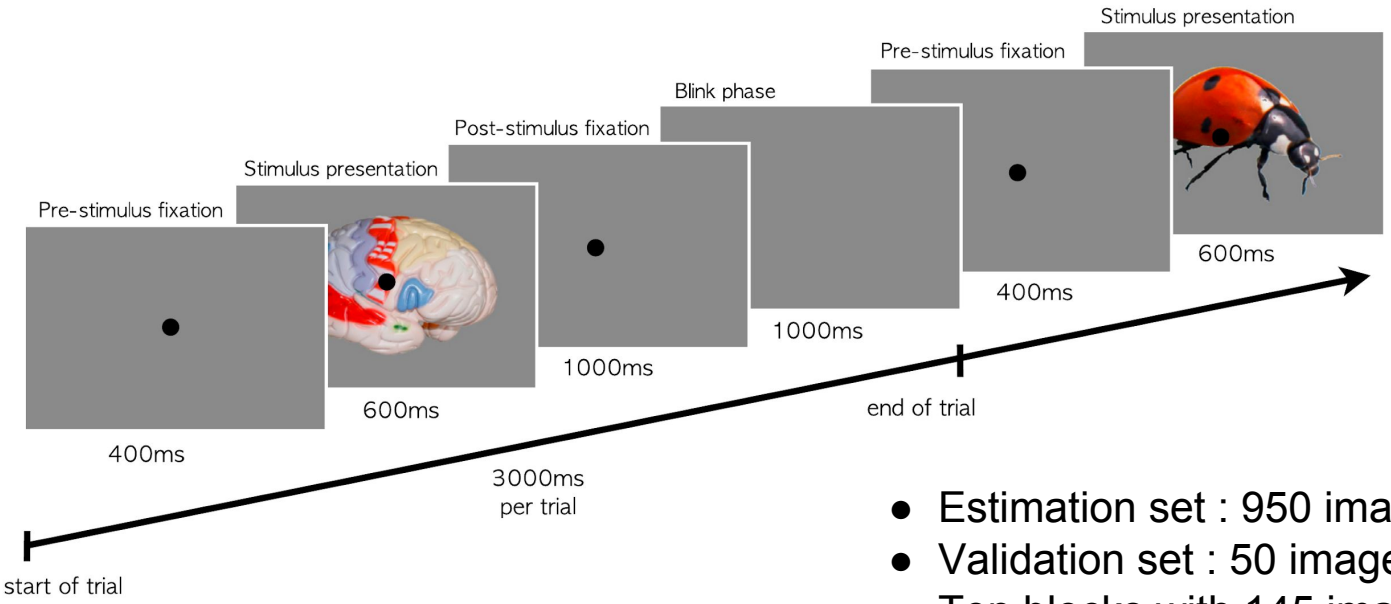
Convolutional neural network-based encoding and decoding of visual object recognition in space and time

K. Seeliger *, M. Fritsche, U. Güçlü, S. Schoenmakers, J.-M. Schoffelen, S.E. Bosch, M.A.J. van Gerven

NeuroImage 2018

Participants and CNN were presented 1450 images

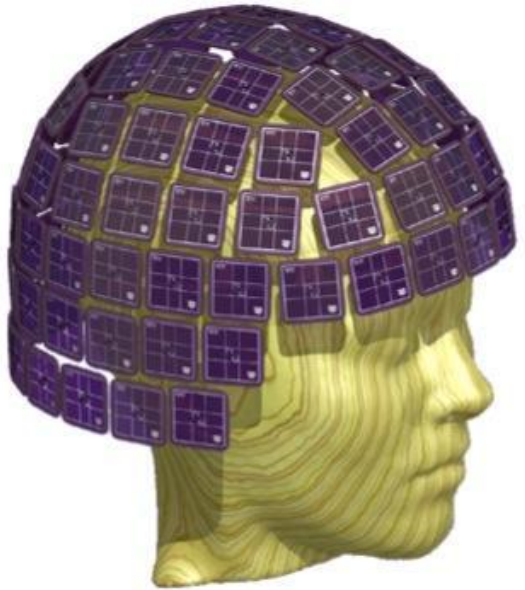


- Estimation set : 950 images
- Validation set : 50 images; repeat 10x
- Ten blocks with 145 images
- Talked to participants between blocks

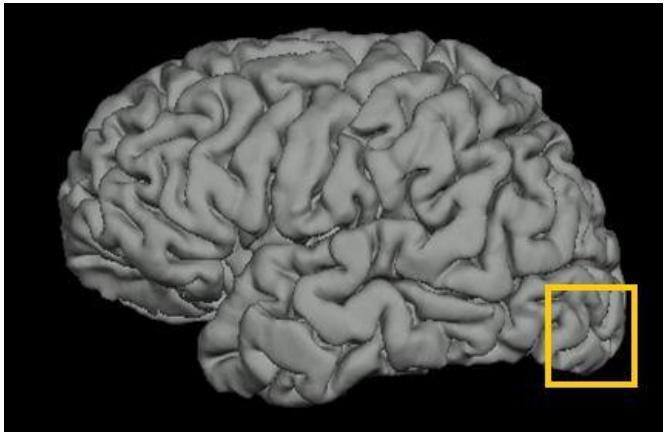
Unusual instructions/methods (to me Ricardo)

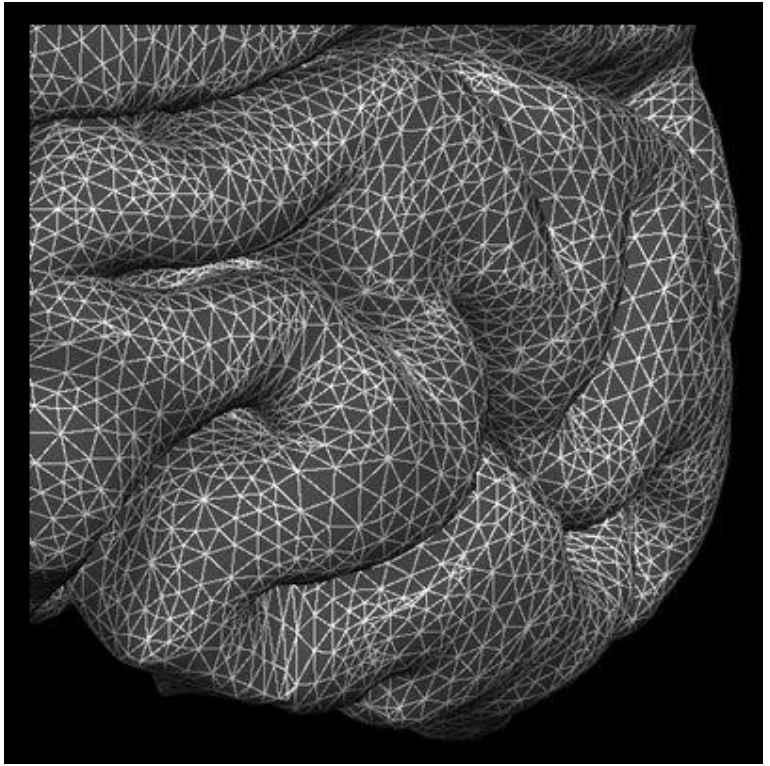
1. (2.2.1) Rejected trials based on preprocessing then average “up to 10 trials”
    a. Was there a minimum number of trials?
2. (2.3.3) Asked participants to “avoid scheduling MRI recordings during the previous few days”
    a. Does MRI affect MEG?
3. (2.3.3) They provided “non-magnetic clothes”
    a. No metal on clothes?
4. (2.3.4) Participants were “informed ... passive nature of experiment”
    a. Seems subjective and impossible to monitor
5. (2.3.4) Participants were “asked ... to adjust their current position”
    a. They do not say how to adjust

Recorded brain activity using 275 channel MEG



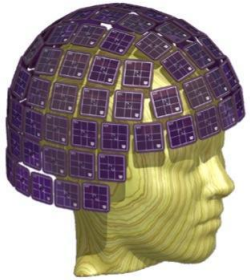
MRI to create constrained source models with mesh



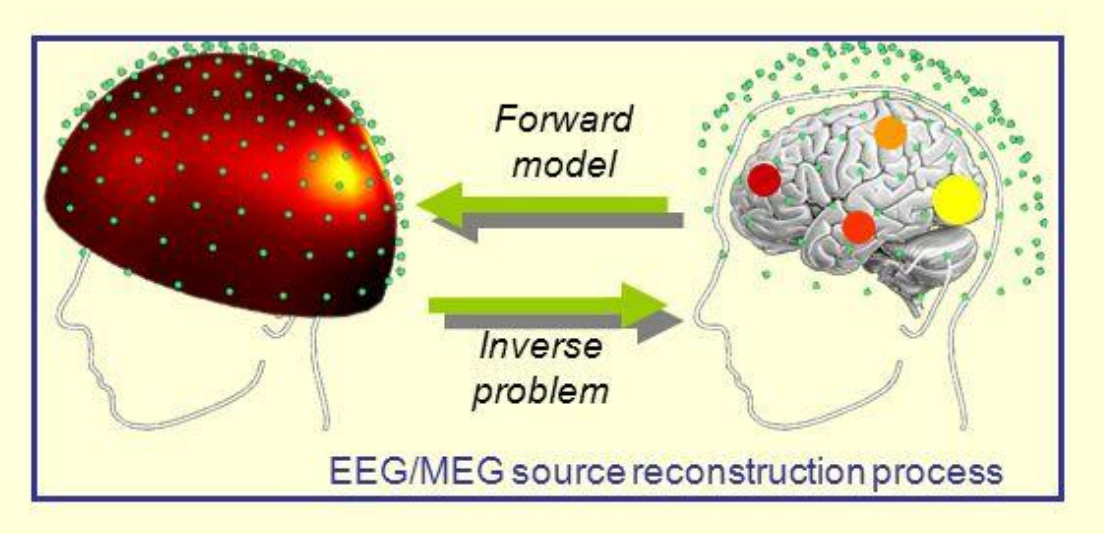


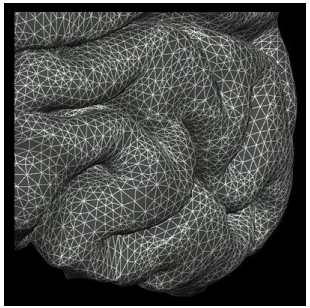
164000 vertices / hemisphere
Downsampled to 4002 vertices / hemi
Removed midline points
7344 dipole sources across entire brain

Estimated source activity with LCMV



sensors = 275



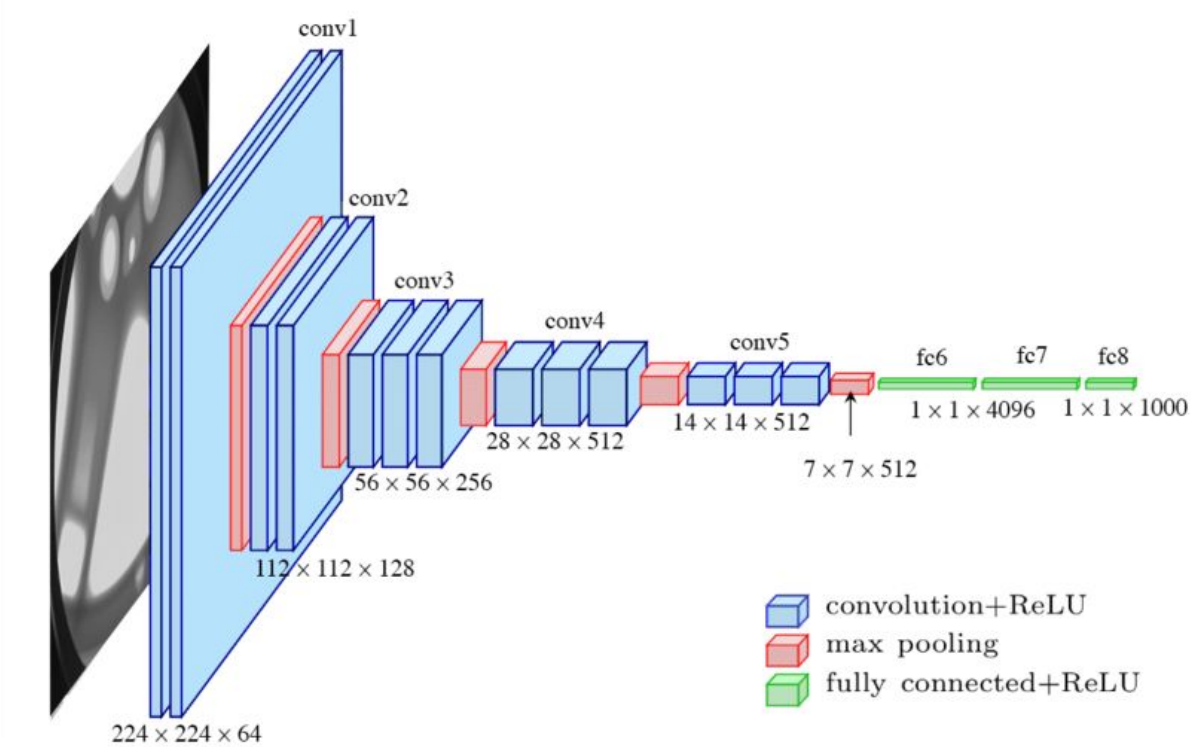


source = 7344

Each image n:
$y_{s,t}$

source response
for time *t*,
for source *s*

Passed each image through the VGG-S (CNN)



#Layers L = 8
Features were size M x 1

Each image n:
$\boldsymbol{x}_{n,L}$

feature representation

Relating VGG-S features to MEG signals

Relevant constants:

- sources **S = 7344**
- timepoints **T = 53**
- layers **L = 8**
- features size **M** (dependent on layer)
- images **N = 1450**

Each image n generated:

- $\boldsymbol{x}_{n,L}$ ***(feature representation)***
- $\boldsymbol{y}_{s,t}$ ***(source response)***

Defined $X_L = (\boldsymbol{x}_{1,L},...\boldsymbol{x}_{N,L})^T$ of size $N \times M$

$$y_{s,t} = \boldsymbol{\beta}_{s,t,L} X_L$$

$$y_{s,t} X_L^T = \boldsymbol{\beta}_{s,t,L} X_L X_L^T$$

$$y_{s,t} X_L^T (X_L X_L^T)^{-1} = \boldsymbol{\beta}_{s,t,L} X_L X_L^T (X_L X_L^T)^{-1}$$

…

$$\boldsymbol{\beta}_{s,t,L} = X_L^T (X_L X_L^T + \lambda I)^{-1} y_{s,t}$$

$\lambda > 0$ ridge regression

"... each of the source **7344** has **53x8 = 424** regression models…"

Results - What we will focus on

1. Mean encoding performance - Fig 2
2. Source-wise layer assignments - Fig 4
3. Prediction-activity correlations over time - Fig 3
4. Temporal onset of layers in different anatomical regions - Fig 5 & 7
5. Spatial distribution of predictive power of layers - Fig 8
6. Stimuli decoding - Fig 10 & 11

Mean encoding performance - Fig 2

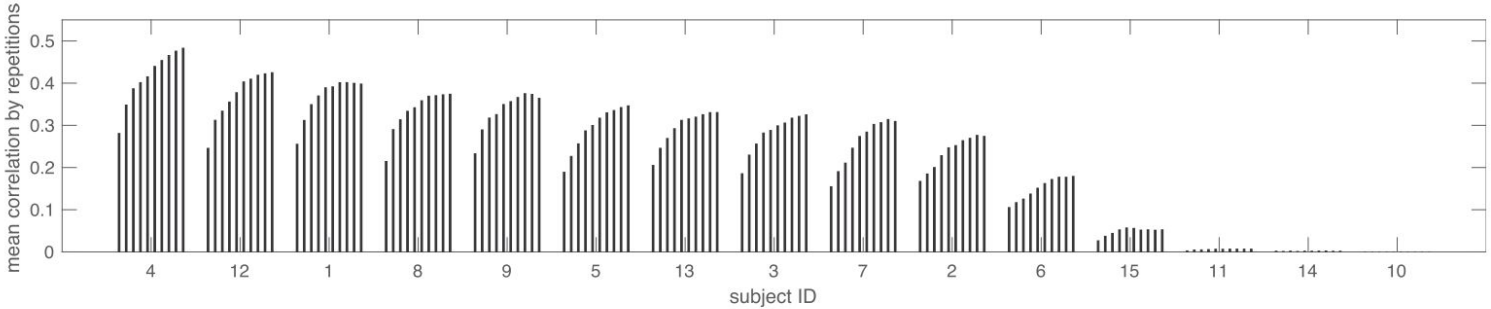


Fig. 2. Mean encoding performance in relation to the number of repetitions on the validation set for sources anatomically assigned to the visual cortex. The model shows considerable mean correlations for 10 out of 15 participants. For 3 participants it is not predictive at all, and for 2 average correlations are low. We predicted the activity for each source for the time bins between 75 ms and 600 ms after image onset for the 50 images in the validation set (before the 75–105 ms time bin, for most subjects no source activity can be predicted). Predictive models were trained for each source-time bin combination on the full estimation set; with significance and optimal layers estimated during cross-validation within the estimation set. Correlations between the predicted and the measured responses per source and time bin were then taken across validation set stimuli. The mean shown here summarizes these correlations for sources assigned to the visual system areas with our anatomical parcellation. The increased SNR from averaging over repetitions improves encoding performance. However for most participants the average over 10 repetitions appears to be close to a performance plateau.

Source-wise layer assignment - Fig 4

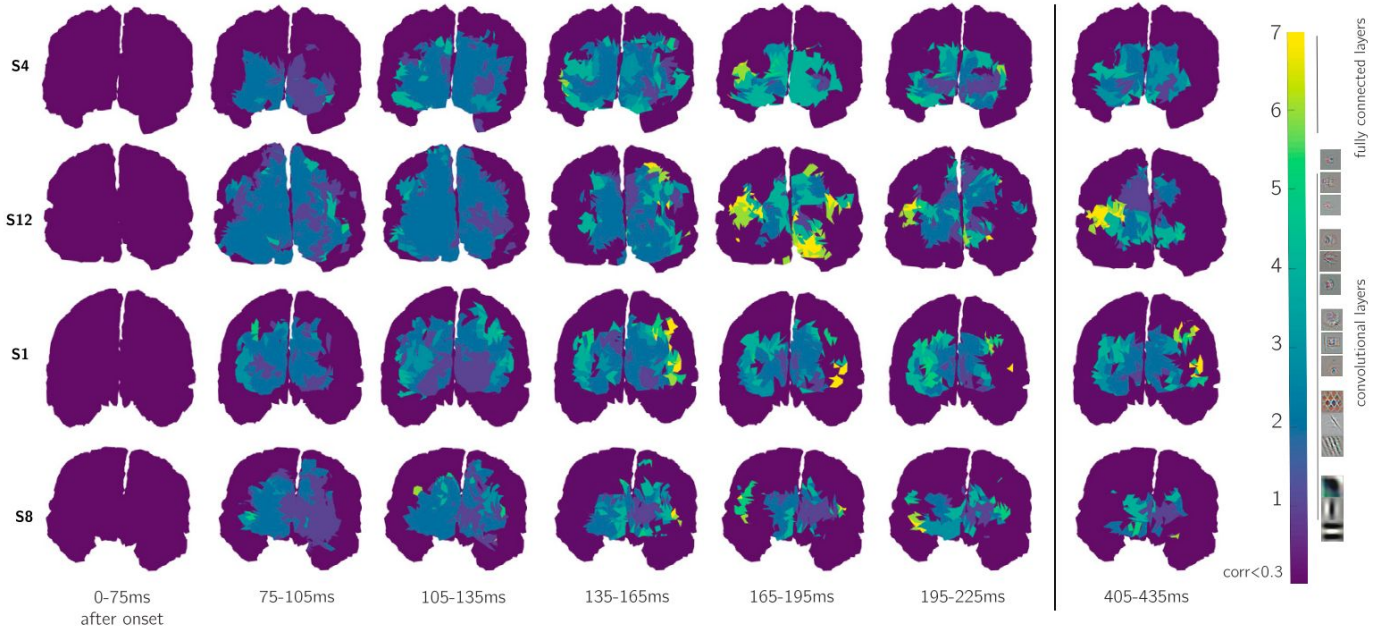


Fig. 4. Source-wise layer assignments over time for four participants. Views are coronal from the posterior onto the occipital lobe. Each source-time bin combination gets assigned the representation layer that explained it best during the nested cross-validation on the estimation set, measured in the average correlation across all folds. Sources are shown on individual brains based on the Freesurfer models. Non-significant sources and significant ones with low correlations (<0.3) are discarded in these maps. The manifestation of the fully-connected layers 6 and 7 first occurs after 135 ms for most participants. Before this time convolutional layers are expressed, starting with a widespread manifestation of layer 1 and 2 in the early visual cortex region. After the expression of fully connected layers for some, but not all participants we see sustained activity, here shown for the time bin 405–435ms. The colormap was chosen to reflect the division between convolutional and fully-connected layers.

Comparison to Cichy et al 2016 study (MEG)

- Used Representational Similarity Analysis (RSA) for stimuli set
- Correlation between RSA matrix of CNN layer and MEG source

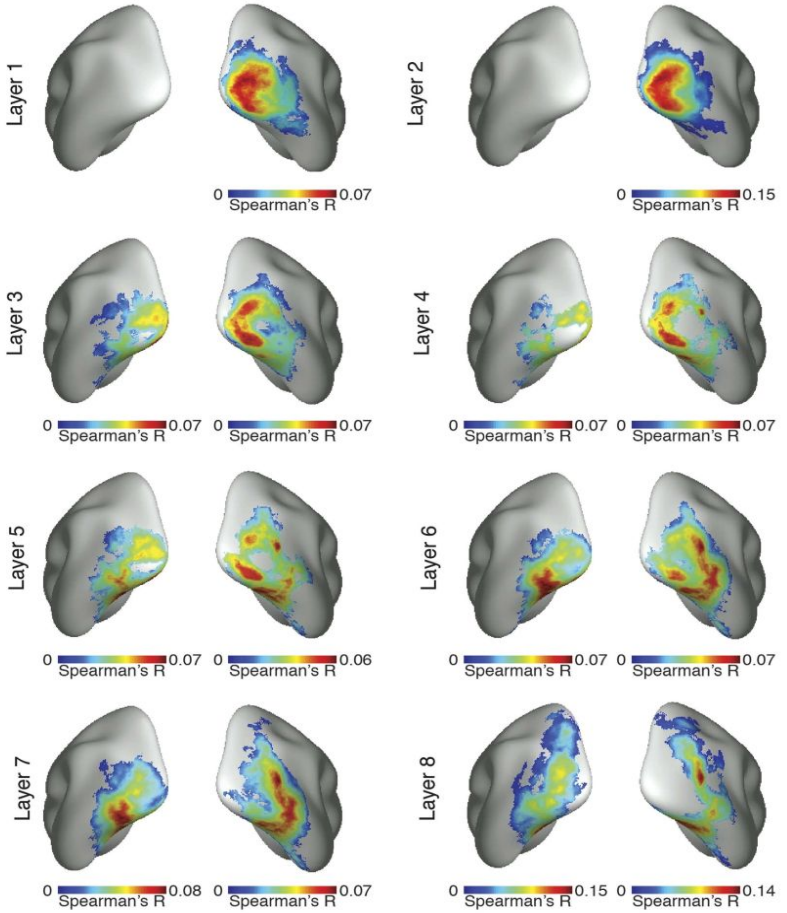


Figure 4. Spatial maps of visual representations common to brain and object DNN. There was a correspondence between object DNN hierarchy and the hierarchical topography of visual representations in the human brain. Low layers had significant representational similarities confined to the occipital lobe of the brain, i.e. low- and mid-level visual regions. Higher layers had significant representational similarities with more anterior regions in the temporal and parietal lobe, with layers 7 and 8 reaching far into the inferior temporal cortex and inferior parietal cortex ($n = 15$, cluster definition threshold $P < 0.05$, cluster-threshold $P < 0.05$ Bonferroni-corrected for multiple comparisons by 16 (8 DNN layers * 2 hemispheres).

Prediction-activity correlations over time - Fig 3

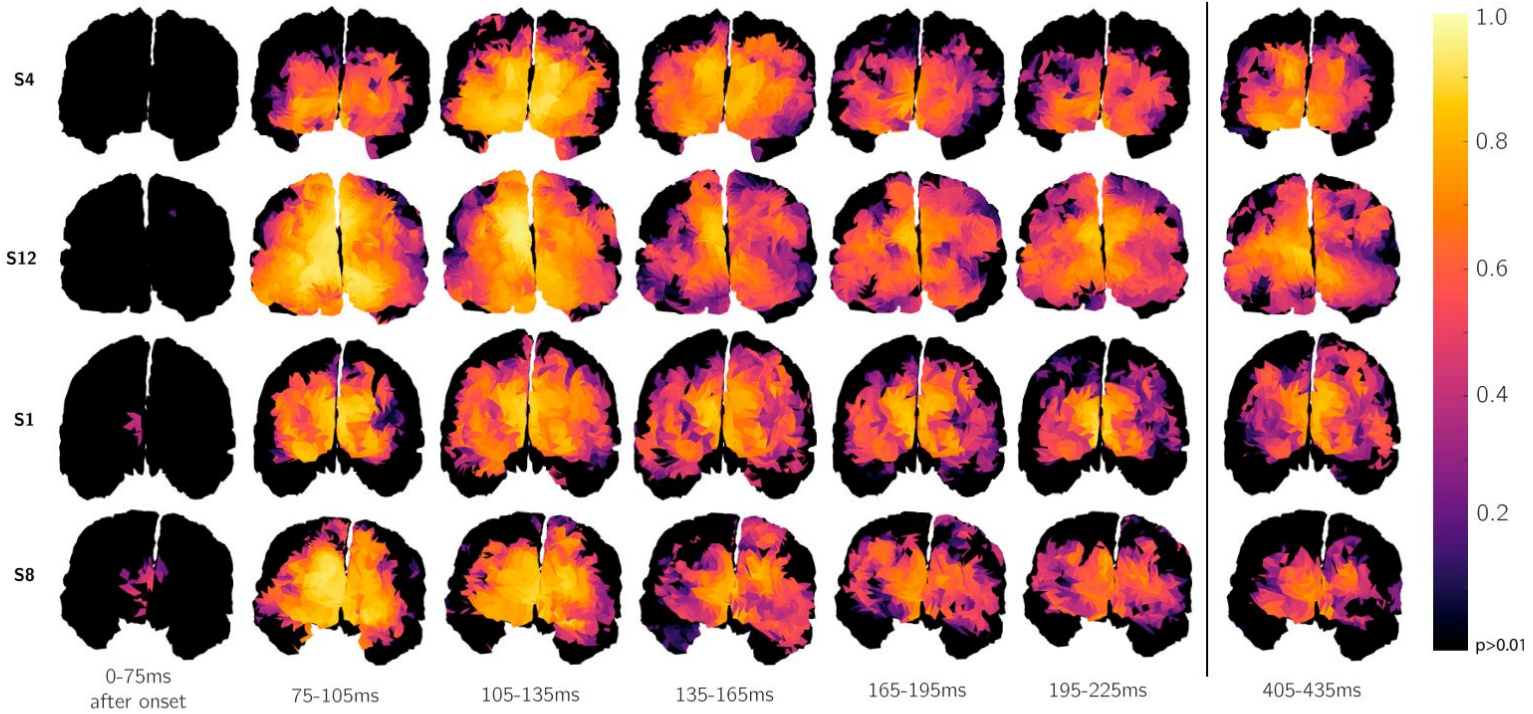


Fig. 3. Prediction-activity correlations over time on the 10-times validation set for the encoding models using the best-explaining layer as in Fig. 4. Views are centered at the occipital lobe. We show all above-chance correlations. We observe the highest correlations around the early visual cortex, and lower correlations in extrastriate areas.

Temporal onset of layers in different anatomical regions - Fig 5

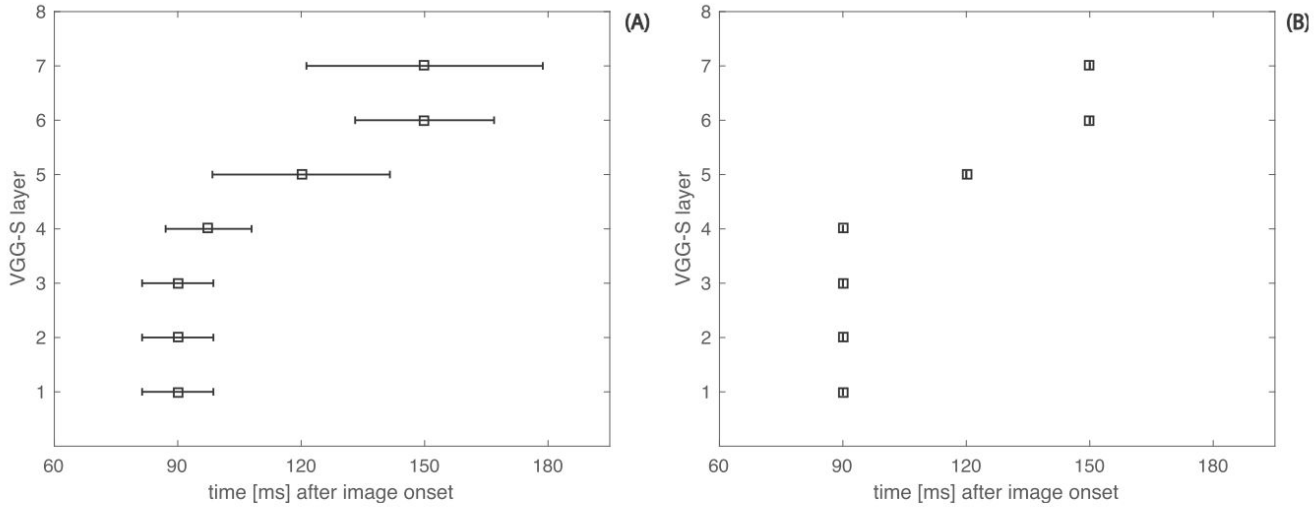
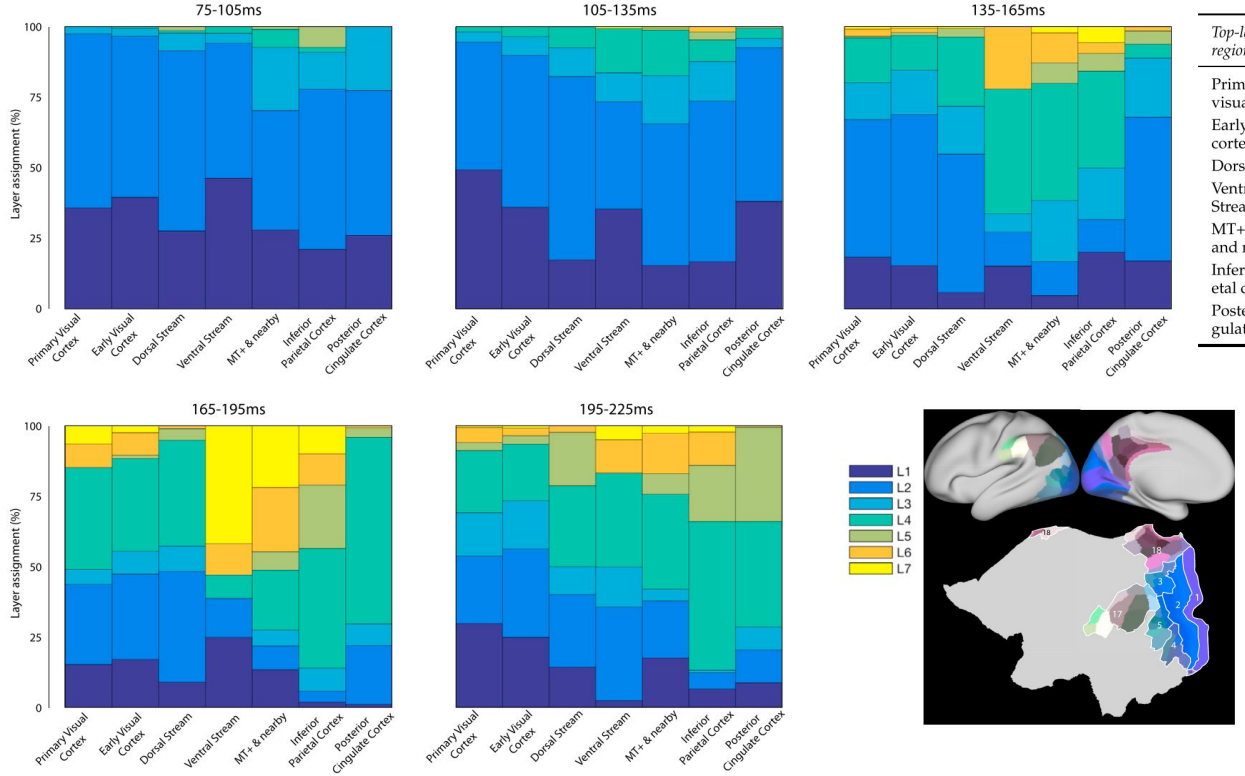


Fig. 5. Temporal onset of layers, average (A) and median (B) over the 12 participants for which the encoding model was predictive. Shown is the first time bin in which the respective layer explains at least one source best with correlations above 0.3. We indeed observe that the hierarchy from Layer 1 to 7 is being expressed sequentially within 200 ms after stimulus onset. The first four convolutional layers first occur within the 75–105 ms time slice. Taking subject-wise Spearman's ρ correlations between layer numbers and time-binned onset times leads to an average rank correlation of 0.854 for the 10 subjects that expressed all 7 layers and the two subject that missed layer 5 or layer 6 and 7 respectively ($p < 0.001$ Bonferroni-corrected, combined over subjects with Fisher's method).

Temporal onset of layers in different anatomical regions - Fig 7

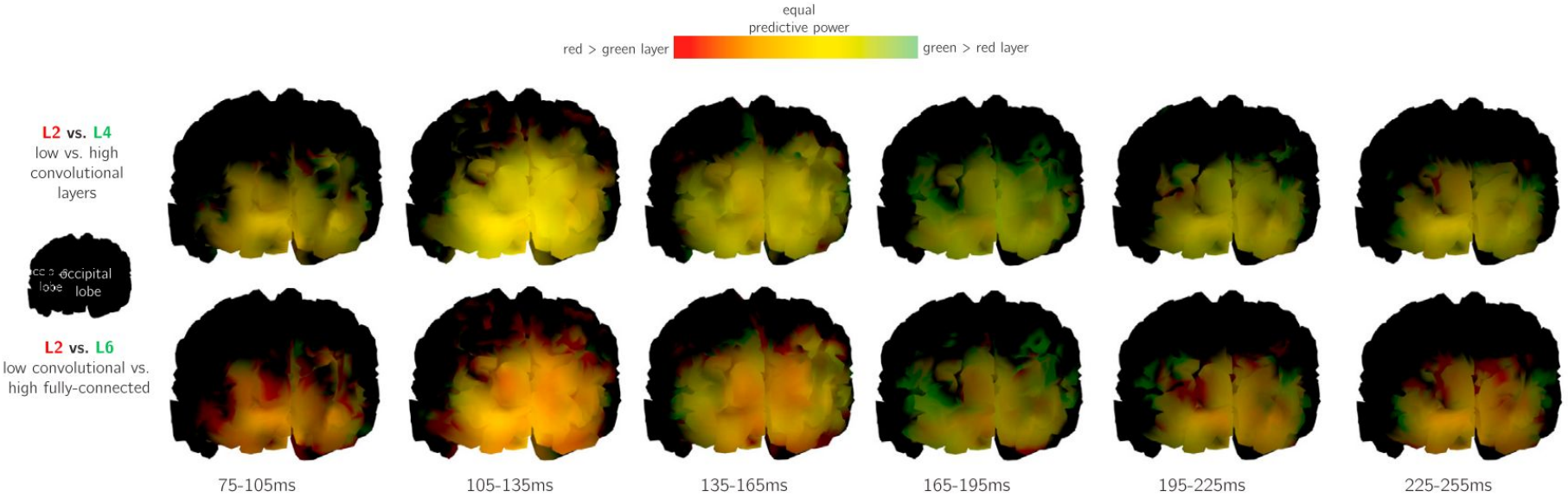


Top-level region	*Anatomical labels from (Glasser et al., 2016)*	*Index*
Primary visual cortex	V1	1
Early visual cortex	V2, V3, V4	2
Dorsal Stream	V3A, V3B, V6, V6A, V7, IPS1	3
Ventral Stream	V8, VVC, PIT, FFC, VMV1, VMV2, VMV3	4
MT+ complex and nearby	V3CD, LO1, LO2, LO3, V4t, FST, MT, MST, PH	5
Inferior parietal cortex	PGp, PGs, PGi, PFm, PF, PFt, PFop, IP0, IP1, IP2	17
Posterior cingulate cortex	DVT, ProS, POS1, POS2, RSC, v23ab, d23ab, 31pv, 31pd, 31a, 23d, 23c, PCV, 7m	18

- The conv feature detectors in Layer 5 already express complex templates of full objects.
- Incorporating the special role of Layer 5, there is a temporal division between lower and mid convolutional layers and the most abstract layers.

Fig. 7. Number of sources in an anatomical region assigned to the network layers, averaged over participants 4, 12, 1, 8. Early convolutional layers up to Layer 4 are expressed across all layers. The most abstract Layers 6 and 7 appear in the ventral stream and neighbouring regions after 135 ms (sources with above 0.3 correlations).

Spatial distribution of predictive power of layers - Fig 8



(a) **Participant 4.**

Fig. 8. Spatial distribution of predictive power of convolutional and fully connected layers over the first 255 ms. Sources explained by early convolutional layers and fully connected layers do not appear in the same regions. Convolutional layers explain similar regions, with the mid-level convolutional layers spreading out into extrastriate areas. For the visualization, correlations for each layer are normalized by the highest correlation observed for each participant and Fisher-z corrected to allow linear comparability. Correlation values for the given layers then fill either the red or the green RGB color channel, highlighting sources were one layer outperforms the other, and leading to a mixture (yellow) if both layers can explain a source equally well.

Stimuli decoding - Fig 10 & 11

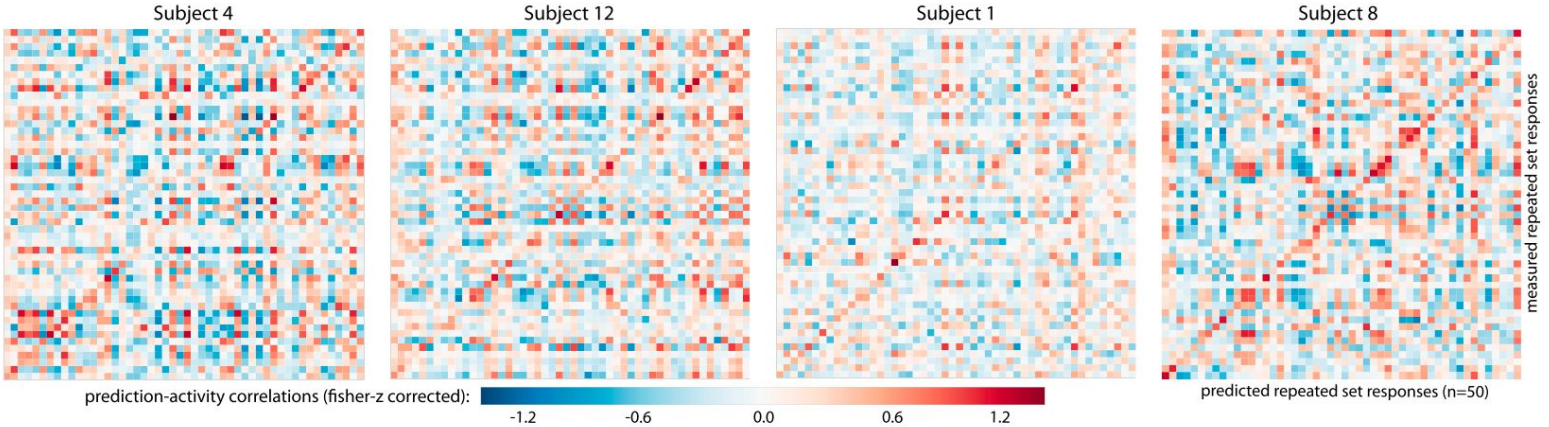


Fig. 10. Correlations (on resampled data) between predicted and measured averaged activity for four participants for the 50 images of the validation set, between 75 ms and 225 ms. Each row represents the measured source response patterns in this time slice for these 50 images, and each column the predicted source response patterns respectively. The matrix color-codes the pairwise correlations between all predicted and measured response patterns. A prominent anti-diagonal thus indicates that predicted and measured response patterns correspond when comparing for the same image, and differ when comparing to all others. Sources that reached higher average correlations than 0.3 during nested cross-validation on the estimation set were selected for decoding. Predicted source responses were compared to measured responses resampled over ten repetitions