shows whether activity in a region of interest is specific to visual areas (see Equation 4): intensive red means that most of the activity in that band and time window happened in visual areas, size of the marker indicates total volume (Equation 2) of activity in all areas. The maximal size of a marker is defined by the biggest marker on the figure.

Finally, the *alignment* between the activity in the visual areas and activity in DCNN is estimated as Spearman's rank correlation between the vector of electrode assignments to visual areas and the vector of electrode assignments to DCNN layers:

$$\rho = \text{Spearman}\left(\begin{array}{ll}\text{Brodmann areas with sig-} & \text{DCNN layers where} \\ \text{nificantly correlating probes} & \text{significantly correlating} \\ \text{ordered by the hierarchy of,} & \text{probes are mapped,} \\ \text{the ventral stream: BA17,} & \text{ordered by the hierarchy} \\ \text{BA18, BA19, BA37, BA20} & \text{of DCNN architecture}\end{array}\right). \tag{6}$$

We note that although the hierarchy of the ventral stream is usually not defined through the progression of Brodmann areas, such ordering nevertheless provides a reasonable approximation of the real hierarchy (Lerner et al., 2001; Grill-Spector and Malach, 2004). As both the ventral stream and the hierarchy of layers in DCNN have an increasing complexity of visual representations, the relative ranking within the biological system should coincide with the ranking within the artificial system. Based on the recent suggestion that significance levels should be shifted to 0.005 (Dienes et al., 2017) and after Bonferroni-correcting for 15 time-frequency windows we accepted alignment as significant when it passed $p < 0.0003(3)$.

Results

Increasing complexity of visual representations is captured by activity in gamma band

We tested the hypothesis that gamma activity carries increasingly complex features along the ventral stream. To that end we assessed the alignment of neural activity in different frequency bands and time windows to the activity of layers of a DCNN.

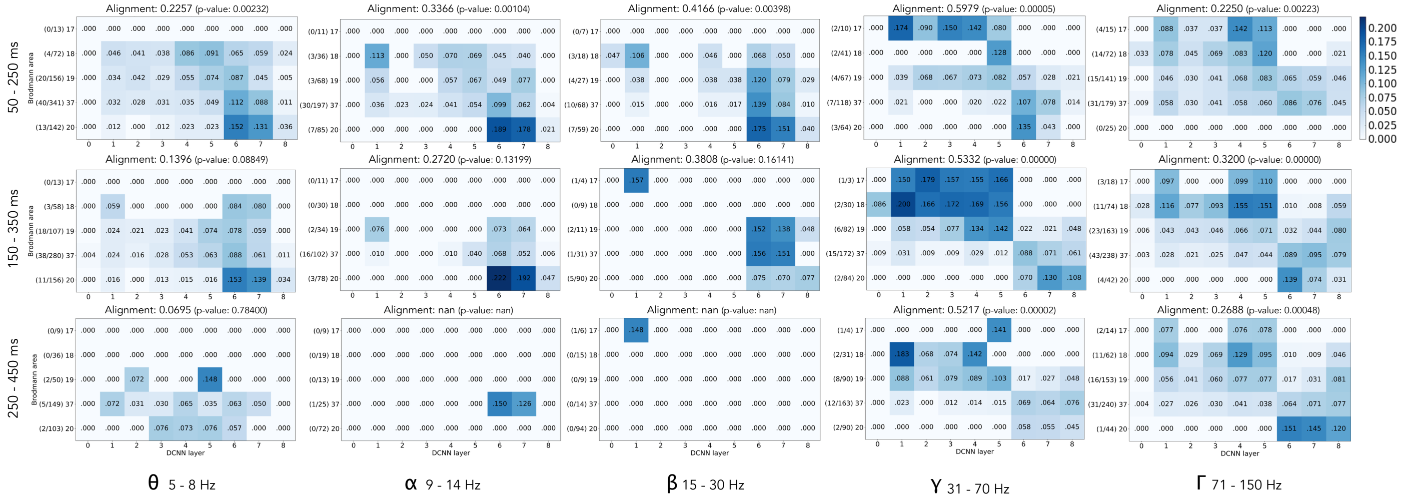

Figure 4 Mapping of activity in visual areas to activations of layers of DCNN across five frequency bands and three time windows. The alignment score is computed as Spearman's rank correlation between electrode assignment to Brodmann areas and electrode assignment to DCNN layers (Equation 6). The numbers on the left of each subplot show the number of significantly correlating probes in each area out of the total number of responsive probes in that area.

In particular, we used RSA to compare the representational geometry of different DCNN layers and the activity patterns of different frequency bands of single electrodes (see Figure 1). We consistently found that signals in low gamma (31 – 70 Hz) frequencies across all time windows and high gamma (71 – 150 Hz) frequencies in 150 – 350 ms window are aligned with the DCNN in a specific way: increase of the complexity of features along the layers of the DCNN was matched by the transformation in the representational geometry of responses to the stimuli along the ventral stream. In other words, the lower and higher layers of the DCNN explained gamma band signals from earlier and later visual areas, respectively.

Figure 2 illustrates assignment of neural activity in low gamma band (panel A) and high gamma band (panel B) to Brodmann areas and layers of DCNN. As one can see most of the activity was assigned to visual areas (areas 17, 18, 19, 37, 20). Focusing on visual areas (panels C, D) revealed a diagonal trend that illustrated the alignment between ventral stream and layers of DCNN. Our findings across all subjects, time windows and frequency bands are presented in table 2 and summarized on the left panel of figure 3. The results in table 2 show the comparison of alignment between DCNN and brain areas with both random and pre-trained networks. We can see that training a network to classify natural images drastically increases the alignment score ρ and its significance. We note that the alignment in the gamma bands is also present at the single-subject level as can be seen in Figure 6.

Apart from the alignment we looked at the total amount of correlation and its specificity to visual areas. On the right panel of Figure 3 we can see that the volume of significantly correlating activity was highest in the high gamma range. Remarkably, 97% of that activity was located in visual areas, which is confirmed by figure 2 where we see that in the gamma range only a few electrodes were assigned to Brodmann areas that are not part of the ventral stream.

		Alignment with layers of randomly initialized `AlexNet`		Alignment with layers of `AlexNet` trained on `ImageNet`		
Band	**Window**	ρ	**p-value**	ρ	**p-value**	
θ	50-250 ms	0.0632	0.71	0.2257	0.00231575	*
θ	150-350 ms	-0.1013	0.59	0.1396	0.08848501	
θ	250-450 ms	0.1396	0.59	0.0695	0.78400416	
α	50-250 ms	-0.2411	0.32	0.3366	0.00103551	*
α	150-350 ms	0.0000	1.00	0.2720	0.13199463	
α	250-450 ms	–	–	–	–	
β	50-250 ms	–	–	0.4166	0.00397929	
β	150-350 ms	–	–	0.3808	0.16141286	
β	250-450 ms	–	–	–	–	
γ	50-250 ms	0.1594	0.62	0.5979	0.00004623	***
γ	150-350 ms	-0.1688	0.34	0.5332	0.00000059	***
γ	250-450 ms	-0.1132	0.56	0.5217	0.00001624	***
Γ	50-250 ms	0.0869	0.42	0.2259	0.00222940	*
Γ	150-350 ms	-0.0053	0.96	0.3200	0.00000051	***
Γ	250-450 ms	-0.1361	0.33	0.2688	0.00047999	*

Table 2 Alignment ρ score and significance for all 15 regions of interest. * indicates the alignments that pass p-value threshold of 0.05 Bonferroni-corrected to 0.003(3) and *** the ones that pass 0.005 (Dienes et al., 2017) Bonferroni-corrected to 0.0003(3). Note how the values differ between random (control) network and a network trained on natural images. Visual representation of alignment and significance is given on the left pane of Figure 3.

Activity in other frequency bands

To test the specificity of gamma frequency in visual object recognition, we assessed the alignment between the DCNN and other frequencies. The detailed mapping results for all frequency bands

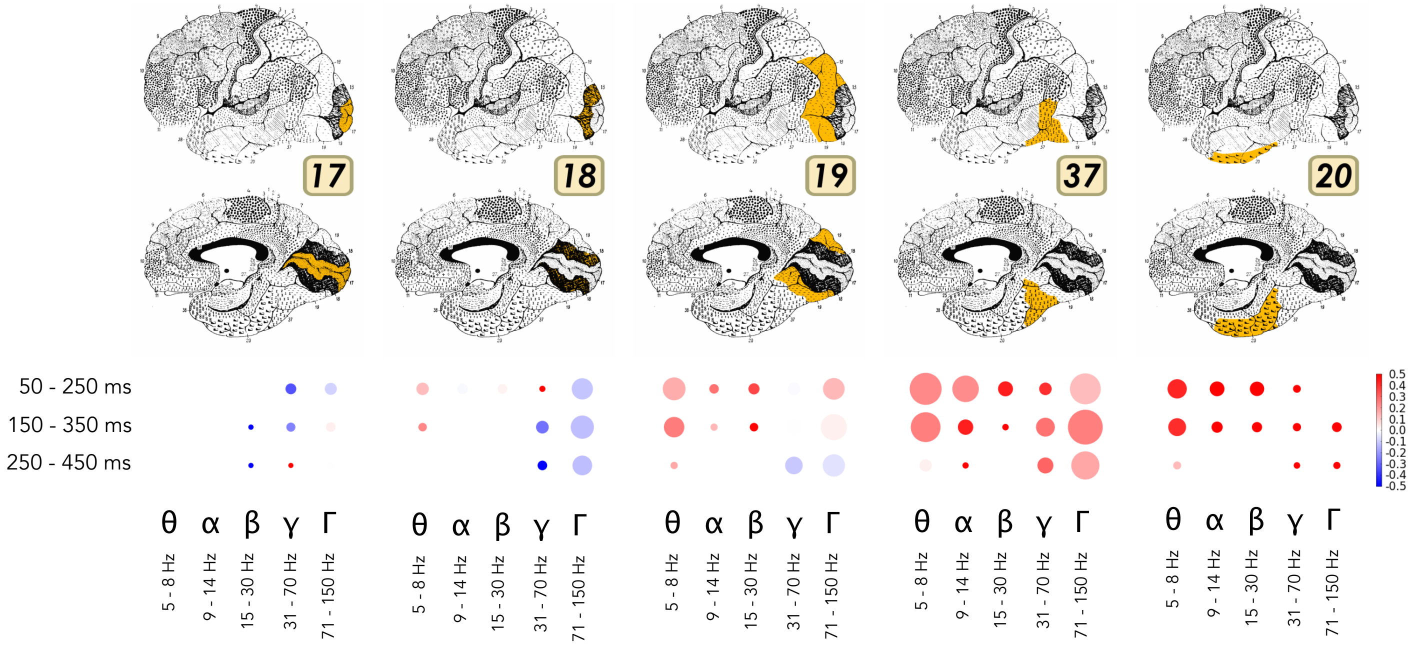

Figure 5 Area-specific analysis of volume of neural activity and complexity of visual features represented by that activity. Size of the marker shows the sum of correlation coefficients between the area and DCNN for each particular band and time window. Color codes the ratio of complex visual features to simple visual features, i.e. the comparison between the activity that correlates with the higher layers (`conv5, fc6, fc7`) of DCNN to the lower layers (`conv1`, `conv2`, `conv3`). Intensive red means that the activity was correlating more with the activity of higher layers of DCNN, while the intensive blue indicates the dominance of correlation with the lower areas. If the color is close to white then the activations of both lower and higher layers of DCNN were correlating with the brain responses in approximately equal proportion.

and time windows are are presented in figures 3 and 4. We can see that weaker alignment (that does not survive the Bonferroni correction) is present in early time window in theta and alpha frequency range. No alignment is observed in the beta band.

To investigate the involvement of each frequency band more closely we analyzed each visual area separately. Figure 5 shows the volume of activity in each area (size of the marker on the figure) and whether that activity was more correlated with the complex visual features (red color) or simple features (blue color). In our findings the role of the earliest area (17) was minimal, however that might be explained by a very low number of electrodes in that area in our dataset (less that 1%). One can see from figure 5 that activity in theta frequency in time windows $50 - 250$ ms and $150 - 350$ ms had large volume and is correlated with the higher layers of DCNN in higher visual areas (19, 37, 20) of the ventral stream. This hints at the role of theta activity in visual object recognition. In general, in areas 37 and 20 all frequency bands carried information about high level features in the early time windows. This implies that already at early stages of processing the information about complex features was present in those areas.

Gamma activity is more specific to convolutional layers, while the activity in lower frequency bands is more specific to fully connected layers

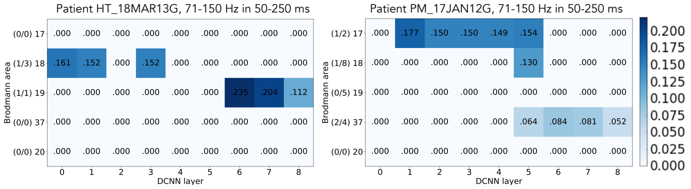

Figure 6 Single subject results from two different subjects. The numbers show the sum of correlations normalized by the number of probes in an area. On the left plot we see how a probe in Brodmann area 18 is mapped to the layers 0, 1, and 3 DCNN, while the activity in Brodmann area 19, which is located further along the ventral stream, is mapped to the higher layers of DCNN: 6, 7, 8. Similar trend is seen on the right plot. The numbers on the left of each subplot show the number of significantly correlating probes in each area out of the total number of responsive probes in that area.

We analysed volume and specificity of brain activity that correlates with each layer of DCNN separately to see if any bands or time windows are specific to particular level of hierarchy of visual processing in DCNN. Figure 7 presents a visual summary of this analysis. In the "Methods" section we have defined total volume of visual activity in layers L. We used this measure to quantify the activity in low and high gamma bands. We noticed that while the fraction of gamma activity that is mapped to convolutional layers is high ($\frac{\bar{V}^{\gamma,\Gamma}_{\{\texttt{conv1}\ldots\texttt{conv5}\}}}{\bar{V}^{\text{all bands}}_{\{\texttt{conv1}\ldots\texttt{conv5}\}}} = 0.71$), this fraction diminished in fully connected layers `fc6` and `fc7` ($\frac{\bar{V}^{\gamma,\Gamma}_{\{\texttt{fc6,fc7}\}}}{\bar{V}^{\text{all bands}}_{\{\texttt{fc6,fc7}\}}} = 0.39$). Note that `fc8` was excluded as it represents class label probabilities and

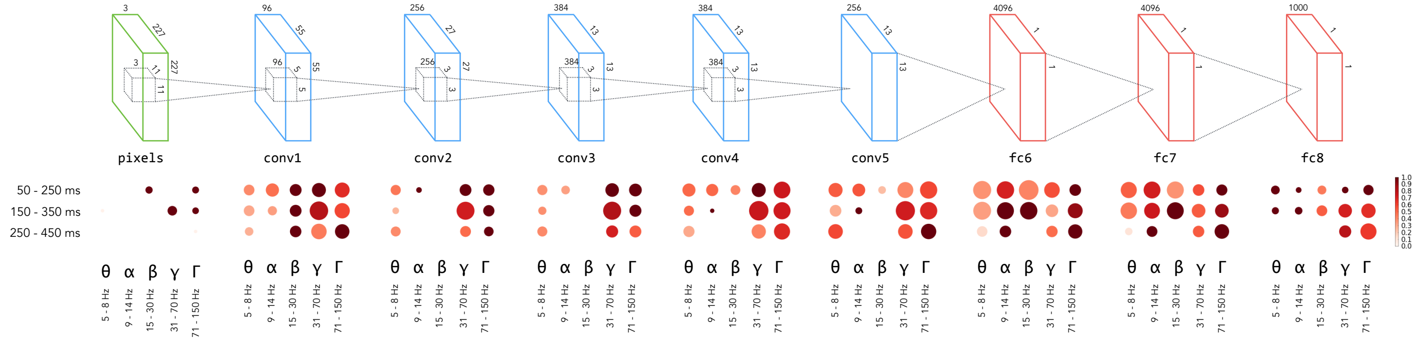

Figure 7 Specificity of neural responses across frequency bands and time windows for each layer of DCNN. Size of a marker is the total activity mapped to this layer and the intensity of the color is the specificity of the activity to visual areas.

does not carry information about visual features of the objects. On the other hand the activity in lower frequency bands (theta, alpha, beta) showed the opposite trend – fraction of volume in convolutional layers was 0.29, while in fully connected it growed to 0.61. This observation highlighted the fact that visual features extracted by convolutional filters of DCNN carry the signal that is more similar to the signal carried by gamma frequency bands, while the fully connected layers that do not directly correspond to intuitive visual features, carry information that has more in common with the activity in the lower frequency bands.

Discussion

The recent advances in artificial intelligence research have been breathtaking. Not only do the deep neural networks match human performance in visual object recognition, they also provide the best model for how biological object recognition happens (Kriegeskorte, 2015; Yamins and DiCarlo, 2016). Previous work has established a correspondence between hierarchy of the DCNN and the fMRI responses measured across the human visual areas (Güçlü and van Gerven, 2015; Eickenberg et al., 2016; Seibert et al., 2016; Cichy et al., 2016b). Further research has shown that the activity of the DCNN matches the biological neural hierarchy in time as well (Cichy et al., 2016b; Seeliger et al., 2017). Studying intracranial recordings allowed us to extend previous findings by assessing the alignment between the DCNN and cortical signals at different frequency bands. As there is a quantifiable increase of the complexity of features along the layers of the DCNN, any signal that is aligned to the DCNN has to carry similarly increasingly complex features built-up during visual object recognition. We observed that the lower layers of the DCNN explained gamma band signals from earlier visual areas, while higher layers of the DCNN, responsive for more complex features, matched with the gamma band signals from higher visual areas. Correspondence between layers of DCNN and visual hierarchy of human brain was present not only at the extremes, but also at the intermediate layers of the hierarchy. Hence, one can conclude that gamma band carries increasingly complex features required for object recognition along the ventral visual pathway. This finding confirms previous work that has given a central role for gamma band activity in visual object recognition (Singer and Gray, 1995; Singer, 1999; Fisch et al., 2009) and feedforward communication (Van Kerkoerle et al., 2014; Bastos et al., 2015; Michalareas et al., 2016). However, importantly, our results show that gamma activity reflects not only object recognition per se but also the feature transformations that are computed on the way towards explicit object representations. Our work demonstrates that the correlation between the DCNN and the biological counterpart is specific not only in space and time, but also in frequency.

Feedforward and feedback computations in object recognition

Visual object recognition in the brain involves both feedforward and feedback computations (DiCarlo et al., 2012; Kriegeskorte, 2015). What do our results reveal about the nature of feedforward and feedback compoments in visual object recognition? We observed that the DCNN corresponds to the biological processing hierarchy even in the latest analysed time-window (Figure 3). In a directly relevant previous work Cichy and colleagues compared DCNN representations to millisecond resolved MEG data from humans (Cichy et al., 2016b). There was a positive correlation between the layer number of the DCNN and the peak latency of the correlation time course between the respective DCNN layer and MEG signals. In other words, deeper layers of the DCNN predicted later brain signals. As evidenced on Figure 3 in (Cichy et al., 2016b), the correlation between DCNN and MEG activity peaked between ca 100 and 160 ms for all layers, but significant correlation persisted well beyond that time-window. However, in the work of (Cichy et al., 2016b) the correlation decreased over time, while in our data we evidenced no such clear drop in the later time windows: even between 250-450 ms the alignment in low gamma was strong and significant.

How could this late alignment be interpreted? In particular, feedforward object recognition is thought to be finished in ca 200-250 milliseconds after stimulus onset (DiCarlo et al., 2012). Hence, one could think that the correspondence between the DCNN, which is a feedforward network and biological visual object recognition should be confined to early time windows. However, although the DCNN is a purely feedforward network it is important to notice that the alignment between electrophysiological signals and the DCNN does not imply that the respective signals have to reflect feedforward computations. Such alignment only means that the progressive changes in representational geometry along the processing hierarchy are similar to the DCNN. In other words, it is possible that the activity patterns observed are a result of recurrent computations, but their outcome representational geometry resembles that of the DCNN. Therefore, the present results together with previous findings (Cichy et al., 2016b) demonstrate that the DCNN is a good model not only for feedforward object recognition, but also for the later phases, which most likely include feedback computations. This fits with the predictive coding framework where the feedback activity is not an unspecific modulatory signal but rather has to signal specific contents from higher to lower levels of the processing hierarchy (Bastos et al., 2012). Hence, within this theoretical framework, a specific representational geometry is expected even from a feedback channel.

Low vs high gamma in object recognition

We observed significant alignment to the DCNN in both low and high gamma bands. However, for high gamma this alignment was more restricted in time, surviving correction only in the middle time window. Previous studies have shown that low and high gamma frequencies are functionally different: while low gamma is more related to classic narrow-band gamma oscillations, high frequencies seem to reflect local spiking activity rather than oscillations (Manning et al., 2009; Ray and Maunsell, 2011), the distinction between low and high gamma activity has also implications from cognitive processing perspective (Vidal et al., 2006; Wyart and Tallon-Baudry, 2008). In the current work we approached the data analysis from the machine learning point of view and remained agnostic with respect to the oscillatory nature of underlying signals. Importantly, we found that numerically the alignment to the DCNN was stronger and persisted for longer in low gamma frequencies. However, high gamma was more prominent when considering volume and specificity to visual areas. The most striking difference between the low and high gamma with regard to specificity was in the earliest time window 50-250 ms where the correlation between the DCNN and high gamma was almost exclusive to visual areas.

Limitations

The present work relies on data pooled over the recordings from 100 subjects. Hence, the correspondence we found between responses at different frequency bands and layers of DCNN is distributed over many subjects. While it is expected that single subjects show similar mappings (see also Figure 6), the variability in number and location of recording electrodes in individual subjects makes it difficult a full single-subject analysis with this type of data. We also note that the mapping between electrode locations and Brodmann areas is approximate and the exact mapping would require individual anatomical reconstructions and more refined atlases.

Future work

Intracranial recordings are both precisely localized in space and time, thus allowing us to explore phenomena not observable with fMRI. In this work we investigated the correlation of DCNN activity with five broad frequency bands and three time windows. Our next steps will include the analysis of the activity on a more granular temporal and spectral scale. Replacing representation similarity analysis with a predictive model (such as regularized linear regression) will allow us to explore which visual features elicited the highest responses in the visual cortex.

Acknowledgements

IK, RV and JA thank the financial support from the Estonian Research Council through the personal research grants PUT438 and PUT1476. This work was supported by the Estonian Centre of Excellence in IT (EXCITE), funded by the European Regional Development Fund.

References

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical microcircuits for predictive coding. *Neuron* 76:695–711.

Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P (2015) Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* 85:390–401.

Brodmann K (1909) *Vergleichende Lokalisationslehre der Groshirnrinde* Barth.

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016a) Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970* .

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016b) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports* 6.

Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory* 36:961–1005.

Delorme A, Makeig S (2004) Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods* 134:9–21.

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? *Neuron* 73:415–434.

Dienes Z, Field A et al. (2017) Redefine statistical significance. *Nature Human Behaviour* .

Eickenberg M, Gramfort A, Varoquaux G, Thirion B (2016) Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* .

Fisch L, Privman E, Ramot M, Harel M, Nir Y, Kipervasser S, Andelman F, Neufeld MY, Kramer U, Fried I et al. (2009) Neural ignition: enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron* 64:562–574.

Gaillard R, Dehaene S, Adam C, Clémenceau S, Hasboun D, Baulac M, Cohen L, Naccache L (2009) Converging intracranial markers of conscious access. *PLoS biology* 7:e1000061.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.

Grill-Spector K, Malach R (2004) The human visual cortex. *Annu. Rev. Neurosci.* 27:649–677.

Güçlü U, van Gerven MA (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience* 35:10005–10014.

Herrmann CS, Munk MH, Engel AK (2004) Cognitive functions of gamma-band activity: memory match and utilization. *Trends in cognitive sciences* 8:347–355.

Hipp JF, Engel AK, Siegel M (2011) Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron* 69:387–396.

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* .

Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol* 10:e1003915.

Kriegeskorte N (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science* 1:417–446.

Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* 2:4.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks In *Advances in neural information processing systems*, pp. 1097–1105.

Lachaux JP, George N, Tallon-Baudry C, Martinerie J, Hugueville L, Minotti L, Kahane P, Renault B (2005) The many faces of the gamma band response to complex visual stimuli. *Neuroimage* 25:491–501.

Lachaux JP, Rodriguez E, Martinerie J, Varela FJ et al. (1999) Measuring phase synchrony in brain signals. *Human brain mapping* 8:194–208.

Lerner Y, Hendler T, Ben-Bashat D, Harel M, Malach R (2001) A hierarchical axis of object processing stages in the human visual cortex. *Cerebral Cortex* 11:287–297.

Levy J, Vidal JR, Fries P, Démonet JF, Goldstein A (2015) Selective neural synchrony suppression as a forward gatekeeper to piecemeal conscious perception. *Cerebral Cortex* 26:3010–3022.

Manning JR, Jacobs J, Fried I, Kahana MJ (2009) Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience* 29:13613–13620.

Michalareas G, Vezoli J, Van Pelt S, Schoffelen JM, Kennedy H, Fries P (2016) Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron* 89:384–397.

Ray S, Maunsell JH (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol* 9:e1000610.

Rorden C (2007) Mricron [computer software].

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115:211–252.

Seeliger K, Fritsche M, Güçlü U, Schoenmakers S, Schoffelen JM, Bosch S, van Gerven M (2017) Cnn-based encoding and decoding of visual object recognition in space and time. *bioRxiv* p. 118091.

Seibert D, Yamins DL, Ardila D, Hong H, DiCarlo JJ, Gardner JL (2016) A performance-optimized model of neural responses across the ventral visual stream. *bioRxiv* p. 036475.

Singer W (1999) Neuronal synchrony: a versatile code for the definition of relations? *Neuron* 24:49–65.

Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience* 18:555–586.

Srinivasan R, Russell DP, Edelman GM, Tononi G (1999) Increased synchronization of neuromagnetic responses during conscious perception. *Journal of Neuroscience* 19:5435–5448.

Talairach J, Tournoux P (1993) *Referentially oriented cerebral MRI anatomy: an atlas of stereotaxic anatomical correlations for gray and white matter* Thieme.

Tallon-Baudry C, Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. *Trends in cognitive sciences* 3:151–162.

Tallon-Baudry C, Bertrand O, Delpuech C, Pernier J (1997) Oscillatory γ-band (30–70 hz) activity induced by a visual search task in humans. *Journal of Neuroscience* 17:722–734.

Van Kerkoerle T, Self MW, Dagnino B, Gariel-Mathis MA, Poort J, Van Der Togt C, Roelfsema PR (2014) Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences* 111:14332–14341.

Vidal JR, Chaumon M, O'Regan JK, Tallon-Baudry C (2006) Visual grouping and the focusing of attention induce gamma-band oscillations at different frequencies in human magnetoencephalogram signals. *Journal of Cognitive Neuroscience* 18:1850–1862.

Vidal JR, Ossandón T, Jerbi K, Dalal SS, Minotti L, Ryvlin P, Kahane P, Lachaux JP (2010) Category-specific visual responses: an intracranial study comparing gamma, beta, alpha, and erp response selectivity. *Frontiers in human neuroscience* 4:195.

Wyart V, Tallon-Baudry C (2008) Neural dissociation between visual awareness and spatial attention. *Journal of Neuroscience* 28:2667–2679.

Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* 19:356–365.

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111:8619–8624.