École Polytechnique Fédérale de Lausanne
School of Life Sciences

Master Project
in
Life Sciences Engineering

Combining neurophysiology and computational modeling through VGG19

Done by:
Leonardo Pollina

Carried out in the Kreiman lab
at Harvard Medical School
Under the supervision of Prof. Gabriel Kreiman

Under the direction of Prof. Silvestro Micera
Translational Neural Engineering Laboratory
EPFL

Boston, MA, USA, March 2021

Abstract

During the last years neurophysiological findings and computational modeling have benefited from one another: new discoveries about the brain's mechanisms inspired some powerful models and new advances about these same models allowed to gather more insights on how neural circuits might work. In particular, the sense of vision has been largely studied in both domains. Crucial regions for object recognition in the brain have been identified and deep convolutional neural networks have achieved amazing performance in classifying images in computer vision tasks. Current computational models of visual processing focus on the bottom-up cascade of sequential operations from the retina onto primary visual cortex and higher-level processing. However, feed-back in the brain is abundant and its role is crucial, especially for goal-directed behaviors. Here, we investigated the extent to which deep convolutional neural networks can account for the variance of intracranial neural data recorded in subjects while presented with images. Moreover, we sought to understand how task goals modulate the responses along the ventral visual pathway through the development of computational models that incorporate top-down feed-back signaling carrying task-specific information. We show here that feature maps extracted from VGG19, a feed-forward deep convolutional neural network, are capable of predicting neural data with satisfying results in two different human data-sets. Additionally, we suggest a simple yet effective way to incorporate top-down modulation propagating through subsequent layers in VGG19. The top-down modulation is proven effective both for target-modulation and for task-modulation, the latter being here characterized in a new unpublished data-set. The versatility of our approach shows the potential of this implementation method to be applied to a wider range of top-down signals, hence making artificial neural networks more similar to biological neural circuits.

Acknowledgements

This work is the result of an amazing year spent doing research in the Kreiman lab at Harvard Medical School in Boston. It is also the end of an incredible journey that lasted more than 5 years. I could not be more grateful for everything I have had.

I would like to thank Prof. Gabriel Kreiman for having given me the chance to conduct this research in his lab and for having guided me through a fascinating field. Thank you also for our weekly meetings and for having supervised me thoroughly while also letting me take my independent decisions. Katarina, thank you for having joined this project, for your guidance and for having enthusiastically engaged in a lot of scientific (and not) conversations with me. I probably owe you half of my mental health during this year of working from home. Thanks to the Bertarelli Foundation and to all its members for the huge opportunity offered me and for the continued supervision.

A huge thank you also to Prof. Silvestro Micera, who supervised my project from the other side of the ocean and who has been a mentor for me during these last few years.

I surely owe a lot to the amazing people, now friends, that I have had the chance to meet here in Boston. Thank you Alice for having been my ally against all these sporty friends and for having endured, without never complaining, all the technical discussions about this project. Thanks Charlotte for always having brought a ray of sunshine in the group. Your energy, your joy and your will to discover are inspiring. Thanks to Hugo for having been my working-from-home-mate, for all our random conversations and for having shown me that we are still a bit crazy even when we are 24. And a special thanks goes to Camille, who accompanied me until the very end. We went through a lot of emotional ups and downs together and we experienced every single season that this year has offered us. People came and went, but we always stayed. Unfortunately, I cannot name everyone here to avoid writing a book, but I really thank you all from my heart.

I also want to thank my friends back in Switzerland: Ali, Gaia, Giovanna and Hugues among others. 2020 has seen us apart, but I am sure that plenty of new experiences are awaiting for us. Thank you for having always reminded me that Europe was just to the other side of the phone.

Finally, I must conclude by thanking my parents. Thank you for having taught me and still teaching me everything you know and for being my home in this world sometimes too big.

Leonardo

Contents

1 Introduction

1.1 Object Recognition in Neurophysiology

Vision has for a long time been considered as the dominant sensory modality [1, 2]. However, this has often been associated to uni- or bisensory experiments and it has been argued that the dominance effect is lost when multi-modality (in particular visual, audio and haptic) stimuli are used [3, 4]. Nonetheless, the importance of vision as a sensory modality is without doubt emphasized by its role in one of the most important tasks that the human visual system solves everyday and at any time: object recognition. This is defined as our amazing ability to quickly identify and classify a multitude of different objects hidden in natural and sometimes even chaotic scenes [5].

The human visual system is commonly described as a sequence of different structures and pathways. The first important structure is the retina, which is the thin photosensitive tissue allowing us to detect light. The information received from the retina is then sent via the optic nerve and the optic tract to the Lateral Geniculate Nucleus (LGN) of the thalamus, a center where visual information begins to be processed. From the LGN, the signals travel through the optic radiation and arrive to the primary visual area, also known as V1 and located in the occipital cortex. V1 constitutes the first area of substantial processing of visual information and of feature extraction. In fact, in this region, populations of neurons have been shown to be specialized in the detection of light's orientation and direction [6, 7, 8]. From V1, the visual system is then commonly described as separating into two main streams, a dorsal and a ventral stream. These two streams are also often referred to as the *Where* and the *What* pathways [9], because of their role in the detection of motion and visual details, respectively. For example, the dorsal pathway processes high temporal frequencies and the signals it receives come from all across the retina, while the ventral pathway processes high spatial frequencies and it involves signals coming mainly from the fovea, the point of highest visual acuity in the human retina [10].

It should then not come as a surprise that there is a common consensus about the significant involvement of the visual ventral pathway in the task of object recognition [11]. In particular, the inferior temporal (IT) cortex has been shown to be crucial for object recognition in primates [11, 12, 13]. The exact homology for such an area is still being investigated in humans: likely candidates are the regions inside and around the lateral occipital cortex [14, 15]. Moreover, the importance of the temporal lobe in humans and its role in object and face recognition has also been shown by some studies based on the temporary deactivation of neural circuits through transcranial magnetic stimulation(TMS). [16, 17].

The mechanism and circuitry underlying object recognition seem to rely upon a highly hierarchical and feed-forward organization of the several regions located along the visual ventral pathway. In particular, the ventral pathway is often divided in the following areas: V1, Visual area 2 (V2), Visual area 4 (V4), posterior, central and anterior IT cortices (pIT, cIT, aIT) [11, 18]. These areas are believed to play different roles in visual signal processing and feature extraction, going from more basic features in earlier areas (such as V1, V2) to more complex representations in the IT. This aspect is also supported by the increasing size of the neurons' receptive fields (RF) from V1 to IT, with neurons in the latter having large RF as compared with RFs in V1 [19].

We already mentioned how V1 was proven to be linked to the extraction of basic features such as orientations and directions, which can be used for the detection of edges. Some studies with non-human primates have indicated that V2 uses the information gathered from V1 to analyze and identify more complex features, such as contours and textures [20, 21, 22]. Although it also presents selectivity to other visual features such as orientation [23], V4 is largely associated with color processing [24]. In addition, V4 has shown to be crucial for color constancy, which gives the capability of perceiving a color constant despite different levels of illumination and brightness [25, 26]. Finally, the most complex features and the proper object recognition task seem to be solved in the highest area of the ventral

pathways, that is the IT cortex. Here, neural populations have been proven to be often category-selective, with categories such as fruits, bodies, faces and places [5, 27].

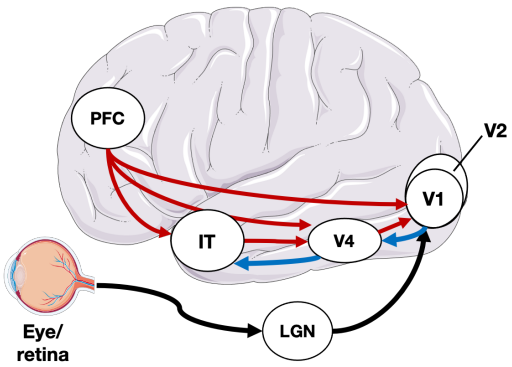


Figure 1: An illustration of part of the visual pathways. The pathway indicated by blue arrows represents the ventral pathway, also called the *What* pathway for its crucial role in object recognition. The red arrows indicate the general idea of top-down modulation originating from the frontal lobe.

As it appears from the aforementioned sequence of processing in consecutive brain regions, part of the visual pathway is a mainly *feed-forward* bottom-up process, where information flows from a lower area to a higher one in the hierarchy. This is also the most commonly used structure. Bottom-up here refers to the processing of an external stimulus (a natural scene) by the different brain areas until identification of the main objects in the scene is achieved. Nonetheless, although more difficult to understand and to study, recurrent and top-down *feed-back* connections are also important and abundant in the visual pathways. One form of feed-back can be represented by the so-called horizontal connections, which constitutes intra-region connections. This type of connectivity is thought to be involved in the creation of highly specialized sets of neurons by linking together neurons with similar feature preferences. An example is the finding of orientation selective neurons in V1 [28, 29]. In addition, top-down connections are believed to be essential for the integration of higher cognitive functions, such as attentional modulation and working memory among others. Often, these top-down connections are thought to origin from the frontal lobe [30, 31]. Recurrent and/or top-down processing has also been suggested to play an important role in the case of object recognition including occluded objects [32, 33].

Another example having been associated with top-down modulation is the task where a subject has to recognize a certain target object (or category of objects) when presented with multiple images showing objects belonging to several categories. In fact, when deciding whether an object is or is not part of the category we were told to consider as a target, we have first to recognize the object and then to compare it to our goal in order to solve the task. In their work, Bansal and colleagues, show that neural signals present no difference between target and non-target trials for around the first 250 ms after stimulus onset. The target and non-target signals start to diverge only after that time, probably indicating the shift from an object-recognition phase to a target-modulation phase, with the latter involving top-down connections likely coming from the prefrontal-cortex [34]. A simple overview of some of the structures involved in the visual pathways is outlined in Fig.1.

1.2 Object Recognition in Deep Learning

In recent years, there has been an increasing interest in the domain of pattern recognition, machine and deep learning and artificial intelligence. In particular, in the field of computer vision, Deep Convolutional Neural Networks (DCNNs) have proven to be extremely successful [35].

Computer vision is generally considered as belonging to the field of computer science and artificial intelligence, since it is concerned with dealing with computational models performing tasks over im-

ages. The most common tasks are: object detection, semantic segmentation and object recognition [36]. However, it is important to note that in all cases we are talking about *supervised learning*, which means that a model is trained in an iterative way by comparing the model's output with the ground truth value. All the mentioned tasks have different purposes and use different metrics to assess the performance. Object detection consists of building a model capable of finding the object(s) present in the scene. This is done by giving as output a location box around the object. One of the metrics most commonly used in this case is the *Intersection over Union*, which is the intersection between the area of the predicted box and the ground truth one over the union between these two same areas. Image segmentation corresponds to the classification of individual pixels. An example of such a task is road segmentation, where satellite images are segmented in street pixels vs background [37]. The performance is often assessed by computing the *segmentation accuracy*, that is the pixel-wise accuracy. Computationally speaking, object recognition represents the task in which a model is supposed to predict the category of the object in an image. A simple metric such as the classic *accuracy* can be used here to validate the model.

DCNNs are computational models that somehow take inspiration from some of the features of the human visual system [7]. These models are composed of three main types of layers: convolutional layers, pooling layers and fully-connected layers. Convolutional layers consist of a certain number of kernels (also called filters) which exhaustively scan the image via a convolution operation that takes several pixels into consideration at once. These layers are supposed to mimic the concept of the receptive fields found in simple and complex cells along the visual pathways. In fact, the area of the input image covered by one filter could be thought as being a sort of computational receptive field that extract useful information in the input image through a "context" analysis rather than a pixel-wise one. This also allows the model to extract local basic features (such as edges or corners) which are then assembled into more complex features in subsequent layers, hence mimicking once again the hierarchy found in the visual pathways going from the LGN to IT in the ventral stream [18]. Moreover, it is worth mentioning that in such layers a certain filter will have the same weights while scanning the input image. This means that this filter will specialize in the extraction of a specific feature, which is the reason why several filters are used. The output of every convolutional filter is called a feature map[35, 38, 39].

Pooling layers are special layers that are commonly inserted after a certain number of consecutive convolutional layers. These layers, among which the most common ones are max pooling and average pooling, perform a certain operation, such as taking the maximum or computing the average over a certain region of their input image. This operation is usually done through a sliding pooling filter. Again, this filter can theoretically be associated to a neuron's receptive field. However, the interesting feature of pooling layers is the generation of spatial invariance regarding the output. Spatial invariance means that a slight change in the object position inside the input image will not generate any difference output-wise. Finally, fully-connected layers often constitute the final part of a model, where the features extracted are combined together in order to obtain a final output. Because of this peculiar structure within a Convolutional Neural Network model, we often refer to the part containing the convolutional and pooling layers as the *feature-extraction* part and to the final part containing the fully-connected layers as the *classifier* one [35, 39].

Another fundamental aspect to outline when talking about DCNNs and more generally about Artificial Neural Networks is the importance of the presence of non-linearity inside the model. In fact, Deep Learning models have been so successful and have outperformed more shallow and linear classifiers thanks to this particularity. A non-linear model will be able to partition the feature space more freely with respect to a linear one. A non-linear function is applied inside each layer after the product between the weights and the input is performed. This function is usually called *activation function*. For instance, the operation performed by a fully-connected layer can be summarized by Eq.1:

$$\mathbf{x}^l = \Phi(\mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^l) \tag{1}$$

where $\mathbf{x}$ represents the vector of activations of a certain layer, l is the superscript indicating the layer, $\mathbf{W}$ is the weights matrix, $\mathbf{b}$ is the bias vector and $\Phi()$ is the activation function. In DCNNs the most common activation function is $ReLU(\mathbf{x}) = max(\mathbf{x}, 0))$ [39]. This function has been seen as a method to preserve and to propagate some information, while discarding some other. Often, when it comes to the last layer, also called the prediction layer, it is common to use $sigmoid()$ or $softmax()$ as the activation function, since the output of these two functions is in the range [0,1] and can hence be interpreted as the probability of belonging to a certain class rather than to another.

Historically, one of the first models used to perform image classification is *LeNet-5*, which is a Convolutional Neural Network. This model has been successfully used in the classification of hand-written digits (MNIST database) [38]. Modern successors of this initial model are *AlexNet* and the *VGGNet* (such as VGG16 and VGG19 models) family. Both models are known for their amazing performances obtained over the ImageNet database. ImageNet is a large dataset composed of ≈15 million labeled high-resolution color images. This database of images is the object of an annual competition on object recognition models: the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This competition uses a subset of the dataset, limiting the number of different classes to 1000. *AlexNet* was presented as the state-of-the-art and the winner of the ILSVRC-10 and ILSVRC-12 competitions at that time. The model is composed of 5 convolutional layers, 3 max Pooling layers and 3 fully-connected layers, of which the last one coupled with a $softmax()$ activation function to perform the proper classification. *VGGNet* is a family of DCNNs presented in 2014 as the one achieving the best results in the ILSVRC in 2014. VGG19, so named because of the presence of 19 weight (trainable) layers, is formed by 16 convolutional layers and 3 fully-connected ones, with the last one having 1000 units and using the $softmax()$ activation function like in the case of *AlexNet*. The convolutional layers are divided in 5 convolutional blocks, each of them followed by a Max Pooling layer. Both VGG19 and *AlexNet* make use of the $ReLU()$ activation function for intermediate layers [40, 41]. The architecture of VGG19 is outlined in Fig.2.

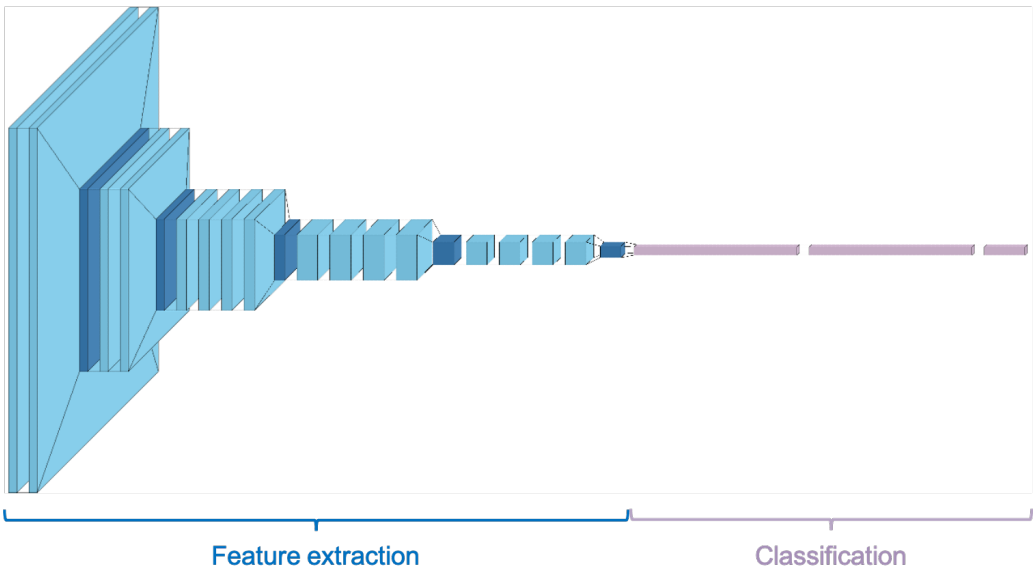


Figure 2: Architecture of VGG19 Convolutional Neural Network. The light blue layers represent the 16 convolutional layers while the dark blue ones are the max pooling layers. The pink blocs correspond to the 3 fully-connected layers.

1.3 Linking Neurophysiology and Computational Modeling

In the previous section we have explained how Deep Convolutional Neural Networks are structured and how this architecture closely mimics the functioning of the ventral visual stream. During the last years, research in neuroscience and computer science have influenced and benefited from one another's progresses: the new findings about the ways our brain works have inspired new computational algorithms and, conversely, these powerful algorithms have helped us to postulate and test new computational theories about how object recognition might function in the human brain. A clear example of how computational models can be leveraged to gather some insights about neural representation in the visual system is given by the work of Ponce and colleagues [42], where a Deep Generative Network was used to shed light on the encoding characteristics of some neural populations inside the IT cortex in monkeys.

A lot of work has been done to try to finally link neurophysiology and computational modeling. In fact, a possible hypothesis is that if two different approaches manage to solve the same tasks (such as object classification), it is then reasonable to think that they have evolved towards developing the same strategies to achieve such results. This is the reason why it is interesting to study the similarity between the features extracted by computational models and the ones computed from neural data.

Amazing results have been obtained regarding the capability to link neural data recorded in monkeys and computational activations extracted via a DCNN. Examples of such analyses are given by the works performed by Yamins and O'Connel and their respective colleagues. Yamins trained a linear regression model associating artificial activation maps given by different layers of a computational model (optimized through hierarchical modular optimization) to neural features. It was found that the last layers correlated extremely well with neural data recorded in the IT cortex, while the model's intermediate layers showed a better correlation with neural data recorded in intermediate regions in the visual hierarchy, such as area V4 [43]. The research group of O'Connel focused on the generalization capabilities of such a linear model. Their results show how the representation learnt between VGG16 and neural features can extrapolate to categories of objects left out from the training of the model. This indicates that the mathematical representation manages to capture important generic visual features [44]. All these results highlight once again the potential similarity between these two systems.

Some interesting results have also been found when using human data. In a work including 61 subjects and a total of 8916 intracranial electrodes, Grossman and colleagues [45] found that 96 sites distributed across 33 subjects were selective to faces. These sites were then used to compare the neural representation of faces to the artificial representation obtained by feeding VGG-Face (a DCNN having the same architecture as VGG16 but trained with a different dataset and achieving human-like performance in face recognition) with the same images. A face-space was built by computing the Euclidian distance between pairs of face images. Significant correlations between the computational and the neural face-spaces were found for the intermediate layers of the model, while chance-level performances were present at the extremes of the architectural hierarchy [45]. On a more behavioral side, the similarity between perceived dissimilarity measurements in 269 humans subjects were compared to a set of computational models in [46]. Among the tested models, the Convolutional Neural Network was found to be the best one based on the amount of variance explained in the human data.

It is important to point out though that all this work often takes into account only the feed-forward components of the visual pathways, hence disregarding the abundant feed-back connections mentioned in Section 1.1. This is because classical DCNNs also present a simple feed-forward sequential architecture. As we already discussed, top-down modulation seems to play a crucial role in visual tasks including attentional components or concerning recognition of occluded objects. Regarding this latter aspect, Tang and colleagues showed how visually selective responses were delayed when the subjects

were presented with images containing occluded objects with respect to when presented with whole object images. This finding might point to the requirement of feed-back recurrent connections when recognizing partially visible objects. Moreover, while state-of-the-art feed-forward convolutional models did not achieve a good classification performance when presented with occluded object images, the performance was restored if recurrent connections were added to the model's architecture [47].

Recently, some efforts have been put in order to implement computational models including such feed-back components. For instance, Zhang and colleagues attempted to create a model mimicking natural visual search. VGG16 was used to model the bottom-up ventral visual stream and to extract the information related to the target object. This information was then stored in a module mimicking the pre-frontal cortex, which in turn top-down modulated a second DCNN finally generating an attention map. Using such modulation it was possible to select the region of space containing the target object, hence succeeding in the visual search task [48]. Another interesting model trying to account for goal-directed behavior and architecturally inspired by brain anatomy is presented by Adeli and Zelinsky. In their Deep-BCN model, they make use of a pre-trained AlexNet to mimic the bottom-up visual processing. They then couple it to a series of blocs representing the pre-frontal cortex, the superior colliculus and the frontal eye field by employing both feed-forward and feed-back connections. The feed-back connections are implemented by using the same connections as the feed-forward signal, but by computing their gradient [49].

The possibility of convergence towards the same mechanisms related to object recognition in the ventral system and in state-of-the-art Deep Convolutional Neural Networks is still extensively studied and investigated. However, it is now clear that object recognition represents but a single aspect of the complexity of vision. In order to be able to account for the large variety of visual dynamics and tasks, a model should both contain a bottom-up phase and a top-down phase carrying the information about important feed-back [50].

1.3.1 Our Focus

In the current work, we sought to further investigate the possible links between computational models and neurophysiological data. Two data-sets, already published and presented in [51, 52, 34], were further analysed by trying to predict neural features starting from computational activations via a linear regression model, similarly to what has been previously discussed and implemented by [43]. In fact, analyses of this kind using intracranial neural data in human subjects are still scarce. In order to study also the role of top-down connections in a target object recognition task, the data-set presented in [52, 34] was further investigated and used to build a proof-of-concept computational model accounting for the target-modulation present between target and non-target trials already observed in neural data.

Finally, a third, unpublished, data-set was studied in the frame of our project. In this case, the patients had been presented with images all belonging to the same category and at every trial they had been asked to perform one of two possible tasks. Here, we report some of our findings regarding the neurophysiological analysis and we show how our model including top-down modulation for target-modulation can also be used in the frame of task-modulation.

2 Experimental Protocols and Data-sets

The present work has been performed on three data-sets. This was done in order to investigate the same phenomenon in the frame of different experimental protocols. A brief description of these data-sets is presented below. A summarizing table containing the main pieces of information for each one of them is shown in Tab.1. Also, note that while data-sets 1 and 2 have already been the object of investigation in previous works, data-set 3 is currently unpublished. All participants in the following experimental protocols were implanted with intracranial electrodes because they were suffering from pharmacologically intractable epilepsy.

2.1 Data-set 1

This data-set (**D1**) was composed of intracranial field potentials (IFP) recorded by 960 total subdural electrodes implanted in 11 subjects. The subjects were presented with a visual stimulus lasting 200 ms and consisting of gray-scale and contrast normalized images representing objects belonging to 5 different categories (*animals*, *chairs*, *faces*, *fruits* and *vehicles*). There were 5 different objects per category and each object could be shown in 5 different forms: standard image, 2 scaled versions (where the object was scaled 2× or 0.5× with respect to the standard size) and 2 rotated versions. This amounted to a total of 25 images for each category and to a total number of 125 images in the whole experiment(25 × 5 categories). The images were presented in a pseudo-random order and the interval between two images was 600 ms. For further details about the experimental protocol and for electrodes location, please refer to [51]. Examples of images belonging to the category *fruits* are shown in Fig.3a. Subjects were asked to perform a one-back task, that is to indicate whether an exemplar was the same as the previous one or not regardless of scale or viewpoint changes.

2.2 Data-set 2

In data-set 2 (**D2**) 10 subjects (for a total of 776 electrodes) were presented with gray-scale images containing either one or two objects belonging to 5 possible categories (*face*, *car*, *chair*, *animal* or *house*). There were 5 objects per category. The images were presented for 100 ms and there was a inter-trial interval of 500 ms between consecutive presentations. Each experimental session was separated into different blocks, each one containing 50 images. For each block, the name of the target category was displayed on the screen. For each trial, i.e. for each image presented, the subject had to press a button if the image contained the target category. The number of trials where at least one of the possible two images belonged to the target category was around 1/3 of the total number of trials inside one block. For further information about the experimental procedure and for electrodes

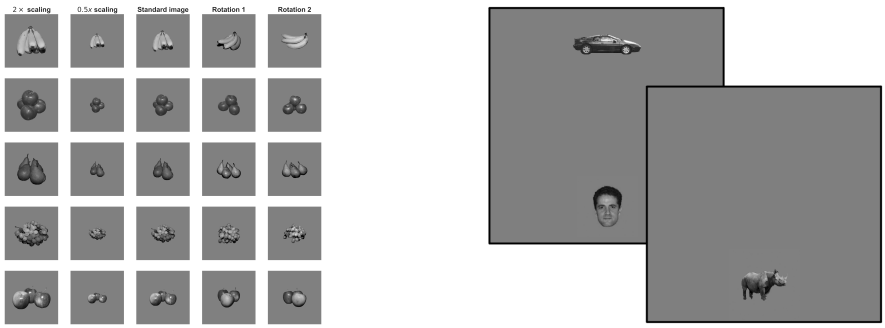


Figure 3: a) Example of the 25 images belonging to category *fruits* in **D1**. There are 5 versions × 5 objects. Each column represents a possible form in which the object was presented (either standard, scaled or rotated). **b)** Example of a two-object image and of a one-object image in **D2**. In particular, examples of objects of three out of five total categories are showed: *animals*, *faces*, *cars*.

location, please refer to [52, 34]. Examples of a single-object and two-objects images are presented in Fig.3b.

2.3 Data-set 3

Data-set (**D3**) is composed of IFP signals recorded by a total of 1201 intracranial electrodes (Ad-Tech, Racine, WI, USA; each recording site was 2 mm in diameter with 1 cm separation) implanted in 11 subjects (6 females, 12-43 years old). Recording sites were arranged in grids or strips. Patients were presented with gray scale and contrast normalized images containing a human face in the center. The faces were artificially generated via `FaceGen`. Patients had to respond to three different tasks: *Age*, *Gender* *Caricature*. The stimulus images were presented for a duration of 415 ms and the inter-trial time was 1250 ms. For each task, patients had to choose between binary labels, such as young-vs-old and male-vs-female. Notice that in our analysis we analysed only the *Age* and *Gender* trials. Also, for each task a *Pre* or *Post* condition was applied to the trial, meaning that the task to be performed could be revealed either before or after the stimulus presentation. The images were presented in a pseudo-randomized order and the experiment was divided in blocks. The data-set was perfectly balanced between *Pre* vs *Post* conditions and *Age* vs *Gender* trials, meaning that an image was always presented in the four possible combinations. The experimental protocol for both a *Pre* and a *Post* trial is presented in Fig.4. Electrodes were located according to the procedure already performed in [51, 34].

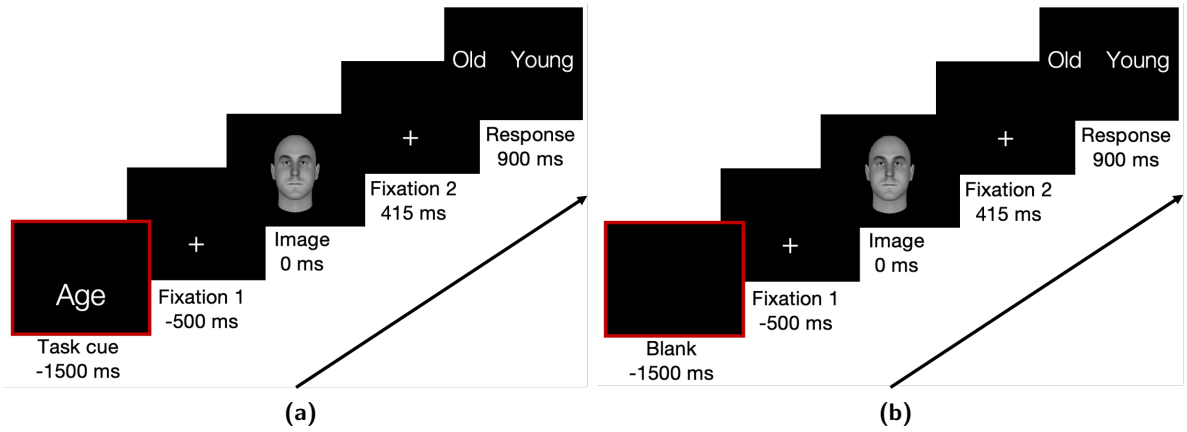


Figure 4: Timeline of the experimental protocol in **D3**. **a)** An *Age-Pre* trial and **b)** an *Age-Post* trial. The red frame points to the main difference in the experimental protocol: while in the *Pre* condition subjects were told which task they were going to perform, in the *Post* condition they found out only when the response screen was shown.

	D1	**D2**	**D3**
Human data	Yes	Yes	Yes
Type of recordings	IFP	IFP	IFP
Task	One-back task	Target recognition	Age or Gender (binary task)
Categories of objects	5	5	1 (only faces)
N. images per category	25 (5 conditions per object exemplar)	5	//
Color images	gray-scale	gray-scale	gray-scale
N. objects per image	1	1 or 2	1
Stimulus-presentation time	200 ms	100 ms	415 ms

Table 1: Summary of the main characteristics of each data-set.

3 Methods

3.1 Data Preprocessing

D1 and D2 were preprocessed in accordance with the studies where they were first published in, see [51] (**D1**) and [34] (**D2**). For a more detailed explanation refer to [51, 34]. In summary, in both cases the neural signals were band-passed filtered in the range [0.1, 100] Hz with a 4^{th}-order Butterworth filter. Also, a notch filter was applied at 60 Hz to remove electrical noise. Regarding artifacts detection, in **D1**, the trials where the total power, defined in Eq.2 (where x_t represents the voltage recorded at a time points t), was higher than the mean total power (over trials per one electrode) plus 4 standard deviations in at least 3 electrodes were removed. A similar approach was used for **D2**, but instead of the total power, the IFP_{range}, which is defined in Eq.3, was employed as a metric. t_i and t_f refer to the initial and final time of the time interval taken into consideration in the computation. Also, in this case, trials were removed also on the basis of the reaction time (RT). That is, those trials where the RT was either $>$ than 2 seconds or $<$ than 200 ms were discarded. In **D2**, only correct trials were selected for further analysis.

$$broadband\ power = \frac{\sum_t^T x_t^2}{T} \tag{2}$$

$$IFP_{range\ [t_i,t_f]} = max(IFP_{[t_i,t_f]}) - min(IFP_{[t_i,t_f]}) \tag{3}$$

Concerning **D3**, the signals were also band-pass filtered with a 4^{th}-order Butterworth filter and a notch filter was applied at 60 Hz. An artefact detection analysis was performed with the same criterion as in the **D2**. Trials considered as containing artefacts in 10% of the electrodes were removed from all electrodes. Only *Age* and *Gender* trials (correct trials only) were selected for further analysis.

In all data-sets, the IFP signals were aligned to stimulus onset and divided into trials corresponding to a time window around the stimulus onset. This process referred to as epoching. In our case, the time windows were [-100,800] ms (**D1**), [-250, 800] ms (**D2**) and [-600, 1000] ms (**D3**). All times mentioned from now on are going to be given as times relative to stimulus onset (fixed at 0 ms).

3.2 Neurophysiological Analysis

Below, all the main methods used to perform the analysis on the neurophysiological data are described. The differences among the data-sets will be explained when needed. Most of the following methods allowed us to investigate phenomena related to visual processing.

3.2.1 Noise Estimation

In **D3**, in order to get rid of noisy channels and of channels showing very low amplitude responses, a noise level of 28.8 μV was estimated. This value was computed as twice the average of the standard error of the IFP signal across all electrodes, across subjects and across all time points, separated by *Pre-Post* condition and by task (four subsets in total). This noise threshold will be used as a condition in several analysis steps in **D3**.

3.2.2 Visual Responsiveness

The neurophysiological recordings we analysed were all obtained from patients implanted with intracranial electrodes to monitor pharmacologically intractable epilepsy. This means that the locations of the electrodes were dictated by clinical reasons. It is then likely that a lot of electrodes were not even located in regions related to visual processing. In order to find visually-responsive (VR) electrodes in

D2 and **D3**, a permutation test was performed for each electrode independently.

The IFP_{range} for all trials was computed in the time window of interest (WOI) [50, 300] ms . This time window was chosen accordingly to [51]. The IFP_{range} was also computed over the time-window [-250, 0] ms, which was used as baseline. A customized metric, shown in Eq.4, was then calculated. $\overline{IFP_{range\,\{\,\}}}$ represents the average IFP_{range} across trials and over a certain time window. VR electrodes were then identified through a permutation test, where the procedure just explained was repeated 1000 times by shuffling the values between the baseline and the WOI.

$$VR\,metric = \frac{\overline{IFP_{range\,\{WOI\}}} - \overline{IFP_{range\,\{baseline\}}}}{SD(IFP_{range\,\{WOI\}}) + SD(IFP_{range\,\{baseline\}})} \tag{4}$$

Additionally, in **D3**, the IFP_{range} in the chosen time window [50, 300] ms was required to be $> 28.8\mu$V according to the noise estimation explained in Section 3.2.1.

3.2.3 Visual Selectivity

D1: A one-way ANOVA was used to test the visual selectivity (VS) of electrodes. This technique tests whether more groups belong to the same population (thus have the same mean) or not. This is done by analysing the variance within groups and the variance across groups' means. More precisely, the null (**H**) and alternative hypotheses (**A**)are as following,

$$\mathbf{H} : \mu_1 = ... = \mu_i = ... = \mu_N$$
$$\mathbf{A} : \exists\, i, j : \mu_j \neq \mu_i,\ \ 1 \leq i, j \leq N$$

where N represents the number of groups to be compared. Notice that the alternative hypothesis of the ANOVA test does not give any information about which group is statistically different from the others. In our case, the one-way ANOVA was applied over the IFP_{ranges} extracted in the time window [50, 300] ms, as it was done in [51]. We assumed that the basic requirements for the ANOVA test (normal distribution of the data, sampling independence and an equal variance among groups) were fulfilled. In order to find *category-selective* electrodes, Tukey's method was used as a *post hoc* analysis. This method, which is based on a studentized range distribution, allows to perform multiple comparisons among all possible group pairings in a single step by controlling for the overall family-wise error. It compares the mean of every group to the means of all other groups, using as null (**H**) and alternative (**A**) hypotheses:

$$\mathbf{H} : \mu_i - \mu_j = 0$$
$$\mathbf{A} : \mu_i - \mu_j \neq 0$$

where i and j represent two different groups. The same hypotheses as for the one-way ANOVA apply here. In order to be considered as a *category-selective* electrode, a category had to show a significant difference with all the other categories.

D2: In the case of data-set 2, a non-parametric permutation test was preferred to asses the visual selectivity of the electrodes recorded. In fact, a non-parametric test requires no assumptions about the data distribution. First, all trials were assigned to a certain category according to the objects presented in the image. Trials corresponding to a two-objects image were associated with both categories whenever the two objects belonged to different categories. This was done in order to avoid a bias towards a certain category and because it is not possible to know whether the signal recorded is more affected by one object rather than by the other. The IFP_{range} (Eq.3) was then computed for each trial in the time window [50, 300] ms.

In order to find VS electrodes, a customized metric inspired by the ANOVA test was implemented. The criterion is defined in Eq.5, where c is the category index, C is the number of categories, $\overline{IFP_{range\{c\}}}$

indicates the average IFP_{range} across trials within a category and Δ represents the deviation of the means. The idea here was to compute the difference between the average neural response (quantified as the IFP_{range}) computed within every category and the mean average response computed across categories. If the value is high for one category, it means that that category diverges from the others. Moreover, we divided by the standard deviation within a category. This leads to a preference for categories of images that show a large deviation from the mean across categories and whose recordings are not widely spread. Finally, we took the maximum value across categories. The same metric was computed 1000 times by shuffling the category labels to build a null (chance) distribution.

$$VS\ metric = max_{\{c\}} \left(\frac{\Delta_{\{c\}}}{SD_{\{c\}}} \right)$$
$$\Delta_{\{c\}} = \left| \overline{IFP_{range\{c\}}} - \frac{\sum_{c=1}^{C} \overline{IFP_{range\{c\}}}}{C} \right| \tag{5}$$

3.2.4 Top-Down Modulation

One of our main focuses in the current project was top-down modulation, both from a neurophysiological and a computational point of view. In **D2** we investigated target-modulation (TM), that is the top-down modulation due to the presence of the target on the current stimulus image (see Section 2.2 for details about the experimental protocol). Bansal and colleagues found that the target vs non-target modulation begins on average at around 250 ms after stimulus onset [34]. In the case of **D3**, we can distinguish two possible top-down modulations. The first one is what we called task-modulation (TkM) and refers to the possible modulation introduced by the simple fact of knowing that there is a task. This modulation could be probed by comparing *Pre* vs *Post* trials (for a certain task). The second one is what we called task-dependent modulation (TDM), and refers to the potential difference observed between *Age* and *Gender* trials in the *Pre* condition (so when the task was known).

D2: The approach we followed to find target-modulated electrodes was similar to the one applied in [34]. Electrodes showing significant difference between target and non-target conditions for more than 70 consecutive milliseconds were considered to be target-modulated. The 70 ms minimal threshold was computed as the amount of consecutive ms necessary to ensure a False Discovery Rate (FDR) $<$ 1%. To compute the threshold, the "target" vs "non-target" labels were shuffled and a t-test for every time point was performed for every electrode and every subject. Consecutive significant ($p < 0.01$) points were found and the duration of these intervals was computed. This procedure was performed 1000 times and the lengths of the significant intervals were all concatenated and sorted to find the 99% quantile.

D3: The same procedure as for **D2** was implemented for the comparison of *Pre* vs *Post* (for both *Age* and *Gender* tasks) and *Age* vs *Gender* (for the *Pre* condition). However, in order to be even more sure to avoid false positives, the minimal time threshold was found in a way to ensure a FDR $<$ 0.1%. In addition, this analysis was performed on three time windows independently: Fixation 1 (F1, [-500,0] ms), Stimulus Presentation (SP, [0, 415] ms) and Fixation 2 (F2, [415,930] ms). To be considered as being modulated, an electrode had to pass our test in at least one of the three time windows. All time windows were equalized to have the same duration, that of the SP time window (we wanted to make sure to analyse time windows having the same lengths in order to discard the potential confusion factor arising from the difference in duration). For **D3** the minimal time threshold required in order to be considered a TkM or a TDM electrode was computed to be 93 ms. In addition, a minimal difference of 28.8 μV between the two conditions (*Pre* vs *Post* or *Age* vs *Gender*) was required (see Section 3.2.1).

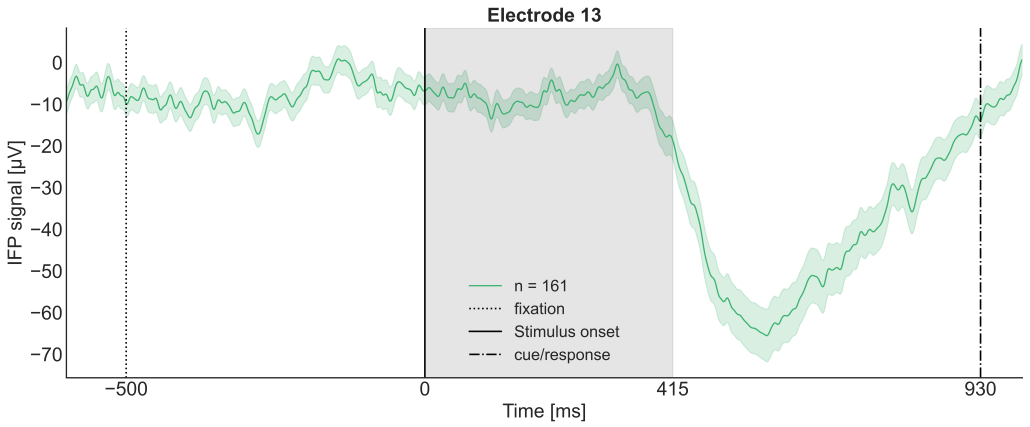


Figure 5: Example of a Delayed-Responsive electrode found in subj1 and implanted in the frontomarginal sulcus and gyrus. The average IFP signal ± the standard error computed across trials is shown. Here, *Age* and *Gender* trials are mixed, but only the *Post* trials are considered. We can see how there is no response during the stimulus presentation window, while a response is present around stimulus offset (at ≈ 400 ms).

3.2.5 Delayed Responsiveness Analysis

By looking at the IFP signals across electrodes and across subjects, we observed an interesting phenomenon: some electrodes presented a response after stimulus offset but did not show any response during the stimulus presentation time window. An example of this observation is shown in Fig.5. We decided to look more into this phenomenon and we called these electrodes delayed-responsive (DR). For this analysis we considered only those electrodes having failed the VR test (Section 3.2.2). A permutation test identical to the one performed to find VR electrodes was used by selecting as WOI the interval [415, 665] ms and as baseline [50, 300] ms. Note how the current baseline time window corresponds to the WOI used for the VR test. Moreover, in order to avoid as much as possible false positives, the IFP_{range} in the WOI was additionally required to be $> 28.8\mu$V and the IFP_{range} in the interval [300, 415] ms was required to be $< 28.8\mu$V (see Section 3.2.1).

This analysis was performed by considering *Age* and *Gender* trials together, but by separating the *Pre* and *Post* conditions. In fact, such a late response could potentially be linked to a modulation due to the knowledge of the task. It was then crucial to investigate the *Pre* and *Post* phenomenon independently.

3.3 Linear Mapping Analysis

We were interested in understanding how well the activation maps extracted from different layers in VGG19 can correlate to neural data extracted from electrodes located at different positions along the visual pathways. In order to do so, VGG19 was chosen as DCNN of reference because of its high performance in object recognition in computer vision. VGG19 was fed with the same images as the ones the subjects participating in the study were presented with. Our analysis focused on 7 main layers: the 5 pooling layers at the end of every convolutional block and the 2 final fully-connected layers. The architecture of VGG19 is shown in Fig.2. Once the activation maps had been extracted, a feature selection process was performed in order to reduce the dimensionality of the data and to find the most suitable features. Finally, these selected features were used to linearly regress the neural metrics associated to the same images. The linear regression is formulated in Eq.6.

$$\mathbf{Y}_{neural\ responses} = \beta_0 + \beta_1 \mathbf{PC}_1 + \beta_2 \mathbf{PC}_2 \tag{6}$$

$\mathbf{Y}_{neural\ responses}$ represents the vector containing the neural responses extracted for each image (by averaging over repetitions of the same image), $\beta_{0,1,2}$ are the coefficients of the linear regression and and $PC_{1,2}$ are the two PCs computed from the activation maps extracted in VGG19 and selected

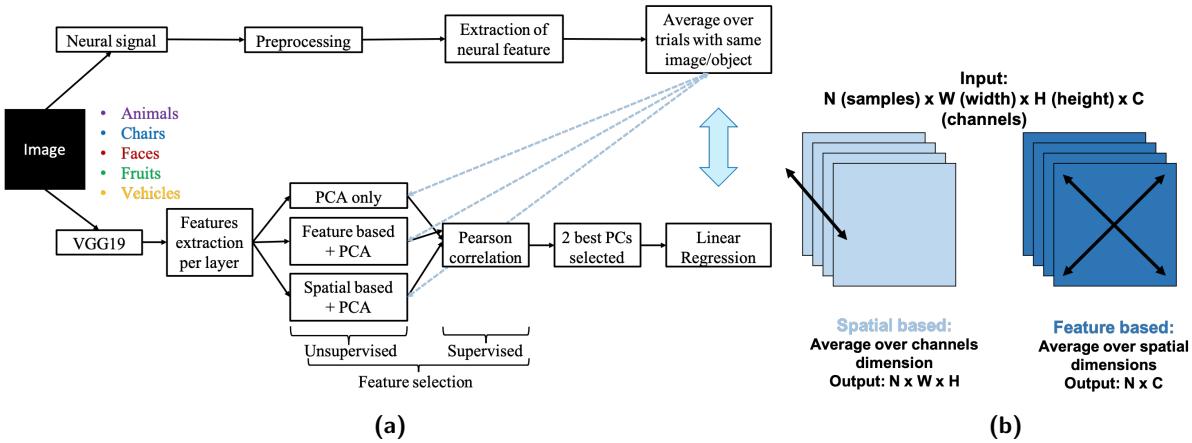


Figure 6: a) Scheme of the overall analysis leading to the linear mapping between neural features and computational activations. The final link between neural data and computational activations is indicated here by the big light blue arrow. On the top the steps performed on the neural data are shown, while on the bottom the ones related to VGG19 are presented. The blue dotted lines highlight the fact that the Pearson correlation is computed between the extracted PCs and the final vector containing neural features.**b)** A sketch to explain more in detail the feature-based and the spatial-based averaging approaches implemented in **D1**.

because of the high Pearson correlation. For both data-sets **D1** and **D2**, the neural metric used was the IFP_{range}. The train-test split was 80-20% and the procedure was repeated 20 times (20-folds). Note that in the case of **D2** the split was not completely random: because of the low number of samples we made sure to always have 4 out of 5 objects per category in the training set and 1 out of 5 in the test set. The most important steps in the linear mapping analysis are detailed below and an overview of the whole procedure can be found in Fig.6a.

3.3.1 Split-Half Self-Consistency

To compute the split-half self-consistency (SHSC) the trials (repetitions) corresponding to the same presented image were divided in two halves. For same image we mean here an image containing the exact same object(s). The average neural response (defined as the IFP_{range}) was then computed for both halves. While theoretically speaking one should expect a correlation value of 1 (since the two average values for each image should represent the same thing), this is not the case with experimental data. The correlation value found can be considered as an estimation of the consistency of the recorded data where the deviation from 1 can be attributed to intrinsic noise given by the huge variability in neurophysiological data.

The split into the two halves was performed randomly and the estimation of the split-half self-consistency was repeated 100 times in order to obtain a more robust estimation of the Pearson correlation. In addition, since some images were repeated only a few times, a threshold on the number of trials required for each image was set to 4 to ensure reliability in **D1**. Significant SHSC values were found with a randomization test: a chance-level SHSC was computed by repeating this procedure 500 times by shuffling the labels (the image identities) among halves at each repetition.

The split-half self-consistency value can be interpreted as a measure of an upper bound on the regression performance when trying to build a linear model to predict neural data. In fact, no model should be able to predict neural data better than neural data itself. However, it is possible that the value obtained for the SHSC is an underestimation of this upper bound because of the fact that the SHSC computation relies on only half of the repetitions. In order to correct for this when performing the linear mapping analysis, the Spearman-Brown correction, which can be found in Eq.7, was applied. ρ

represents the Pearson Correlation.

$$SB\_corrected\_SHSC = \frac{2 * \rho}{1 + \rho} \tag{7}$$

While the images in **D1** contained only one object, this was not the case for **D2**. The extremely high number of different combinations of two objects possibly contained in one image was the cause of the low number of repetitions for a specific image. In order to increase the robustness of our approach, we decided to follow an "object-based" approach for the linear mapping analysis in **D2**. In the object-based approach, all images containing two objects were assigned to one of the two objects randomly. This led to 25 neural measurements (equal to the number of objects). Note that two-object images were not assigned to both objects in order to avoid repeating trials in this analysis. In fact, since the linear model needs to be validated by splitting the data in train and test sets, it is suitable not to repeat trials to prevent the test set to contain a part of data already considered in the training set and used to fit the model.

3.3.2 Feature Extraction in VGG19

When extracting the activation maps from the different layers of VGG19, three different approaches were followed in **D1** in order to get the final computational features to use for the linear regression analysis: PCA-only, feature-based and spatial-based. These three methods differ only in the case of pooling or convolutional layers. In the first approach, the activation maps were flattened and a Principal Component Analysis (PCA) was performed. The PCs accounting for 95% of the variance were kept. The test set activation maps were projected into the new space according to the PCA transformation matrix computed on the train set. The second and third approach both contain this step. However, before flattening the activation maps, an averaging was performed along a certain dimension. In the feature-based averaging, the activation maps are averaged along the spatial dimensions of the image. This means that if the input has dimensions $N(samples) \times W(width) \times H(height) \times C(channels)$, the output will have dimensions $N \times C$. Note that we call it feature-based because the information related to the different features extracted (i.e. the different channels) is preserved. In the spatial-based approach the averaging is performed along the channels dimension, hence giving as output a structure having dimensions $N \times W \times H$. The two averaging approaches are summarized in Fig.6b. Because of the object-approach followed in **D2** for the linear mapping analysis, the activation maps for the different images chosen to represent a certain object were averaged similarly to what was done on the neural data. After this averaging, only the PCA-only approach was implemented in **D2**.

Following the step of dimensionality reduction, we selected 2 features for further analysis in a supervised manner. The Pearson Correlation between the Principal Components (PCs) extracted from the train set and the final vector of neural metrics (i.e. the values to regress) was computed for each PC independently. The 2 PCs showing the highest correlation in absolute value were selected and used to build the final linear regression model.

3.3.3 Linear Regression

A simple least squares linear regression was implemented by using the module `LinearRegression` from the `python` module `sklearn.linear_model`. The features were normalized before fitting the model through the use of a `StandardScaler` instance from the module `sklearn.preprocessing`. The test dataset was normalized according to the mean and standard deviation estimated over the training set.

3.3.4 R^2 Score

The R^2 score (also called coefficient of determination) was chosen as the metric measuring the performance of the linear regression because of its relatively straightforward interpretability. It represents the variance in the dependent variable that is predictable from the independent variable(s) for the

specific model. Despite that it is called R "squared", the R^2 value can range between $(-\infty, 1]$, where the values in the range (0,1] mean that the model is better than a model defined as just the average of the dependent variable data points. The possibility for negative values comes from the definition of R^2 as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \qquad SS_{tot} = \sum_i (y_i - \bar{y})^2$$

where $\hat{y}_i$ represents the linear regression estimate for a certain point i and $\bar{y}$ stands for the average over all dependent variables. However, notice that in some cases, especially when referring to R^2_{train}, i.e. the R^2 score computed on the set of points used to fit the linear model, $R^2 = r^2$, where r represents the Pearson correlation.

3.4 Computational Modeling

The bulk of studies trying to link and to mimic neurophysiological data via computational modeling has been focusing on the feed-forward component of the visual pathways. As previously mentioned, feed-back connections are abundant in the brain and can come as both lateral (horizontal) and top-down connections. Here, we sought to investigate these latter by implementing a top-down modulated DCNN as a proof-of-concept model building from the VGG19 architecture. The main aspects of this computational analysis are detailed below.

3.4.1 Model Architecture

While top-down connections linking non-consecutive regions in the brain could be present (for example feedback connections going from IT to V2), for the sake of simplicity we decided to focus on top-down connections between adjacent regions. Computationally speaking, this could be implemented by adding some sort of modulation coming from a higher layer in the hierarchy down to the previous layer. We focused on the highest layers in the hierarchy, i.e. the fully-connected (FC) layers, which are usually referred to as the classifier part of the network and as the part mimicking the IT cortex because of their role in combining features to allow object identification. We associated the prediction layer of VGG19 to be a prefrontal cortex (PFC)-like layer, where the comparison between the object identified from the input image and the task-goal occurs. From the prediction layer a target(task)-dependent modulation was sent to FC2, which in turn sent back some modulation to FC1, as shown in the top part of Fig.7. In order to do so, we used VGG19 pre-trained on ImageNet and we transposed the weights matrix to allow the modulation to flow backward. This strategy allowed us to avoid the training of the new connections created. Moreover, since our purpose was not object recognition, we changed the activation function in the prediction layer from $softmax()$ to $ReLU()$. Eq.1 represents the general function applied in FC layers in the feed-forward sweep. Our implementation of top-down modulation for two consecutive layers is mathematically summarized in Eq.8.

$$\begin{aligned} \mathbf{x}^l(t=2) &= \Phi\left(\mathbf{W}^{l+1^\mathbf{T}}\left(\mathbf{x}^{l+1}(t=2)\right)\right) + \mathbf{x}^l(t=1) \\ \mathbf{x}^{l-1}(t=2) &= \Phi\left(\mathbf{W}^{l^\mathbf{T}}\left(\Phi\left(\mathbf{W}^{l+1^\mathbf{T}}\left(\mathbf{x}^{l+1}(t=2)\right)\right) + \mathbf{x}^l(t=1)\right)\right) + \mathbf{x}^{l-1}(t=1) \\ \mathbf{x}^L(t=2) &= \mathbf{x}^L(t=1) \end{aligned} \tag{8}$$

$\mathbf{x}$ is the activations vector and for a layer l it has the dimensions N_units^l (L = the last layer). Note how $\mathbf{x}(t)$ is a function of time, since it is supposed to model the full temporal dynamics of neural signals. In our case, we only have two phases: $t = 1$ indicates the first object-recognition phase, while $t = 2$ is the second phase, that is when layers receive feed-back from the prediction layer. $\mathbf{W}$ represents the weight matrix and, for instance, $\mathbf{W}^l$ has dimensions $N\_units^{l-1} \times N\_units^l$. $\Phi()$ is the activation function. For each layer, after having applied the activation function to the modulation obtained, the modulation was added to the activations of the layer feed-forward counterpart.

3.4.2 Modulation

The modulation itself was implemented in a target- or task-specific way. In the case of target-modulation (hence for **D2**), we decided to send back as top-down signals the information contained in the activations related to the target category, irrespectively of a target (that is when the image presented contains an object belonging to the target category) or a non-target trial. In fact, the model is supposed to perform object recognition and thus, in the case of a target trial, the activations related to the target category should be higher with respect to the non-target trial case. A similar approach was used in **D3** to model task-modulation: only the information related to the current task was sent backward. The middle and bottom sections in Fig.7 show how the modulation was implemented in both **D2** and **D3**. The colored units in the prediction layers represent the units associated to the different categories. The identification of these units is detailed in Section 3.4.3. Note how the modulation itself occurs only at the level of the prediction layer sending information backward to FC2. From there, the modulation propagates through layers without any further input.

3.4.3 VGG19 Category-Mapping

VGG19 is a DCNN model trained on ImageNet and has a prediction layer containing 1000 units, one per category. As mentioned above, the key function of our top-down modulated model relies on sending back task- or target-specific information. In order to do so it was then important to decide how our categories of interest (*fruits*, *cars*, *faces*, *animals* and *chairs* in **D2** and *Age* and *Gender* in **D3**) were represented in the prediction layer of VGG19.

D2: The VGG19 prediction layer's categories do not include our 5 categories. This is also due to the fact that our categories are rather general (e.g. *Animals*) with respect to the ones present in ImageNet. For instance, ImageNet contains tens of categories related to specific species of dogs and cats. In order to find which units we should use for sending back target category-related information, we fed the original VGG19 model with the 25 single-object images (see Section 2.2 for further details on the experimental protocol), each one showing a different object. For each image, we sorted the activations in the prediction layer and a certain number of units N_units was then selected and assigned to the object contained in the image. The units mapped to all the objects belonging to a category represented the new definition of that category within VGG19's prediction layer. This means that if for instance just 1 unit was selected per object, then the category linked to that object will be represented by 5 units. This mapping between the prediction layer activations and our categories was done in an exclusive fashion, meaning that no object could share the same units, even if belonging to the same category. This was done in order to prevent a category to be represented by more units than another.

Since the mapping between the prediction layer's units and our categories was exclusive, a maximum number of 40 units could be chosen per object (in fact, $25 \times 40 = 1000 =$ number of units in VGG19's prediction layer). The modulation analysis was performed repeatedly by incrementing N_units by 1 unit until $N\_units = 40$ was reached. To measure the impact of N_units, three different measures were computed: the p-value of the t-test between the target and the non-target activations during the feed-back step, the absolute difference between the average target activation ($\overline{Target}$) and the average non-target activation ($\overline{Non\_target}$) and a customized metric shown in Eq.9.

$$N\_units\ metric = \frac{|\overline{Target} - \overline{Non\_Target}|}{SD(Target) + SD(Non\_Target)} \tag{9}$$

D3: In the case of data-set 3, the mapping between VGG19's prediction layer and our categories was more complicated. In fact, while the images in **D2** contained different objects leading to different patterns of activations in the prediction layer, this was not the case in **D3**. Here, both tasks were applied

to the same sets of images, which made it impossible to use the images to map out the corresponding activations and create new artificially chosen categories as we did for **D2**. For the sake of clarity, think of an image showing a cat and one showing a car in **D2**. These two images are going to give two different patterns of activations in VGG19's prediction layers and it is then possible to select the units showing the highest activations for both images. In **D3**, the image of a face was associated to both *Age* and *Gender* and the pattern of activations in the prediction layer will not change depending on the task.

All the possible images in **D3** (no matter the *Age*, *Gender* or even *Caricature* parameters), were fed to the VGG19 model (containing the $ReLU()$ activation function instead of the $softmax()$ one in the prediction layer) and the activations were stored. This gave a matrix having the shape $N\_images \times 1000$. This matrix can be considered as being a new data-set where each sample is associated with 1000 features. For each sample a label related to the *Age* task (i.e. young vs old) and one related to the *Gender* task (male vs female) was stored. The order of samples was randomized and the data-set was divided in a train and test sets with a 90-10% split. Linear Discriminant Analysis (LDA) based on a least squares approach was used to independently build two classifiers for each feature: one to solve the binary task related to *Age* and one for *Gender*. The balanced accuracy (BA), which is the arithmetic average of sensitivity and specificity, was preferred to the usual accuracy as a metric to evaluate the performances of the classifiers since the two classes for each task were not perfectly balanced in the data-set. You can find details about balanced accuracy in Eq.10. The BA values were sorted and the units giving a BA > 0.7 were chosen to represent the task. TN/FN/TP/FP stand for True-False Negatives-Positives. This entire procedure was repeated 20 times in order to ensure robustness in the selection process of the units.

$$
\begin{aligned}
Balanced\ Accuracy &= \frac{sensitivity + specificity}{2} \\
Sensitivity = \frac{TP}{TP + FN} \qquad & Specificity = \frac{TN}{TN + FP}
\end{aligned}
\tag{10}
$$

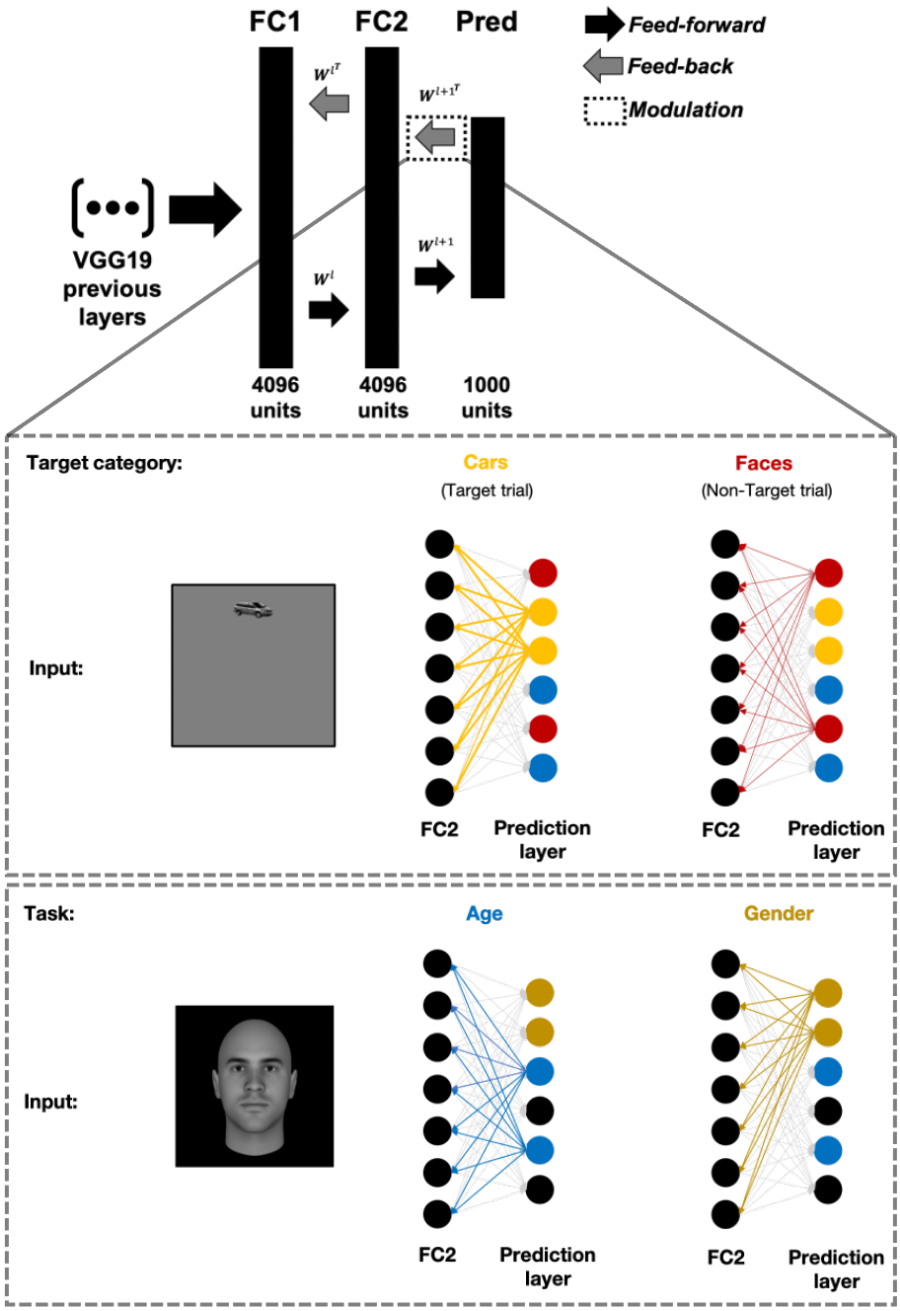


Figure 7: Illustration of our VGG19 top-down modulated model. On the top the general concept of top-down modulation is shown. Our analysis focused on the fully-connected layers of VGG19. The prediction layer here acts like a PFC-like layer, meaning that it sends a feed-back signal to the previous region to modulate the feed-forward signals. The feed-back is implemented by using the same weight matrices as in the feed-forward counterparts, but by transposing them. Note that the modulation itself originates at the level of the prediction later (indicated by the gray rectangle) and then it is simply propagated backward. For a mathematical formulation see Eq.8. On the bottom part, a zoom-in over the implementation of the modulation is shown. For illustration purposes, only few units for the FC2 and the prediction layer are showed. The colored units in the prediction layer represents the newly defined categories. For **D2**, no matter if the image represents a target or a non-target trial the prediction layer sends back the information stored in the units representing the target category. In the case of a target trial, like when the target category is *Cars*, the activations sent backward are supposed to be relatively high, since the object was indeed present in the input image. On the contrary, in the case of a non-target trial, as when the target category is *Faces*, the activation values in the prediction layer should be lower because of the actual absence of the target category in the input image. The supposed lower activations in a non-target trial are indicated here through thinner backward connections with respect to those sent back in the case of a target-trial. Light gray connections simply indicates the feed-forward connections. Finally, a similar implementation is used for **D3**, where information related to the *Age* vs *Gender* tasks are sent backward accordingly to the task we wanted to model. No hypotheses about the magnitude of activations were present for **D3**.

4 Results

4.1 Visual Selectivity

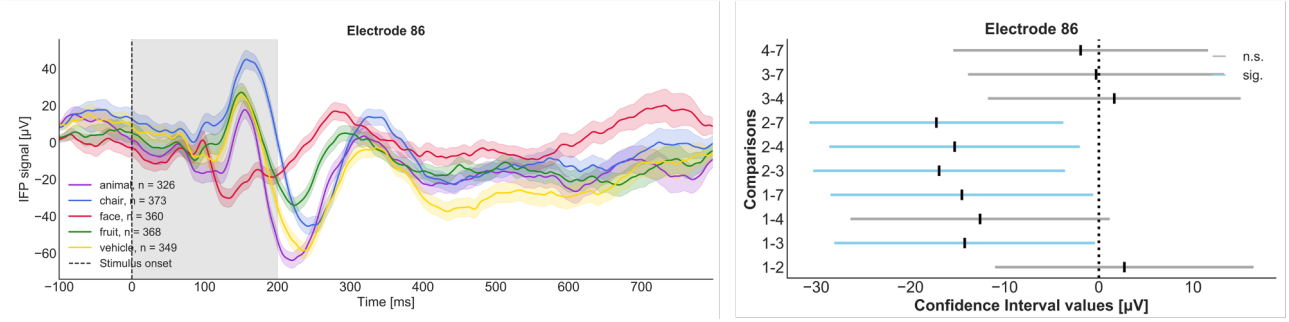
In order to assess whether some of our electrodes were selective to a specific category of objects among the ones presented, some statistical tests were run over **D1** and **D2**. In particular, in **D1** a one-way ANOVA was used to assess visual selectivity and it was then followed by a *post hoc* Tukey test to find a potential favorite category. A category was considered as being the favorite one if the comparisons with all other categories resulted significant. In the case where more than one category satisfied the condition, the one being overall more significantly different from the others was chosen. Regarding **D2**, a first selection aiming at finding visually-responsive (VR) electrodes was performed. Then, a permutation test was applied over the VR electrodes to identify visually-selective electrodes.

D1: Subject 7 was removed from the analysis because the neural recordings were particularly noisy. Out of the 856 electrodes analysed (960 total - 104 from subject 7), 128 ($\approx 15\%$) resulted as visually selective ($p < 0.05$). 10 ($\approx 8\%$) of these selective electrodes did not have a well established location. The visually-selective (VS) electrodes identified did not span uniformly the different lobes, with the occipital and temporal ones containing most of them. Among the electrodes placed in a defined region, 22 ($\approx 19\%$) were found in the inferior temporal gyrus, 16 ($\approx 14\%$) in the mid-temporal gyrus, 10 ($\approx 8\%$) in the fusiform gyrus and 9 ($\approx 8\%$) in the mid-occipital gyrus.

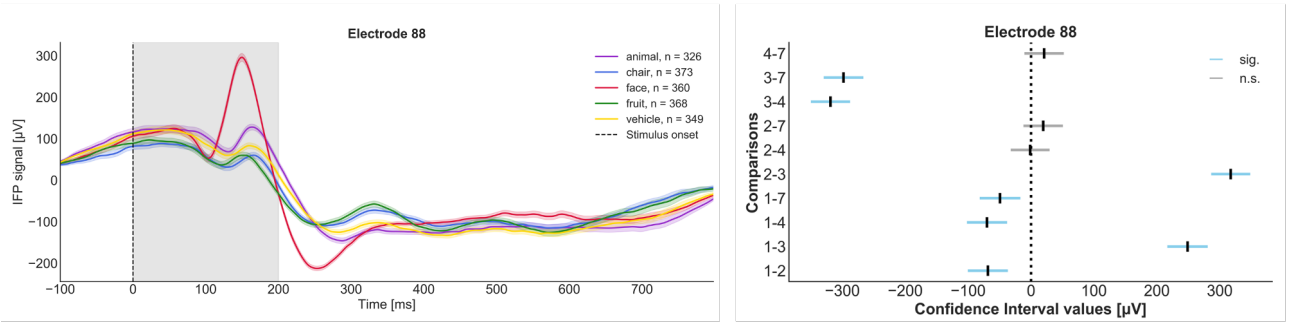
Out of the 128 VS electrodes, 18 ($\approx 14\%$) resulted as being also category selective (overall family-wise error rate for Tukey's tests fixed at $\alpha = 0.05$). Category selective electrodes were found only in 5 subjects out of 10. No electrodes were found to be selective for category 1 and 7 (i.e. *animals* and *vehicles*), while most of them resulted to be selective for category 3 (i.e. *faces*, $\approx 67\%$). The locations of the identified category selective electrodes spanned several areas, mainly in the occipital and temporal lobes, with 4 electrodes ($\approx 22\%$) in the inferior occipital gyrus and sulcus, 4 electrodes in the medial occipitotemporal gyrus and 3 electrodes ($\approx 17\%$) in the inferior temporal gyrus. Fig.8 shows an example of a VS-only electrode and an example of an electrode being also category-selective. Note for example how electrode 86 showed a significant difference for all pair-wise comparisons related to category 2 with the exception of the comparison with category 1. On the other hand, electrode 88 satisfied the requirement to be a category-selective electrode for both categories 1 and 3. Category 3 was finally considered as the favorite one because of the more significant difference.

D2: Out of a total of 776 electrodes analysed, 152 ($\approx 20\%$) were considered as being visually-responsive ($p < 0.01$). Visually-responsive electrodes were found in all subjects. The locations of electrodes for subject 10 were not available (53 VR electrodes, $\approx 35\%$). Out of the located 99 electrodes, 16 electrodes ($\approx 16\%$) were in the superior temporal gyrus, 15 electrodes ($\approx 15\%$) in the inferior temporal gyrus, 10 electrodes ($\approx 10\%$) in the fusiform gyrus and 8 ($\approx 8\%$) were placed in the middle temporal gyrus.

Out of the 152 VR electrodes, 22 ($\approx 14\%$) passed our randomization test ($p < 0.01$) and were considered as being visually-selective. VS electrodes were found in 5 subjects out of 10. In this case we considered the favorite category as being the one maximizing our customized metric (see Eq.5). All categories except for category 4 (i.e. *animals*) were represented across the visually-selective electrodes identified, with the majority of them ($\approx 59\%$) being selective to category 1, i.e. *faces*. Regarding the locations, the placement of 1 electrode (belonging to subject 10) was not known. Among the located electrodes, 12 electrodes ($\approx 57\%$) were in the inferior temporal gyrus, 8 electrodes ($\approx 38\%$) were in the fusiform gyrus and 1 electrode ($\approx 5\%$) was located in the medial occipitotemporal gyrus. This means that with respect to the VR electrodes found, the 80% of the electrodes in the inferior temporal gyrus and the 80% of the electrodes in the fusiform area were also visually-selective. An example of a visually-selected electrode implanted in the inferior temporal gyrus is shown in Fig.11a.



(a)



(b)

Figure 8: Example of **a):** a visually-selective electrode located in the lingual gyrus and **b):** a category-selective electrode for category *faces* located in the parahippocampal gryus. Both electrodes are from subj10. On the left, the IFP signal divided for the five different categories. The standard error (colored shaded area) is computed over trials belonging to the same category. The shaded gray part shows the period of time during which the stimulus was presented. On the right, the simultaneous confidence intervals given as output by the Tukey's test are reported.

4.2 Linear Mapping Analysis

One of the main goals of the current work was to study the similarity between the mechanisms underlying object recognition in the brain and the ones allowing object recognition in deep learning. This was done by trying to regress neural data starting from activation maps extracted from VGG19. Our analysis focused on the main layers of VGG19 architecture, that is those layers marking an important step. The layers considered were the 5 pooling layers and the 2 final fully connected layers. In **D1** a deeper analysis was performed, since three different ways to extract important information from VGG19 were investigated: PCA-only, feature-based and spatial-based. These three methods diverge only when considering the pooling layers. In **D2** only the PCA-only method was implemented.

Those electrodes having passed the Tukey's tests for **D1** and the visually-selective electrodes identified on **D2** were considered for this analysis. In order to be finally selected for the linear mapping procedure, an electrode had to pass an additional permutation test ($p < 0.05$) aiming at selecting only those electrodes showing a significant SHSC. All electrodes, both in **D1** and **D2**, passed this test.

Qualitatively speaking, the performance obtained by the linear regression was satisfying. One way to qualitatively assess the goodness of our model for a certain electrode is to consider good a R^2 being consistently higher than zero for at least one VGG19 layer, i.e. with the error bar not touching zero. By following this criterion we can say that 18 out of 18 electrodes in **D1** (by considering at least one averaging method satisfying the criterion) showed good performances. In addition, when comparing the three different averaging methods applied in **D1**, the PCA-only and feature-based looked in general very similar, while the performance of the spatial-based method was generally worse. If focusing on the pooling layers and using our qualitative criterion, 17 electrodes out of 18 for PCA-only ($\approx 94\%$), 18 electrodes out of 18 for feature-based and 10 electrodes out of 18 ($\approx 55\%$) for spatial-based gave

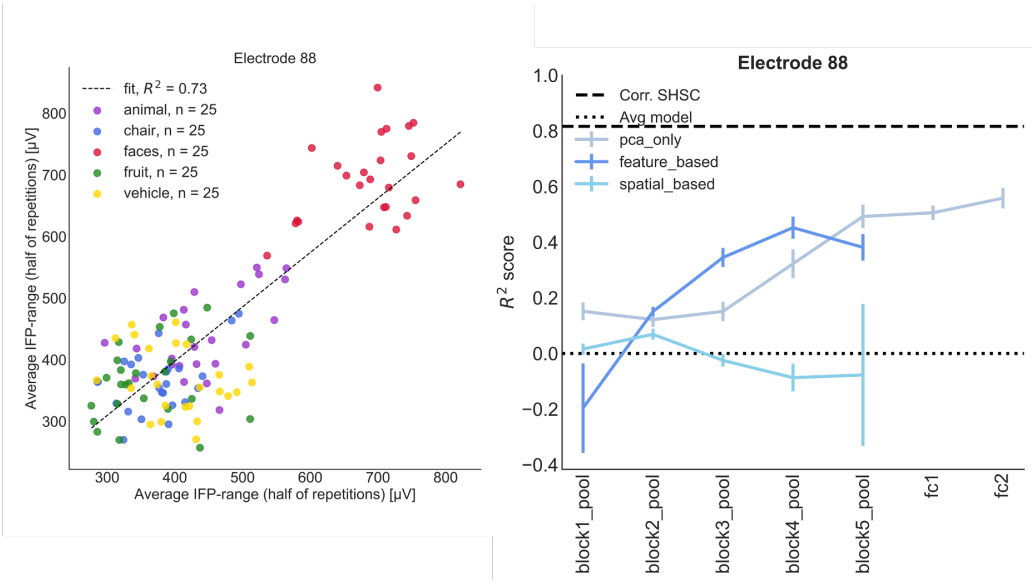


Figure 9: On the left the Split-Half Self-Consistency for electrode 88 in subj10 is shown. This electrode was located in the parahippocampal gyrus. Each color represents a different category and each point in the graph is a specific image in the data-set 1. Note that this represents only one of the 100 repetitions performed to compute a robust SHSC value. On the right, the test performance from the linear mapping analysis for the same electrode is shown. The different colors represent a different approach during the features extraction from VGG19. The error bars are computed over 20 folds. The thick dashed line represents the squared Spearman-Brown corrected SHSC.

good results. For some electrodes and for some layers, negative R^2 scores were consistently found. This implies that the linear regression model was actually worse than the horizontal line being the average of the test set samples.

When looking at the performance through the different VGG19 layers analysed, a general increasing pattern was found in most electrodes in **D1** (similar to the one observed in Electrode 88 in subj10 in Fig.9). However, this was not always the case, as it is illustrated by electrodes 26 (from subj17) and 82 (from subj10) in **D1**, whose performances are shown in Fig.10. These electrodes were implanted

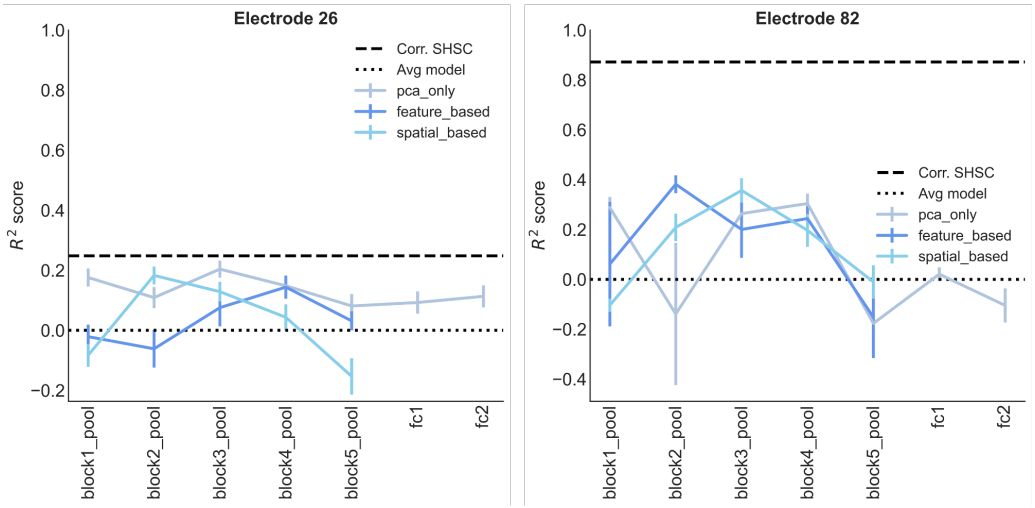


Figure 10: Linear mapping analysis performances computed on the test set for electrodes 26 (subj17) and 82 (subj10). Electrode 26 was located in the inferior occipital gyrus and sulcus, while electrode 82 was placed in the occipital pole. The different colors represent a different approach during the features extraction from VGG19. The error bars are computed over 20 folds. The thick dashed line represents the squared Spearman-Brown corrected SHSC.

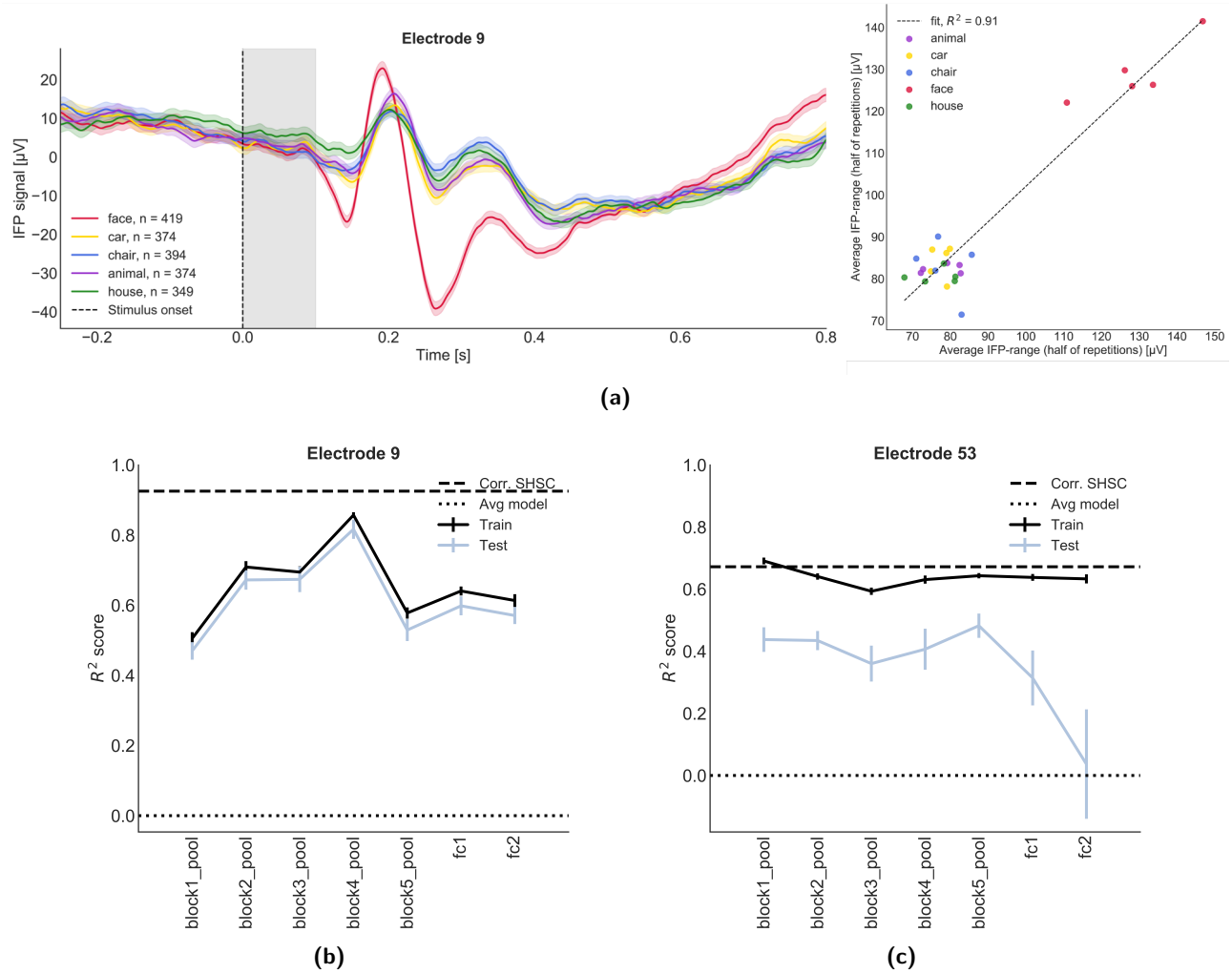


Figure 11: a) Electrode 9 (subj6) in **D2**. On the left the IFP signal divided per category is shown. The standard error (the shaded) area is computed over trials corresponding to images containing objects of the same category. The shaded gray area represents the stimulus presentation period. On the right the corresponding SHSC is shown. **b)** The linear mapping analysis performance for the same electrode, which presented the inverted V shape pattern found in several electrodes. **c)** The performance of the linear mapping analysis for another electrode (Electrode 53 from subj4) showing a decreasing performance throughout layers. Both electrodes were implanted in the inferior temporal gyrus. The black line represents the training performance, while the gray one is the test performance. The error bars are computed over a number of folds here ranging from 17 to 20 (depending on the number of outliers removed). The thick dashed line represents the squared Spearman-Brown corrected SHSC.

in the inferior occipital gyrus and sulcus and in the occipital pole, respectively. It is also worth noting how in these electrodes the spatial-based averaging method resulted comparable to the other two methods. Interestingly, two other electrodes showed a satisfying spatial-based performance (especially for the first layers). They were located in the inferior occipital gyrus and sulcus and in the medial occipitotemporal gyrus. Note though that some other electrodes placed in the same areas presented the more common pattern consisting of an increasing performance when moving from the first pooling layer to the second fully-connected layer.

By following the same qualitative criterion used for **D1**, we can say that 19 out of 22 electrodes ($\approx$ 86%) showed a satisfying linear mapping performance in **D2**. Note that a post-processing was implemented in **D2** because of the low number of samples (i.e. 25). In fact it was highly likely to randomly select a "bad" fold, which then could have hidden the overall average performance. The interquartile range (IQR), which is defined as the difference between the 3^{rd} and the 1^{st} quartiles ($Q3 - Q1$), was

computed and the folds giving a performance being either > Q3 + 3×IQR or < Q1 - 3×IQR were discarded. The number of outliers removed ranged from 0 to 4, hence still allowing to compute a robust average over 16 folds at least.

When looking at the performance behavior throughout the different layers, the most common pattern in **D2** consisted of an inverted V shape, where the highest performance was reached in the 4^{th} pooling layer. This pattern was present in 10 electrodes ($\approx 45\%$). 5 of these electrodes ($\approx 42\%$) were located in the inferior temporal gyrus, 4 (50%) in the fusiform gyrus and 1 in the medial occipitotemporal gyrus. Examples of mapping performances in **D2** are presented in Fig.11b-11c. Electrode 9 from subj6 shows the common inverted V pattern, while electrode 53 from subj4 follows a decreasing pattern throughout the layers. Both electrodes were implanted in the inferior temporal gyrus.

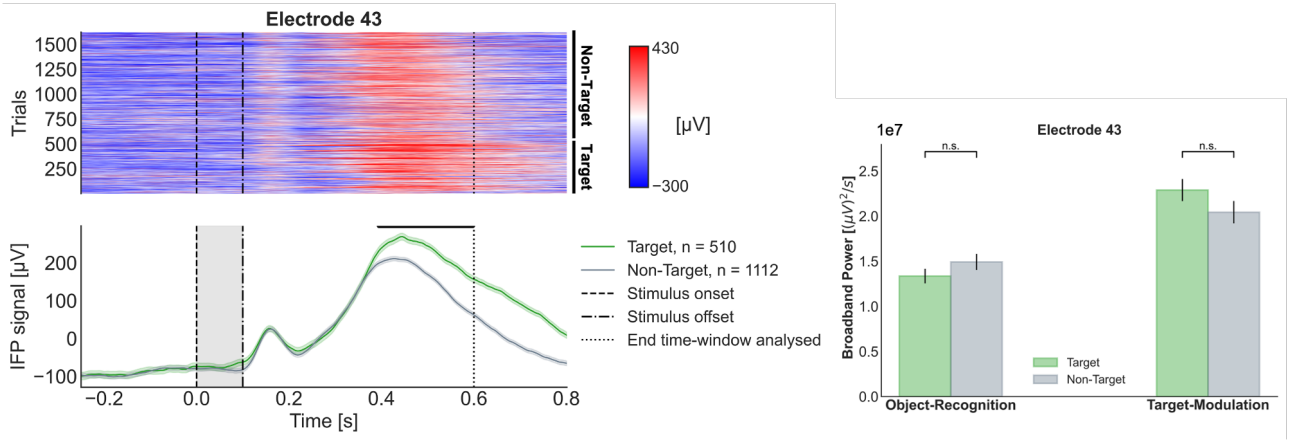
4.3 Target Modulation

In the frame of our project we were interested in the investigation and modeling of top-down modulation. In order to identify target-modulated electrodes, we repeated the procedure followed by Bansal and colleagues [34]. From their previous work, the hypothesis we wanted to validate was to get no significant difference between target and non-target trials in the object recognition phase and to find a significant difference in the target-modulation phase.

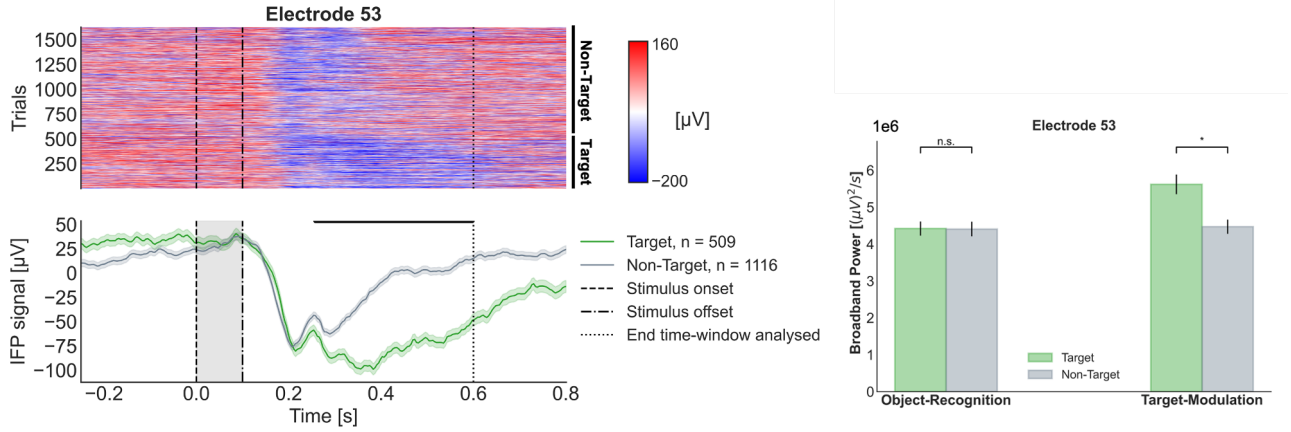
Out of the total 152 VR electrodes in **D2**, 49 electrodes ($\approx 32\%$) were target-modulated (found across 7 subjects). Among the TM electrodes, 20 ($\approx 40\%$) were also visually-selective. This means that $\approx 90\%$ of the visually-selective electrodes also showed target-modulation. 5 of these 49 electrodes belonged to subj10 and thus did not have a defined location. Out of the 44 located electrodes, 12 ($\approx$ 27%) were placed in the inferior temporal gyrus, 9 ($\approx 20\%$) in the fusiform gyrus and 4 ($\approx 9\%$) in the orbital gyrus.

In order to investigate if it was possible to account for the phenomenon of target-modulation with a scalar value, we computed the broadband power, which is defined in Eq. 2, for target and non-target trials and for both the object recognition (OR) phase and the target modulation (TM) phase. We defined the target modulation phase as being the period of time in which there was a significant difference between target and non-target trials. The object recognition phase was defined as the time interval going from 50 ms to the beginning of the TM phase. This means that the two phases could be of different lengths in time. However, note that the broadband power is an average over the duration of the time window considered, which then solves the confusion linked to this factor. We selected only those images having been presented both in the target and non-target cases. We then averaged over the images containing the same objects and corresponding to trials having the same target category. Take for instance an image containing a chair and a car. This image could be presented several times, sometimes being a target trial for the category *cars* and sometimes for the category *chairs*. The trials corresponding to these two different scenarios were averaged independently. In the non-target trials frame, the same image was then considered twice with two different target categories.

We then performed a t-test over the broadband power values obtained to compare target and non-target trials for both object-recognition and target-modulation phases. Out of the 49 TM electrodes, only 21 ($\approx 43\%$, across 6 subjects) passed this test ($p < 0.05$). Of these, 12 were also visually-selective ($\approx 57\%$). Subj10 accounted for 2 of these electrodes, meaning that these locations were unknown. Out of the 19 located electrodes, 8 ($\approx 42\%$) were implanted in the inferior temporal gyrus, 3 ($\approx 15\%$) were in the fusiform gyrus. None of the 4 TM electrodes located in the orbital gyrus passed this test. In Fig.12 we report an example of a TM electrode passing the additional test when "scalarizing" the target-modulation phenomenon and one example of a TM electrode not passing the test. In the same figure the raster plots for individual trials as well as the average IFP signals for both target and non-target conditions are reported.



(a)



(b)

Figure 12: Examples of electrodes showing target modulation: **a)** Electrode 43 from subj4, implanted in the fusiform gyrus and **b)** Electrode 53 from subj4, located in the inferior temporal gyrus. On the left the raster plots for all individual trials are shown together with the average IFP signal for both target and non-target conditions. The faded green and gray areas around the average IFP signal represent the standard error and the black bars on the top indicate statistical significance between the two signals ($p < 0.01$). In the raster plot trials were sorted in order to have all target and non-target trials together. On the right, the results from our "scalarization" method using the broadband power as a neural metric are shown. Error bars are computed across those images having been presented in both target and non-target conditions. * indicates statistical significance at $\alpha = 0.05$ in the t-test we performed between target and non-target conditions in the object recognition and target modulation phases.

4.4 Neurophysiological Analysis in D3

As previously mentioned, **D3** is an unpublished data-set. For this reason and in order to better characterize it, we detail here the neurophysiological results related to this data-set. Note that trials were equalized whenever two signals were compared in the frame of the current analysis.

4.4.1 Behavioral Analysis

In **D3**, subjects were presented with images of faces and had to perform an *Age* task, in which they had to discriminate between young vs old, and a *Gender* task, in which they had to distinguish between male vs female. Moreover, the task could be announced either before or after the stimulus presentation. These two conditions were named *Pre* and *Post*, respectively. The face images were artificially created via `FaceGen`.

All 11 subjects performed well in both tasks for both conditions. Fig.13 shows the averaged perfor-

mance across subjects for all four possible combinations. There was no significant difference between the performances achieved in the *Pre* vs *Post* conditions neither for task *Age* nor for task *Gender* (t-test, $p > 0.05$). However, a significant difference between *Age* and *Gender* (t-test, $p<0.001$) was found.

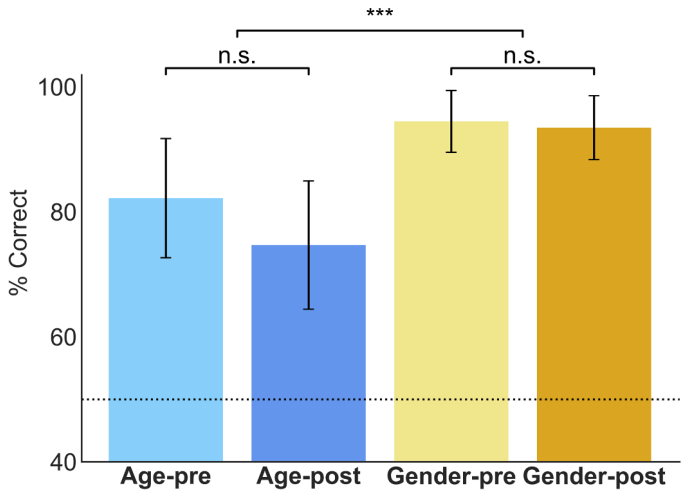


Figure 13: Behavioral performances for the four possible combinations of tasks and conditions in **D3**. Since in all cases it was a binary task, the chance level was 50% (dashed line). Error bars were computed over the performances of $n = 11$ subjects. t-tests were performed between the *Pre* and *Post* conditions independently for each task. A t-test between the performance in all *Age* trials vs the performance in all *Gender* trials was also done. n.s. indicates "not significant", while *** indicates a significant difference at the level $\alpha = 0.001$.

4.4.2 Visual Responsiveness Analysis

In order to find which electrodes were responsive to a visual stimulus, a similar permutation test to the one employed on **D2** was performed in **D3**. Moreover, in the case of **D3**, the minimal IFP_{range} in the time window [50,300] ms was required to be > 28.8 μV (see Section 3.2.1).

Out of 1201 electrodes, 70 electrodes ($\approx 6\%$) were considered visually-responsive ($p < 0.01$). Visually-responsive (VR) electrodes were found in 5 subjects out of 11. These electrodes spanned several brain regions, but areas within the temporal and occipital lobes were the most prevalent. 10 electrodes ($\approx$ 14%) were located in the fusiform gyrus, 10 ($\approx 14\%$) in the middle occipital gyrus, 9 ($\approx 13\%$) in the inferior temporal gyrus, 7 ($\approx 10\%$) in the angular gyrus and 5 ($\approx 7\%$)in the lateral occipitotemporal sulcus. An example of a VR electrode from subj4 implanted in the inferior temporal gyrus is shown in Fig.14. From the raster plot it appears that the IFP signal in the last trials (day 3 of recording) was weaker than in the first trials (day 1 of recording). A difference between the trials from day 1 and the ones in day 2 is also visible.

4.4.3 Delayed Responsiveness Analysis

Here we report our findings related to what we called delayed-responsive (DR) electrodes. In order to be considered DR, an electrode had to fail the permutation test assessing visual response and to pass a similar permutation test where the time window of interest was [415,665] ms vs a baseline time window [50,300] ms. Moreover, the IFP_{range} between 415 and 665 ms had to be higher than the noise threshold previously computed (see Section 3.2.1) and the IFP_{range} in the time window [300, 415] ms had to be lower than this same threshold. Note that 415 ms corresponds to the stimulus offset in **D3** experimental protocol. This analysis was performed on *Age* and *Gender* trials mixed,

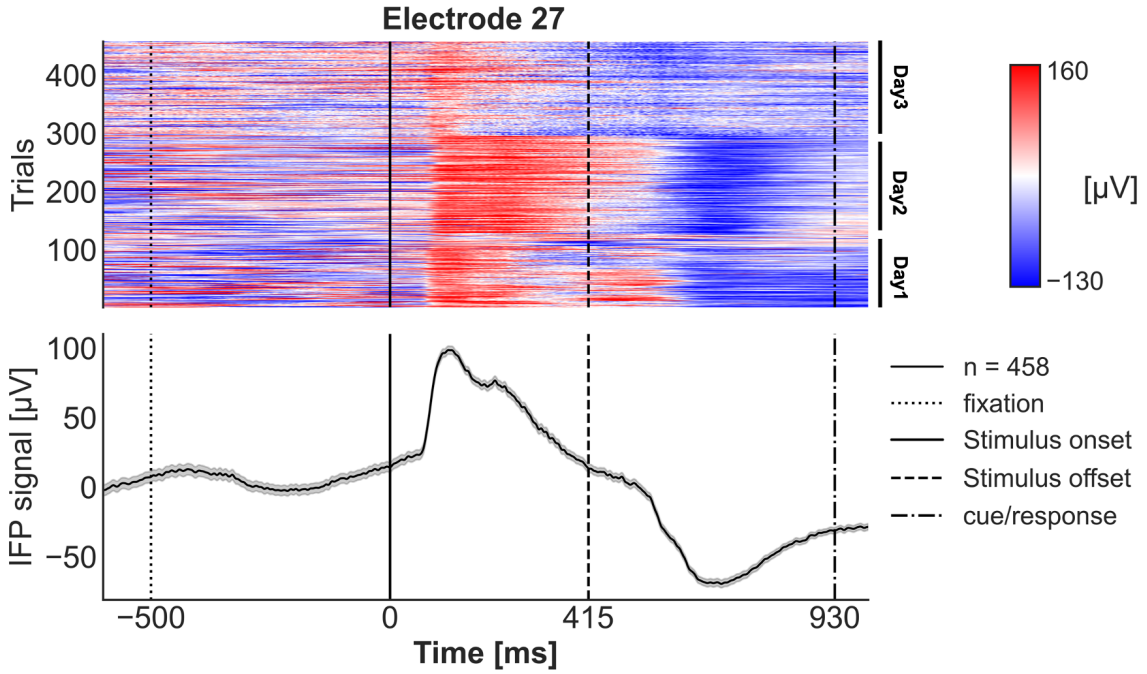


Figure 14: Example of a visually-responsive electrode from subj4 (**D3**). This electrode was implanted in the inferior temporal gyrus. On the top the raster plot shows the amplitude of the IFP signal for each trial. The trials belonging to different days of recording are indicated on the right. On the bottom the average IFP ± the standard error is presented.

but independently on *Pre* and *Post* trials, hence differentiating DR-*Pre* from DR-*Post* electrodes.

Out of 1131 electrodes (1201 total - 70 VR electrodes), 18 electrodes ($\approx 2\%$) were considered DR-*Post* (spread across 7 subjects, $p < 0.01$). 5 electrodes ($\approx 28\%$) were located in the inferior temporal gyrus, 3 ($\approx 17\%$) in the superior temporal gyrus and 3 ($\approx 17\%$) in the middle temporal gyrus. Only 5 electrodes ($\approx 0.5\%$) were considered DR-*Pre*. 4 out of 5 electrodes were located in the superior temporal gyrus. Only one electrode, which was located in the superior temporal gyrus, was considered both DR-*Post* and DR-*Pre*.

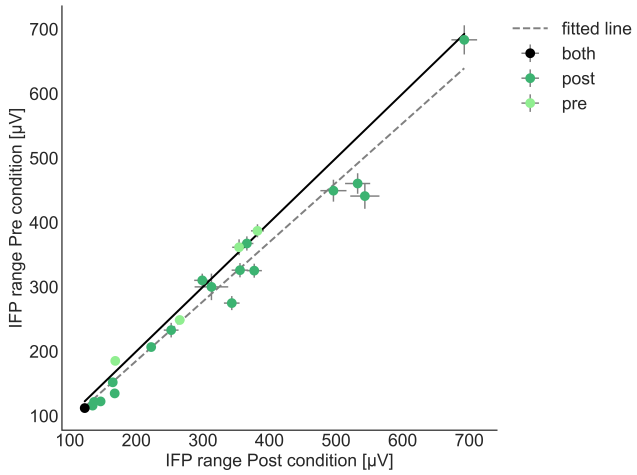


Figure 15: Comparison of average IFP_{range} of IFP signals for *Pre* vs *Post* condition for the DR electrodes identified. The dark green indicates DR-*Post* identified electrodes, while the light green stands for DR-*Pre* electrodes. A black point represents an electrode being considered as DR for both *Pre* and *Post* conditions. The continued line is the diagonal and the dashed line is the line fitted on our data. Error bars across trials are shown in gray. 22 points, corresponding to the total number of DR electrodes identified, are shown in the plot.

In Fig.15 we plotted the scatter plot of the average IFP_{range} of IFP signals for the *Post* vs *Pre* condition for the total number of DR electrodes identified. The Pearson correlation obtained when fitting a line to our data was 0.97.

4.4.4 Task and Task-Dependent Modulation

While in **D2** we were interested in the kind of top-down modulation related to the presence of the target category in the images the subjects were presented with, in **D3** we wanted to investigate the top-down modulation related to the presence (task modulated electrodes, TkM) and to the identity itself of the task (task-dependent modulated electrodes, TDM). In both cases a similar approach to the one used in **D2** to identify TM electrodes was used. However, note that in this case the FDR was set to be $< 0.1\%$. The time threshold, that is the required minimal length of the time interval in which the two signals were significantly different, was 93 ms. This analysis was performed on three different time windows independently: fixation 1 [-415,0] ms, stimulus presentation [0,415] ms and fixation 2 [415,830] ms. To be considered as a TkM or TDM, an electrode had to show a significant difference between the two signals longer than the time threshold for at least one of the three time windows. Also, the minimal difference between the two average signals was required to be $> 28.8\ \mu$V for at least one point (see Section 3.2.1).

Age-TkM: To find TkM electrodes, the *Pre* and *Post* conditions were compared within each task independently. Out of 1201 electrodes, there were 72 Age-TkM ($\approx 6\%$) electrodes. Age-TkM electrodes were found in 7 subjects out of 11. 12 electrodes ($\approx 17\%$) were located in the middle occipital gyrus, 7 ($\approx 10\%$) in the angular gyrus, 5 ($\approx 7\%$) in the inferior temporal gyrus, 5 in the left cerebral white matter, 4 ($\approx 6\%$) in the parahippocampal gyrus, 4 in the orbital gyrus and 4 in the lingual gyrus.

When looking at the different time windows, 33 electrodes ($\approx 46\%$) were Age-TkM during fixation 1, 29 ($\approx 40\%$) during stimulus presentation and 27 ($\approx 38\%$) during fixation 2. Out of the 33 electrodes found in fixation 1, 10 ($\approx 30\%$) were in the middle occipital gyrus and 6 ($\approx 18\%$) were in the angular gyrus. Regarding the electrodes found to be Age-TkM during stimulus presentation, 6 ($\approx 21\%$) were in the middle occipital gyrus, 4 ($\approx 14\%$) in the lingual gyrus and 4 in the angular gyrus. For fixation 2, 4 ($\approx 15\%$) electrodes were found in the orbital gyrus, 3 ($\approx 11\%$) in the opercular gyrus, 3 in the frontomarginal gyrus and sulcus and 3 in the inferior temporal gyrus.

Gender-TkM: Out of 1201 electrodes, there were 99 Gender-TkM ($\approx 8\%$) electrodes. Gender-TkM electrodes were found in 7 subjects out of 11. The subjects where TkM electrodes were not found were the same as for Age-TkM electrodes. 15 electrodes ($\approx 15\%$) were implanted in the middle occipital gyrus, 14 ($\approx 14\%$) in the inferior temporal gyrus, 9 ($\approx 9\%$) in the temporal pole, 9 in the angular gyrus and 8 ($\approx 8\%$) in the middle temporal gyrus.

When considering the different time windows, we found 46 ($\approx 46\%$) electrodes being Gender-TkM during fixation 1, 45 ($\approx 45\%$) during stimulus presentation and 67 ($\approx 68\%$) during fixation 2. Among the 46 electrodes significant during fixation 1, 12 ($\approx 26\%$) were located in the middle occipital gyrus, 5 ($\approx 11\%$) in temporal pole, 5 in the lingual gyrus, 5 in the angular gyrus and 5 in the inferior temporal gyrus. Regarding the stimulus presentation time window, 13 electrodes ($\approx 29\%$) were in the middle occipital gyrus, 8 ($\approx 18\%$) in the angular gyrus, 4 ($\approx 9\%$) in the lingual gyrus and 4 in the inferior temporal gyrus. Finally, out of the 67 electrodes identified as Gender-TkM during fixation 2, 13 ($\approx 19\%$) were located in middle occipital gyrus, 9 ($\approx 13\%$) in the inferior temporal gyrus, 7 ($\approx 10\%$) in the angular gyrus, 7 in the middle temporal gyrus.

TDM: TDM electrodes were found by comparing *Age* vs *Gender* trials for the *Pre* condition. Out of 1201 total electrodes, 24 ($\approx 2\%$) were considered task-dependent modulated electrodes. TDM elec-

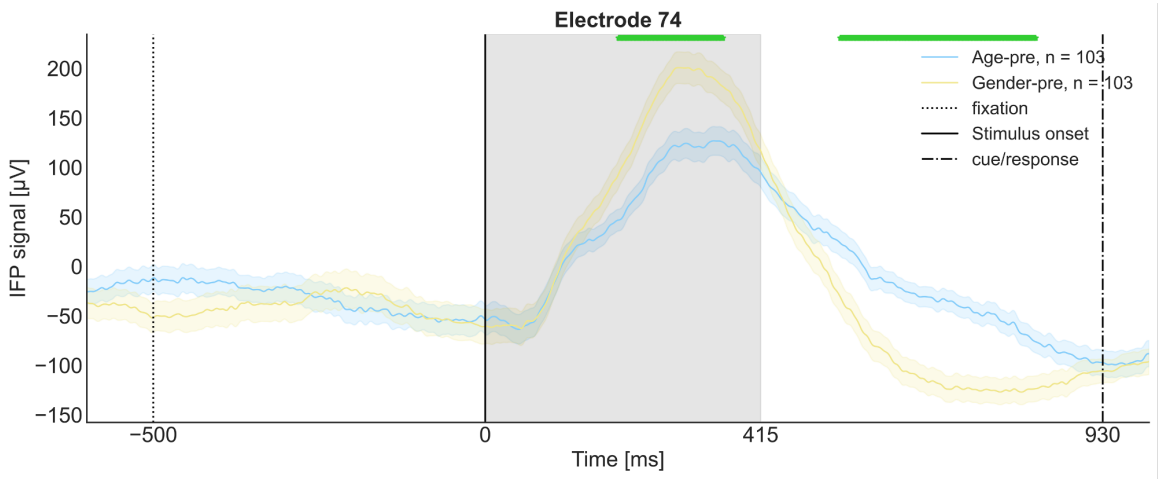


Figure 16: Example of an electrode being identified as task-dependent modulated in **D3**. The average signals ± the standard error are shown for both *Age-Pre* and *Gender-Pre* trials. This electrode was located in the angular gyrus. The green bar indicates a significant difference. This examples shows modulation during both stimulus presentation and fixation 2.

trodes were found in 3 subjects out of 11, with 22 electrodes belonging to subj4. 7 electrodes ($\approx 29\%$) were placed in the angular gyrus and 5 ($\approx 21\%$) were in the middle occipital gyrus.

In particular, 5 electrodes ($\approx 21\%$) were TDM during fixation 1, 6 (25%) during stimulus presentation and 18 (75%) showed a modulation during fixation 2. The 5 electrodes significant in fixation 1 were spread across different areas, including the lingual gyrus, the angular gyrus and the inferior temporal gyrus. Out of the 6 electrodes found to be TDM in the stimulus presentation time window, 3 (50%) were in the angular gyrus. Regarding fixation 2, 7 electrodes ($\approx 39\%$) were located in the angular gyrus and 5 ($\approx 28\%$) in the middle occipital gyrus.

It is worth mentioning that this analysis aiming at finding TDM electrodes was conducted also on *Age-* and *Gender-Post* trials as a negative control. Only 1 electrode in subj4 was found to be TDM in the *Post* condition. Finally, Fig.16 shows an example of an electrode in subj4 identified as task-dependent modulated (in the *Pre* condition) during both the stimulus presentation and the fixation 2 time windows.

4.5 Top-Down modulated model

As previously mentioned, the ventral visual pathway is commonly simplified to a feed-forward only stream. However, this is not true and feed-back in the brain is abundant. In particular, feed-back connections are thought to be important for goal-directed behaviors and attentional modulation [30, 31] as well as when performing more difficult object recognition tasks including for example occluded objects [32, 33].

In the frame of this project we focused on the implementation of a computational model containing also top-down modulating signals. Our approach involved a top-down modulation induced from the prediction layer and sent to the next to last layer in VGG19, that is the second fully connected layer. From there the top-down modulation could also travel to previous layers by propagating backward as formulated in Eq.8. The object of our interest is the target modulation occurring in **D2** during the target modulation phase and the task-dependent modulation observed in **D3** when comparing *Age* and *Gender* trials in the *Pre* condition (TDM electrodes).

Our modulation implementation consists in sending back the information related to the target category (for **D2**) and the information related to the task category (for **D3**). Conceptually speaking, this means

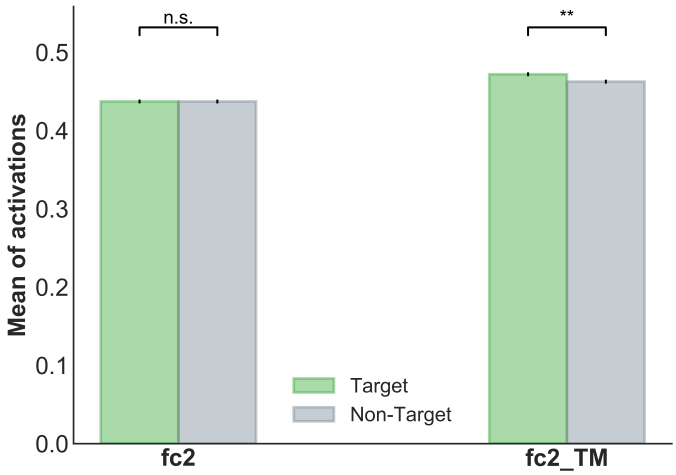


Figure 17: Results obtained from our top-down modulated model when fed with the same images used for subj4 for the barplots in Fig.12. Every object was mapped to a single unit in the prediction layer. Error bars are computed across images. ** represents a significant difference at the level $\alpha = 0.01$ (t-test).

that we gave the prediction layer a role similar to the one the prefrontal cortex is believed to play in triggering top-down modulation in our brain. In fact, the information about the target or task category identity is integrated as an input (as it was in the case of the experimental protocols) at the level of the prediction layer. An overview of the mechanism of our model is shown in Fig.7.

4.5.1 Target-Modulated model

What we aimed at qualitatively reproducing here was a similar pattern to the one observed when comparing the broadband power between the target and non-target conditions during the object-recognition and the target-modulation phases in Fig.12.

As it was explained in Section 3.4.3, in order to send back information specific to the target category, new categories needed to be redefined within the prediction layer of VGG19, which accounts for 1000 categories. 25 single-object images containing each a different object were used to perform the mapping between our categories and the units in VGG19's original prediction layer. What we observed is that the object recognition capabilities of VGG19 were limited, since almost no object was classified with a category similar to the one assigned in the experimental protocol. It is in fact important to remember that ImageNet is a large database and there are no huge general categories such as *chairs*, *faces*, *animals*, *houses* and *cars* in it. Also, some objects shared the same category as the most highly activated one. We then decided to force the mapping between our categories and the units in VGG19's prediction layer to be exclusive, meaning that no object could share the same units. This was done in order to avoid overlap between units assigned to different categories. In this way we also ensured that the amount of computation performed for each category was the same (and not for instance that category *chairs* was linked to 5 units while category *cars* to only 3).

All layers in VGG19 except for the prediction layer, which is coupled to $softmax()$, use as activation function $ReLU()$. In order to keep the activation values in the prediction layer similar in magnitude to the activations functions of the other fully-connected layers, we decided to change the activation function in the prediction layer from $softmax()$ to $ReLU()$. In Fig.17 the results obtained from our model are shown for the modulation applied on layer FC2. FC2_TM stands for fully-connected 2 target-modulated and refers to the values extracted after modulation (so at $t = 2$ in Eq.8). In this

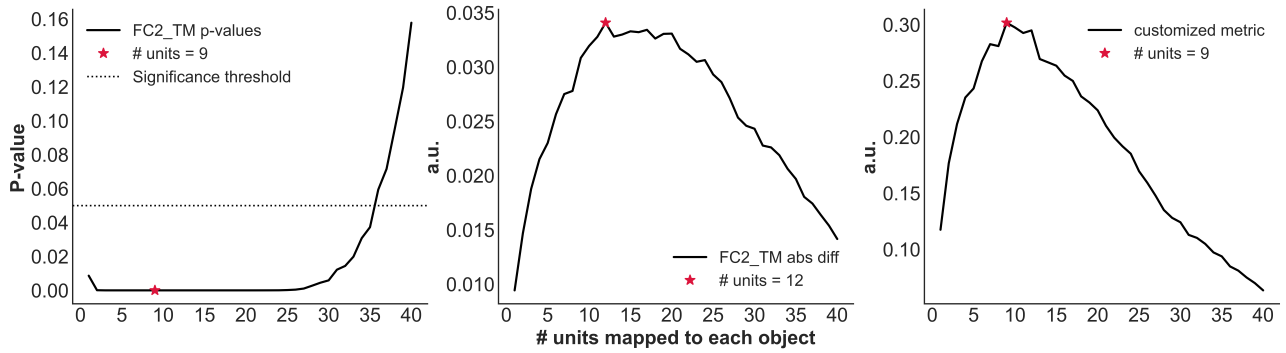


Figure 18: The results for the optimization of *N_units* for subj4. On the left the p-value of the t-test performed between target and non-target trials in the feed-back step is shown. In the middle we report the absolute difference between the average target and the average non-target activations. Finally, on the right our customized metric, as defined in Eq.9 is plotted. The red star indicates the minimum or the maximum of the curves.

case every object was mapped to a single unit in the prediction layer, meaning that each category was represented by 5 units. We fed the model with different sets of images corresponding to the different images the subjects were presented with. Only subjects having shown target-modulation in the neurophysiological data and having passed the t-test on the broadband power (see Section 4.3) were considered. In order to obtain a single scalar from the activations extracted for each picture, the activation vectors were averaged across units. In the feed-forward step the two average activations for target and non-target are identical as expected. However, a statistical difference ($p < 0.01$) was found for the feed-back step.

In order to investigate the effect the number of units associated to each object had, we performed the same analysis by increasing subsequently *N_units* from 1 to 40, which represents the highest number of units possible per object (if considering an exclusive mapping). In the frame of this analysis, we reported three different measures: the p-value of the t-test performed between the target and the non-target conditions for FC2_TM, the absolute difference between the mean target and non-target activations and the customized metric in Eq.9.

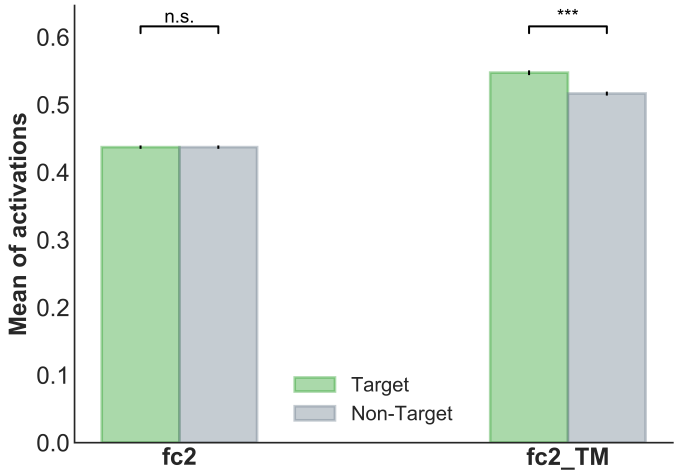


Figure 19: Results obtained from our top-down modulated model when fed with the same images used for subj4 for the barplots in Fig.12. Every object was mapped to 9 units in the prediction layer. Error bars are computed across images. *** represents a significant difference at the level $\alpha = 0.001$ (t-test).

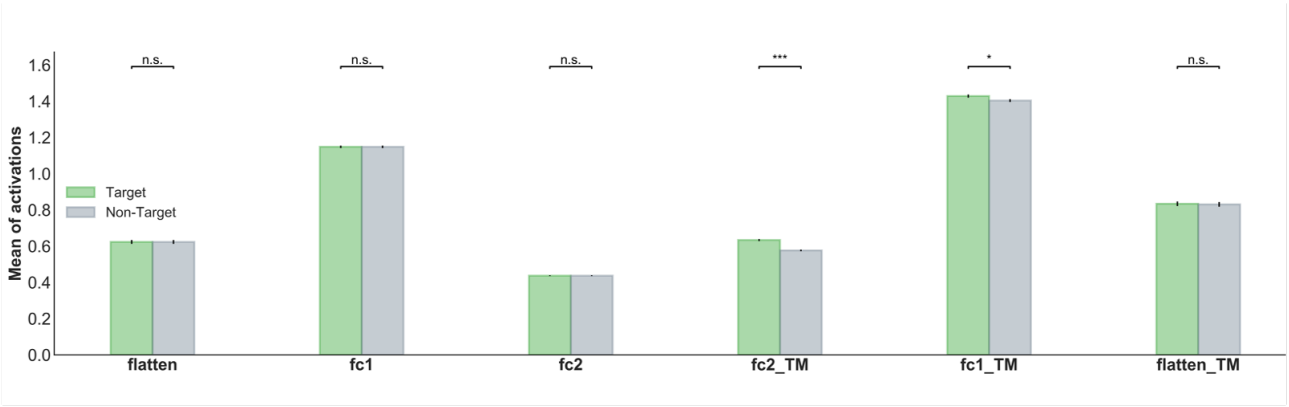


Figure 20: Results obtained from our top-down modulated model when fed with the same images used for subj4 for the barplots in Fig.12 and when going backward until the flatten layer following Eq.8. Every object was mapped to 9 units in the prediction layer. Error bars are computed across images. * and *** represent a significant difference at the level $\alpha = 0.05$ and $\alpha = 0.001$, respectively (t-test). A scaling factor of $2\times$ was applied to the prediction layer's activations.

The results for subj4 are reported in Fig.18. As we expected the three metrics all present a slight U shape (inverted in the case of the center and right plots). This clearly shows the impact of the *N_units* parameter. Regarding the leftmost plot, that is the one showing the p-value, a dashed line representing the common significance level at α=0.05 is shown. While *N_units* seems indeed to affect the effect of top-down modulation via the difference between the target and non-target conditions, the effect over the significance is not dramatic.

We finally decided to use our customized metric (Eq.9) to select the optimal number of units. By majority vote over the 6 subjects considered, *N_units* = 9 was chosen. The other optimal *N_units* were 11 and 12, thus showing robustness around this number of units. The same results as the ones in Fig.17 are shown in Fig.19, with the only exception of having 9 units associated to each object for a total of 45 total units per category. We can see how the modulation effect appears here more evident.

Finally, we investigated how the modulation propagated more backward in previous layers. For sake of simplicity we analysed the effect only on the two fully-connected layers and the flatten layer, which is the intermediate layer linking the convolutional part to the fully-connected part of the model. The modulation itself does not receive further input in more backward layers and it is only propagated as shown in Eq.8. In order to better show our results, a scaling factor $2\times$ was applied to the prediction layer's activations before sending the feed-back in order to increase the magnitude of the top-down modulation. These results are reported in Fig.20. It is interesting to remark how the difference between the target and non-target conditions decreases as we go more backward in the model, until not being significant anymore at the level of the flatten layer. Note that by increasing the scaling factor further it would be possible to obtain a significant difference also in the flatten layer.

4.5.2 Task-Modulated model

In the case of **D3**, we were interested in modeling the modulation depending on the kind of task the subjects were performing, that is the modulation observed in the TDM electrodes (see Section 4.4.4).

In order to identify the units in the VGG19's prediction layer associated to task *Age* and task *Gender*, we followed a different approach with respect to the one used for **D2**. In fact, here it was not possible to divide the images in two subsets representing a task each. The images used for the two tasks were exactly the same and tasks *Age* and *Gender* are not mutually exclusive as for example *chairs* vs *faces* are. Thus, we fed the VGG19 model having the $ReLU()$ activation function in the prediction layer with all the images contained in **D3**. We then considered each unit in the prediction as a feature and

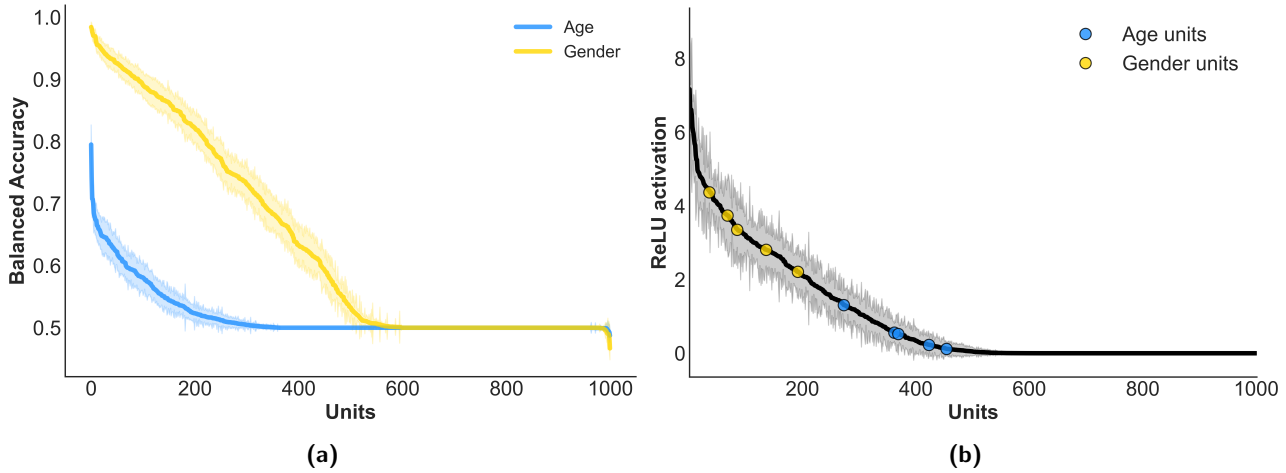


Figure 21: Category-mapping for tasks *Age* and *Gender* in **D3**. **a)** Balanced accuracy performance for all units for both tasks *Age* and *Gender*. The shaded area represents the standard deviation computed over 20 folds. Units are sorted in decreasing order according to the mean balanced accuracy. **b)** ReLu activations sorted in decreasing order for all units in VGG19's prediction layer. Standard deviation is computed across all images (no train-test split). The colored markers indicate the units selected to represent the tasks: light blue for *Age* and yellow for *Gender*.

we computed its discriminant power in performing the binary classification that also subjects were asked to perform during the experiment, that is young vs old and male vs female. The procedure was performed 20 folds with a train-test split of 90-10%. The balanced accuracies (BA) for all units and for the two different tasks were then sorted and the best units were selected to represent the task.

We decided to arbitrarily impose a threshold requiring the minimal balanced accuracy to be ≥ 0.7. We also wanted to have the same number of units representing the two tasks and thus, these two criteria led us to select 5 units for each task. For sake of completeness, it is worth mentioning that the best unit for task *Age* gave a BA of 0.79 ± 0.03, while the best unit for task *Gender* gave a BA of 0.98 ± 0.01. This means that the features extracted in the prediction layer of VGG19 predicted better *Gender* than *Age* given the images of **D3**, but, most importantly, this means that VGG19 shows the capability of adapting to new tasks it was not trained on. Classification performances are shown in Fig.21a. No units were shared between the two tasks. Also, in Fig.21b the $ReLU()$ activations for all units in the prediction layers are shown and the 5 units for each task are indicated in order to give

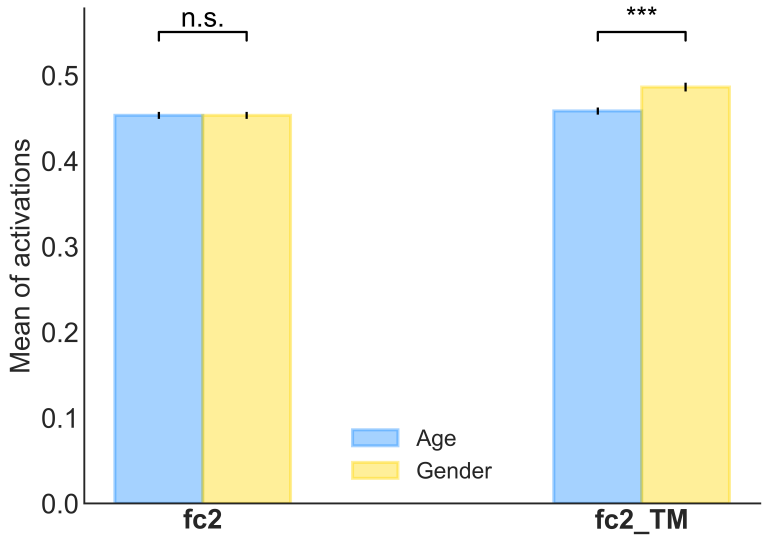


Figure 22: Results obtained from our task-modulated model. Each task was mapped to 5 units in VGG19's prediction layer. Error bars are computed across images. *** represents a significant difference at the level $\alpha = 0.001$ (t-test).

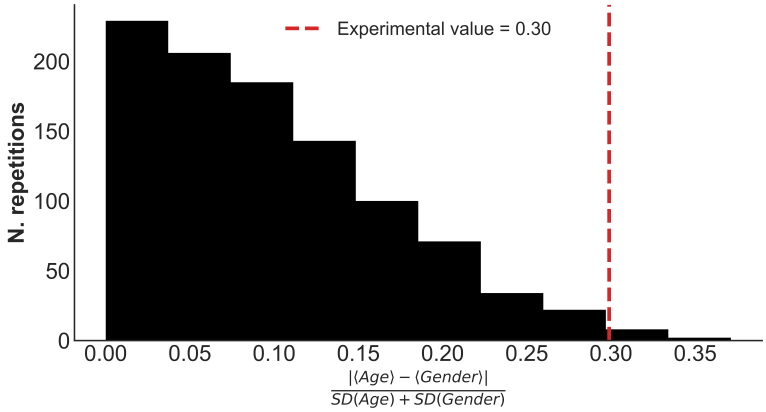


Figure 23: Results from the randomization test we performed in order to assess the importance of the mapping identified between tasks and units in VGG19's prediction layer. The red bar represents the experimental value. p-value < 0.01.

a sense of the actual amplitude of the activations. We can see that *Gender* units were consistently showing higher activations with respect to *Age* units.

We then used the same kind of architecture and feeding paradigm as the one used for the target-modulated model in **D2**. We fed the model with the same images the subjects were presented with and for each image we sent back once the information related to *Age* and once the information related to *Gender*. The results are shown in Fig.22.

In order to rule out the possibility that the modeled difference (modulation) between *Age* and *Gender* was simply due to the selection of two separate subsets of units within the prediction layer, we performed a randomization test. We repeated 1000 times the same procedure followed to obtain the results in Fig.22 but by randomly choosing the units to map to the two tasks (though ensuring no overlap was present). For each repetition we computed a similar metric to the one showed in Eq.9, where we replaced $Target$ and $Non - Target$ by *Age* and *Gender*. We compared the experimental metric found to the null distribution. The value obtained with the real mapping between the two tasks in the feed-back step resulted significant, p = < 0.01 (see Fig.23).

Finally, also in the case of **D3**, we were interested in checking what would happen when the modulation is let propagate more backward through the first fully-connected layer and the flatten layer. A similar result to what found in **D2** was obtained, as shown in Fig.24. However, note that in order to obtain a significant difference in the feed-back step for layer fc1, a scaling of $10\times$ had to be applied on the prediction layer's activations prior to sending back the modulation.

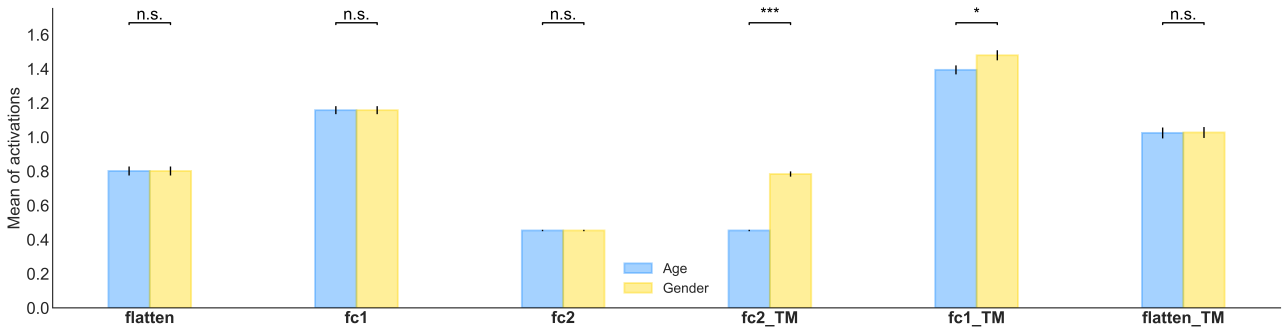


Figure 24: Results obtained from our task-modulated model. Each task was mapped to 5 units in VGG19's prediction layer. Error bars are computed across images. * and *** represent a significant difference at the level $\alpha = 0.05$ and $\alpha = 0.001$, respectively. A scaling factor of $10\times$ was applied to the prediction layer's activations.

5 Discussion

5.1 VGG19 can predict neural responses

One important part of the current project was the investigation of visual selectivity in neural populations recorded via intracranial electrodes. By pushing the analysis even further, it was possible to build a linear regression model capable of predicting neural metrics from computational features extracted by a feed-forward model like VGG19.

Even if the methods used for the identification of electrodes selective to a specific category (the electrodes called category-selective and visually-selective in **D1** and **D2**, respectively) varied a bit between the two data-sets, when considered together, 40 electrodes out of a total of 1632 electrodes were selective to a specific category. Out of these 40, 25 ($\approx 1.5\%$ of the total) were selective to the category *faces*. This percentage is in accordance with what found by Grossman and colleagues [45]. Moreover, the fact that *faces*-selective electrodes were the most abundant ($\approx 63\%$ of neurons showing selectivity) was expected, as it is supported by previous neurophysiological studies [53, 54]. Regarding the locations of visually selective electrodes, these were located mostly in the occipital and temporal lobes. In particular, the inferior temporal gyrus, the medial occipitotemporal gyrus and the fusiform gyrus were among the most represented areas, which is consistent with the structure of and the role played by the ventral visual pathways.

Both **D1** and **D2** were used to perform the Linear Mapping Analysis. Our aim was to link deep convolutional neural networks to neurophysiological data by trying to explain some of the variance recorded in neural data through a computational model. Some works on human data [45, 46] have already shown possible correlations between computational activations and neural data. However, our methodology resembles the one employed in [43, 44], even if both these works were conducted on non-human primate data. The results obtained through our linear regression analysis were overall satisfactory, meaning that they pointed towards the conclusion that it is indeed possible to regress neural data starting from computational activations. A thorough and detailed quantification of the results was difficult to perform because of the high variability across electrodes and across subjects. One possibility could have been to average the performance of electrodes located in the same areas and in the same subject. However, this approach is also debatable, because electrodes generally located in the same areas often show very different activation patterns.

By following the qualitative approach explained above, all electrodes in **D1** and $\approx 86\%$ of the electrodes in **D2** yielded a good performance. However, consistent negative R^2 scores were not rare. This result could be explained by the small size of our data-sets. Not only the number of subjects per data-set was not high, but also the number of repetitions for each image was not high and not balanced across images. The linear mapping analysis in **D1** was performed over 125 samples, each one corresponding to a different image. The number of repetitions for a specific image was higher in **D1** with respect to **D2**. This is also due to the nature itself of **D2**, since most images contained two objects, which made the likelihood of repeating the same image quite low. This is the reason why an object-based approach (rather than an image-based one) was preferred in **D2**. However, while the robustness in the estimation of a neural metric for a certain object increased, the final number of samples was reduced to 25. The basic hypothesis in model validation is that, when a huge number of trials (samples) is present, the mean features for the train and for the test sets should be similar because the samples should conceptually belong to the same population (following the law of large numbers). This is not the case with a population of just 25 samples and it is likely that when split into train and tests sets with a 80-20% ratio, the average of the features of the 20 training samples is not the same as the one computed over the 5 test samples. Very "bad" folds were thus discarded with our outliers procedure implemented in **D2**. Also, it is worth saying that it is not clear how the presence of two objects in the images could affect the activation maps extracted from VGG19's layers.

Moreover, besides the evident limitations of our methods and data, the simple possibility that some layers do not extract features suitable for the prediction of neural data should not be ruled out.

When looking at the comparison between averaging methods performed in **D1**, our result, i.e. the general worse performance of the spatial-based approach, is consistent with what found by O'Connell [44]. This trend can be expected: the averaging over the channels dimension actually cancels out the work of feature extraction performed by VGG19 through the convolutional layers. This means that when averaging a lot of precious information is lost, which ultimately leads to a worse performance when trying to predict neural data. On the other hand the PCA-only and feature-based methods gave similar results. In the feature-based approach, the averaging is performed over the spatial dimensions of the activation maps, hence maintaining the important features extracted by the different convolutional filters. However, it is important to note that in our images there was no real information linked to the spatial dimension, because objects were consistently located either in the center (**D1**) or on the top and/or bottom of the image (**D2**). Thus, this could explain the similarity between the PCA-only and feature-based approaches.

Finally, an interesting aspect to consider when discussing the linear mapping analysis is the trend observed when comparing the performance throughout the different layers in VGG19. Previous work [43] has suggested that the location of the electrodes correlates with the position of the layer in the model giving the best performance. For example, an electrode located in V4 will be better predicted by the features extracted in an intermediate layer, while the variance in neural data recorded by an electrode located in the IT will be better explained by the last layers in the model, i.e. the highest ones in term of hierarchy. This theory is very appealing and it is also based on the correlation existing between the feature processing in the ventral visual pathway in the brain and in convolutional neural networks, with low-level features being extracted in the first areas/layers (V1-V2/conv1-conv2-conv3) and more high-level complex feature are computed by the last regions/layers (IT/fully-connected layers). This theory was hard to validate and to investigate in our data, because of the low number of electrodes analysed and because in **D2** almost all electrodes were located either in the inferior temporal gyrus or in the fusiform area, that is two regions situated high in the visual hierarchy. However, a common pattern consisting in the 4^{th} pooling layer showing the best performance was found. This seems to contradict a bit the results found by Yamins and colleagues, since this layer would be located intermediately and the electrodes would be placed in high visual areas. On the other hand the results found in **D1** correlate better with the Yamins' hypothesis: most of the electrodes showing an increasing performance through the layers were located in the temporal lobe and the few electrodes showing a decreasing pattern were rather located more posteriorly. Interestingly, these electrodes were also the ones showing a spatial-based approach performance comparable to the feature-based and PCA-only ones. We could hypothesize that this is the case because electrodes positioned in lower visual areas record data from neuronal populations extracting simpler features. These would then correlate with the initial convolutional layers and with basic spatial information, which can be encoded also in the spatial-based averaging approach.

However, all of these are conjectures and a deeper analysis should be performed. Of course, the amount of data and the number of different subjects from which data are recorded are a crucial factor in such an analysis. The more visually selective electrodes are identified and the more visual areas concerned, the easier it would be to formulate a clear theory. It should be remembered though that it is highly unlikely that a certain region in the visual pathways can be ascribed to a specific layer in a deep neural network. In fact the architectures of computational models having proved to be efficient in predicting neural data vary a lot. If a proper correspondence between layers and brain regions does not look realistic, we could still think of dividing both the visual and the model hierarchies in an initial, an intermediate and a final section, hence facilitating the quantification of the phenomenon. Finally, more neural codes should be investigated. In our case we focused on the IFP_{range} as a preferred

neural metric by taking inspiration from the work of Liu [51] and Bansal [34], but other possibilities should be explored, such as the broadband power or the power in one specific frequency band (think of alpha, beta and gamma for instance).

5.2 Target-Modulation is hardly summarized in a scalar

Data-set 2 was at the origin of our interest in studying and modeling top-down modulation. Following the work and the results of Bansal and colleagues [34], we tried to reproduce some of their methods in order to select target-modulated electrodes.

The main difference with respect to their methods is that we conduced the target-modulation analysis only on visually-responsive electrodes. The inferior temporal gyrus and the fusiform gyrus were the regions containing most of the TM electrodes we identified. Interestingly, few electrodes were also found to be target-modulated (and thus visually-responsive) in the orbital gyrus, a region not commonly ascribed to visual processing. Interestingly, $\approx 90\%$ of the visually-selective electrodes identified in **D2** showed also target-modulation. This could make us think that rather than being two separate subsets of neurons, visually-selective neurons represent a subset within the target-modulated ones, where the regions involved in object recognition receive some sort of feedback.

Also, similarly to their results, both cases where the IFP signal in target trials during the modulation phase was higher than the IFP signal in non-target trials and viceversa were found. However, in order to later have a representation to compare to the results obtained from our model, our analysis went further and we tried to summarize the target-modulation phenomenon into a scalar value. In order to do so, we used the broadband power as a neural metric. This choice was preferred to the IFP_{range} since the latter would have been affected by the different time window length when dividing the signal in object-recognition (OR) phase and in target-modulation (TM) phase. In fact, we decided to perform such a division in a data-driven fashion, that is by choosing as target-modulation phase the time interval resulted to be significantly different when comparing the target to the non-target condition. The object-recognition phase was then defined as the time interval starting from 50 ms until the beginning of the modulation phase. The 50 ms was fixed in order to account for the latency of visual responses as mentioned in [51]. We then evaluated our model by performing a t-test measuring the difference between the target and the non-target conditions both for the OR and TM phases, where the former was supposed to be not significant and the latter significant. Only $\approx 46\%$ of the TM electrodes passed this test. This showed us that it is difficult to account for a complex phenomenon such as top-down modulation via a simple scalar. This is likely to be due to the loss of temporal dynamics. For example, in the case where the modulation happens around 0 μV, even if the difference between target and non-target trials is large, it is possible that the total broadband power will be similar because of the squaring component of this metric. Thus, if a scalar approach was to be pursued, a better neural feature would need to be engineered, perhaps from the combination of more common metrics.

5.3 Cognitive phenomena might underlie task-modulation

As mentioned several times, **D3** represents a new unpublished data-set. The images presented to subjects were in this case belonging to only one category, and thus the visually-selectivity and linear mapping analyses were not performed.

Visually-responsive electrodes were found in only 5 subjects out of 11. The locations of the electrodes were consistent with the knowledge we have about the visual pathways and were generally in accordance with the locations of VR electrodes found in **D2**. It is difficult to explain the difference in the overall percentage of VR-electrodes found in **D3** with respect to **D2**, especially if we consider that the method used was the same. While it is true that in the case of **D3** we also computed a noise

threshold in order to discard those electrodes showing a weak neural response (the IFP_{range}), we saw that only 6 electrodes were discarded because of this condition. Subjects variability probably is the true reason behind the difference observed between the two data-sets. Moreover, another potential factor to take into account is what was observed in the raster plot in Fig.14. The last trials showing a washed-out signal actually correspond to the trials recorded on the third and last day of recording in subj4. It is not possible to assess what happened, but it is likely to be a problem related to the electrodes themselves. The main reasons of signal degradation are usually mechanical failure and the increase of impedance in the electrodes due to the foreign body response. However, this option does not seem plausible in our case, since these intracranial electrodes should usually be implanted just for the period of time necessary to identify the seizure foci in the patient. Moreover, the first glial immune reaction occurring during the first days after implantation is not believed to affect significantly electrode's ability to record [55]. More details about the timeline of the implantation and the following experiments should be obtained in order to consider the different possibilities. An analysis per day was thus initially considered. However, dividing the data per day would have led to a huge loss of data and to a lower statistical power. Moreover, even if the IFP signal's amplitude is lower, the general pattern of increasing and decreasing voltage often appeared to stay consistent (as it is visible in Fig.14). Finally, it was not possible to compare the amplitude of the signals recorded across days since there was no overlap in the images used. For all these reasons we decided to move on with all data at once, even if it remains unclear to what extent this large difference in amplitude affects our statistical methods.

Very few electrodes were identified as delayed-responsive (DR), that is electrodes showing a remarkable neural response in a late time-window ([415, 665] ms), but none during stimulus presentation. This analysis was performed on *Age* and *Gender* trials mixed but on *Pre* and *Post* conditions independently. While we initially thought that the knowledge of the task (thus the *Pre* condition) could increase the IFP response signal, our results actually point in the opposite direction. The DR phenomenon resulted more prevalent in the *Post* condition than in the *Pre* condition. Also, when plotting the IFP response for *Post* vs *Pre* condition for all the DR electrodes identified, the points were located consistently slightly below the diagonal, which indicate a larger response in DR-*Post* electrodes than in DR-*Pre* electrodes. This could be interpreted as the knowledge of the task actually decreasing the IFP response, perhaps by inhibiting an ongoing cognitive phenomenon. We could speculate that when the task is not known, the subject has to focus more on a wider range of features in the image presented, hence triggering some attentional modulation or some working-memory circuits. However, it is also interesting to note that the locations of the DR electrodes correlate more with areas commonly ascribed to visual processing. It is then still an open question to understand why these electrodes did not show any visual response during stimulus presentation.

Still regarding modulation, a similar analysis to the one involving target-modulation in **D2** was performed in **D3**. In this case we distinguished task-modulated electrodes (called TkM) and task-dependent modulated electrodes (called TDM). Most of the electrodes identified as TkM or TDM were located in areas involved in visual processing. Interestingly, TkM electrodes in the angular gyrus (AG) were also consistently found .The AG has been associated to a wide pool of functions, among which attentional modulation and word reading and comprehension [56], all aspects that could somehow explain the modulation we observed. For example, regarding TkM electrodes modulated in fixation 1, one could think that the modulation derives from the reading of the task in the *Pre* condition. More generally speaking, the finding that TkM electrodes showed modulation in a lot of electrodes during fixation 1 (thus even before stimulus presentation) points to the conclusion that this modulation is highly cognitive. In fact, fixation 1 is the time interval following the presentation of a fixation cross screen, which is identical for both *Pre* and *Post* cases. Thus, the modulation observed should not be associated to visual processing, but rather to a cognitive one. We can speculate that subjects start preparing for the task when they know it, maybe picturing in their mind the features they should

focus on.

On the other hand, TDM electrodes were found only in 3 subjects, with almost all electrodes coming from subj4. This represents the main drawback of our findings, since it is then difficult to understand if this neurophysiological result really generalizes across subjects. The lower number of TDM electrodes identified with respect to TkM electrodes should not surprise. In fact, it should be remembered that the phenomenon investigated here is more subtle than the one of modulation in the *Pre* condition vs the *Post* condition. Naturally, if the condition either on the minimal IFP_{range} (the noise threshold) required or on the amount of consecutive ms (the time threshold) was to be changed, the number of TDM electrode would increase. However, this would also lead to an increase in TDM-*Post* electrodes, which represented our negative control. In the case where the task was revealed after stimulus presentation (*Post* condition), no difference should be expected between *Age* and *Gender* trials. Electrodes showing a statistical difference for a prolonged period of time in the *Post* condition should be better investigated in order to find an explanation for the phenomenon.

5.4 Top-down modulation can be qualitatively modeled in DCNN models

In the present project we suggested a new way of implementing top-down modulation in computational models. In our specific case we worked on VGG19 and we modified its architecture to add a feed-back step.

Our simple approach has proven to generate top-down modulation in both **D2** and **D3**, meaning that the general principle can be adapted to the specific kind of modulation we seek to model. Also, it is worth reminding that our method involves changes from the architectural point of view and in the feeding paradigm of the model. No training was performed to obtain the results shown here. This not only makes the approach more appealing, but also allows us to avoid a computational heavy procedure and potential issues such as overfitting. Moreover, in order to send backward the information related to the target or task of interest, we used the same weight matrices as in the feed-forward counterparts by just transposing them. This means that two neurons were connected by the same synapses both in the feed-forward and in the feed-back steps, an aspect that might resemble more what really happens in the brain.

The fact that VGG19 was not able to correctly classify most of the single-object images could be due to the small size of the objects in the images and to the relatively poor resolution. For example, when using the same objects but shown in the middle of the image and in a bigger format, VGG19 was capable of recognizing accurately most of the *Animals*, *Chairs* and *Cars* objects in **D2**. In particular, the $softmax()$ activations were consistently higher. However, the fact that VGG19 did not classify efficiently our objects into ImageNet categories was not a problem. The whole purpose of the category-mapping procedure was indeed to build new customized categories capable of representing the categories our objects belonged to. In order to counteract the very low activations given by $softmax()$, we changed the activation function of VGG19's prediction layer to $ReLU()$. Notice this change does not affect the mapping results within a certain number of units. In fact, both $softmax()$ and $ReLU()$ are monotonically increasing functions, which means that the order will be preserved. This is not the case for a negative input, since $ReLU()$ will output simply 0, while $softmax()$ will always give a number in the range (0,1). However, the output of $softmax()$ will be extremely close to 0, which thus means that the units chosen will be almost identical in magnitude, no matter if their identity changes. This is notably true when considering the results given by the optimization procedure performed for *N_units*. The optimal number of units to link to an object was in the range [9,12], thus far from the number of units necessary to start involving units having 0 as activation value. The U (and inverted U) shapes of the curves reported in Fig.18 can be explained now from this point of view. At the beginning, the more units are added the more robust the result becomes and the stronger the modulation effect. However, after a certain critic point, random units start to be added to the different categories. When

the number of random units become large, the modulation effect between target and non-target is lost.

Because of the different nature of the images in **D3**, the approach followed to identify our new categories within VGG19's prediction layer had to be different. Thus, as a corollary result we found out that the some units extracted from VGG19's prediction layer actually contain enough information to perform a binary classification for both tasks *Age* and *Gender* with a performance well above chance. For some unknown reasons, *Gender* results were higher than *Age*. Again, this could be a data-set specific effect, meaning that perhaps the *Age* changes produced by `FaceGen` with our parameters were less apparent than the ones obtained for *Gender*. Also, from what is shown in Fig.21b, we see that on average the units selected to represent task *Gender* show a higher activation with respect to the units chosen for *Age*. This could point towards a "Gender bias" present in VGG19. The meaning of our chosen mapping between VGG19's units and the task categories was further validated by the permutation test presented in Fig.23.

Another interesting result found in both **D2** and **D3** is the decreasing pattern observed in the modulation while going backward to lower layers. This phenomenon could be thought to model the reverse hierarchy theory, that is the concept stating that perceptual learning is stronger in high-level visual areas (like IT) and decreases the more we reach low-level visual areas (like V1) [57]. This makes our model's behavior even more similar to what is believed to occur in the brain. It would be interesting to go back also to convolutional layers in order to investigate this phenomenon further. However, some limitations need to be considered: following our approach, it will be difficult to go back to convolutional layers because of the presence of maxpooling layers. Indeed, it will be impossible to mimic the feed-forward flow of information `conv` $\rightarrow$ `maxpool` in the feed-back step `conv` $\leftarrow$ `maxpool`. This step does not involve learned weights and thus the information from the convolutional layer to the maxpooling layer is lost. Since in our approach we are quantifying the result by computing the average over the activations, one solution could be to change the maxpooling layers to averagepooling layers. In this way the feed-back step could be modeled by simply performing an oversampling with the average value. However, such a change in the architecture of the original VGG19 model could potentially lead to the loss of its object recognition capabilities. Re-training or validation over the ImageNet data-set would then be required.

Our approach shows a simple method capable of reproducing qualitatively some important phenomena found in the brain. However, it is also very simplistic and a lot needs to be done in order to veritably reproduce neural signals when top-down modulated. As it was mentioned also in Section 5.2, the top-down modulation is a phenomenon which is difficult to "discretize" into a scalar (as also our model does). In order to reproduce a temporal dynamic through our model, recurrent connections should be added and a recurrent deep convolutional neural network implemented. The curve obtained from the time steps of the recurrency within the model could then be compared to the activity recorded in neural data in order to assess similarity. Recently, Nayebi and colleagues have worked exactly on this aspect [58, 59]. They have been showing that classical recurrent units (such as gating units or Long-Short-Term-Memory units) do not actually increase dramatically the performance of DCNNs when performing a challenging object recognition task and that more *ad hoc* recurrent cells (incorporating both *gating* and *bypassing* functions) have to be implemented. Thus, they built a ConvRNN containing recurrent connections within a layer (similar to horizontal connections in the brain) and long-range connections (similar to top-down modulation across areas). Finally, they linearly mapped the time-dependent output of the computational model to the time dynamics recorded in V4 and in IT in monkeys by showing very promising results [58, 59]. All these findings point strikingly towards the need of recurrency and feed-back connections in computational models to closely mimic the brain. The work of Nayebi and colleagues is inspiring and could represent the next step to undertake for our analysis on human data, even if the amount of data remains an obstacle to deal with.

6 Conclusion

In this project we investigated the link between neurophysiology and computational modeling by focusing on the sense of vision. In particular, our interest was in studying the two different but related phenomena of object-recognition and top-down modulation.

Several studies have shown that deep convolutional neural networks (DCNN) are capable of predicting neural responses recorded in the ventral visual pathways in monkeys. Results of this kind in humans are still scarce. Despite the limitations due in part to the nature of the data at our disposal, we showed here that a DCNN model such as VGG19 can indeed regress neural responses in humans in a satisfactory way. The results were not always robust, but they surely were encouraging and in accordance with previous findings. Also, it is still unclear whether a correlation between the hierarchy of visual regions along the ventral pathway and the hierarchy of layers in a DCNN exists.

In order to more exhaustively model neural data when it comes to more complex visual tasks, feedback components need to be taken into consideration. In the current work we suggested a simple but efficient way to model target- and task-modulation by implementing feed-back connections in consecutive layers in VGG19. Our results showed that this approach was able to mimic in part the pattern observed in neural data.

All results taken together point to the appealing possibility to make deep artificial neural networks more similar to natural neural circuits. However, when working with human data, the limited amount of data and the large inter-subjects variability represent a huge obstacle that needs to be overcome with the implementation of finer analytic tools and more biologically-inspired models.

References

[1] F. B. Colavita, "Human sensory dominance," *Perception & Psychophysics*, vol. 16, no. 2, pp. 409–412, Mar. 1974.

[2] F. B. Colavita and D. Weisberg, "A further investigation of visual dominance," *Perception & Psychophysics*, vol. 25, no. 4, pp. 345–347, Jul. 1979.

[3] D. Hecht and M. Reiner, "Sensory dominance in combinations of audio, visual and haptic stimuli," *Experimental Brain Research*, vol. 193, no. 2, pp. 307–314, Feb. 2009.

[4] C. Spence, "Explaining the Colavita visual dominance effect," *Progress in Brain Research*, vol. 176, pp. 245–258, 2009.

[5] M. Riesenhuber and T. Poggio, "Neural mechanisms of object recognition," *Current Opinion in Neurobiology*, vol. 12, no. 2, pp. 162–168, Apr. 2002.

[6] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, Oct. 1959.

[7] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154.2, Jan. 1962.

[8] D. H. Hubel and T. N. Wiesel, "Shape and arrangement of columns in cat's striate cortex," *The Journal of Physiology*, vol. 165, no. 3, pp. 559–568.2, Mar. 1963.

[9] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences*, vol. 15, no. 1, pp. 20–25, Jan. 1992.

[10] J. Norman, "Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches," *Behavioral and Brain Sciences*, vol. 25, no. 1, pp. 73–144, 2002, place: United Kingdom Publisher: Cambridge University Press.

[11] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012.

[12] C. G. Gross, "How Inferior Temporal Cortex Became a Visual Area," *Cerebral Cortex*, vol. 4, no. 5, pp. 455–469, Sep. 1994.

[13] G. A. Orban, "Higher order visual processing in macaque extrastriate cortex," *Physiological Reviews*, vol. 88, no. 1, pp. 59–89, Jan. 2008.

[14] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher, "The lateral occipital complex and its role in object recognition," *Vision Research*, vol. 41, no. 10, pp. 1409–1422, May 2001.

[15] G. A. Orban, D. Van Essen, and W. Vanduffel, "Comparative mapping of higher visual areas in monkeys and humans," *Trends in Cognitive Sciences*, vol. 8, no. 7, pp. 315–324, Jul. 2004.

[16] D. Pitcher, L. Charles, J. T. Devlin, V. Walsh, and B. Duchaine, "Triple Dissociation of Faces, Bodies, and Objects in Extrastriate Cortex," *Current Biology*, vol. 19, no. 4, pp. 319–324, Feb. 2009.

[17] M. W. Sliwinska and D. Pitcher, "TMS demonstrates that both right and left superior temporal sulci are important for facial expression recognition," *NeuroImage*, vol. 183, pp. 394–400, Dec. 2018.

[18] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex (New York, N.Y.: 1991)*, vol. 1, no. 1, pp. 1–47, Feb. 1991.

[19] A. T. Smith, K. D. Singh, A. L. Williams, and M. W. Greenlee, "Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex," *Cerebral Cortex (New York, N.Y.: 1991)*, vol. 11, no. 12, pp. 1182–1190, Dec. 2001.

[20] C. M. Ziemba, J. Freeman, J. A. Movshon, and E. P. Simoncelli, "Selectivity and tolerance for visual texture in macaque V2," *Proceedings of the National Academy of Sciences*, vol. 113, no. 22, pp. E3140–E3149, May 2016, publisher: National Academy of Sciences Section: PNAS Plus.

[21] J. Hegdé and D. C. Van Essen, "Selectivity for Complex Shapes in Primate Visual Area V2," *The Journal of Neuroscience*, vol. 20, no. 5, pp. RC61–RC61, Mar. 2000.

[22] A. Anzai, X. Peng, and D. C. Van Essen, "Neurons in monkey visual area V2 encode combinations of orientations," *Nature Neuroscience*, vol. 10, no. 10, pp. 1313–1321, Oct. 2007, number: 10 Publisher: Nature Publishing Group.

[23] V. Mountcastle, B. Motter, M. Steinmetz, and A. Sestokas, "Common and differential effects of attentive fixation on the excitability of parietal and prestriate (V4) cortical visual neurons in the macaque monkey," *The Journal of Neuroscience*, vol. 7, no. 7, pp. 2239–2255, Jul. 1987.

[24] S. Zeki, "Colour coding in rhesus monkey prestriate cortex," *Brain Research*, vol. 53, no. 2, pp. 422–427, Apr. 1973.

[25] S. Zeki, "The Distribution of Wavelength and Orientation Selective Cells in Different Areas of Monkey Visual Cortex," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 217, no. 1209, pp. 449–470, 1983, publisher: The Royal Society.

[26] A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel, "Toward a unified theory of visual area V4," *Neuron*, vol. 74, no. 1, pp. 12–29, Apr. 2012.

[27] R. Lafer-Sousa and B. R. Conway, "Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex," *Nature Neuroscience*, vol. 16, no. 12, pp. 1870–1878, Dec. 2013, number: 12 Publisher: Nature Publishing Group.

[28] C. D. Gilbert and T. N. Wiesel, "Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex," *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 9, no. 7, pp. 2432–2442, Jul. 1989.

[29] A. Das and C. D. Gilbert, "Long-range horizontal connections and their role in cortical reorganization revealed by optical recording of cat primary visual cortex," *Nature*, vol. 375, no. 6534, pp. 780–784, Jun. 1995.

[30] F. Baluch and L. Itti, "Mechanisms of top-down attention," *Trends in Neurosciences*, vol. 34, no. 4, pp. 210–224, Apr. 2011.

[31] S. R. Lehky and K. Tanaka, "Neural representation for object recognition in inferotemporal cortex," *Current Opinion in Neurobiology*, vol. 37, pp. 23–35, Apr. 2016.

[32] H. Tang, C. Buia, R. Madhavan, N. E. Crone, J. R. Madsen, W. S. Anderson, and G. Kreiman, "Spatiotemporal dynamics underlying object completion in human ventral visual cortex," *Neuron*, vol. 83, no. 3, pp. 736–748, Aug. 2014.

[33] J. S. Johnson and B. A. Olshausen, "The recognition of partially visible natural objects in the presence and absence of their occluders," *Vision Research*, vol. 45, no. 25, pp. 3262–3276, Nov. 2005.

[34] A. K. Bansal, R. Madhavan, Y. Agam, A. Golby, J. R. Madsen, and G. Kreiman, "Neural Dynamics Underlying Target Detection in the Human Brain," *Journal of Neuroscience*, vol. 34, no. 8, pp. 3042–3055, Feb. 2014, publisher: Society for Neuroscience Section: Articles.

[35] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, p. 7068349, 2018.

[36] R. K. Sinha, R. Pandey, and R. Pattnaik, "Deep Learning For Computer Vision Tasks: A review," *arXiv:1804.03928 [cs]*, Apr. 2018, arXiv: 1804.03928.

[37] C. Henry, S. M. Azimi, and N. Merkle, "Road Segmentation in SAR Satellite Images with Deep Fully-Convolutional Neural Networks," *arXiv:1802.01445 [cs]*, Aug. 2018, arXiv: 1802.01445.

[38] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, conference Name: Proceedings of the IEEE.

[39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, number: 7553 Publisher: Nature Publishing Group.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[41] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, arXiv: 1409.1556.

[42] C. R. Ponce, W. Xiao, P. F. Schade, T. S. Hartmann, G. Kreiman, and M. S. Livingstone, "Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences," *Cell*, vol. 177, no. 4, pp. 999–1009.e10, May 2019.

[43] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, Jun. 2014, publisher: National Academy of Sciences Section: Biological Sciences.

[44] T. P. O'Connell, M. M. Chun, and G. Kreiman, "Zero-shot neural decoding of visual categories without prior exemplars," *bioRxiv*, p. 700344, Jul. 2019, publisher: Cold Spring Harbor Laboratory Section: New Results.

[45] S. Grossman, G. Gaziv, E. M. Yeagle, M. Harel, P. Mégevand, D. M. Groppe, S. Khuvis, J. L. Herrero, M. Irani, A. D. Mehta, and R. Malach, "Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks," *Nature Communications*, vol. 10, no. 1, p. 4934, Oct. 2019, number: 1 Publisher: Nature Publishing Group.

[46] R. T. Pramod and S. P. Arun, "Do Computational Models Differ Systematically from Human Object Perception?" in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1601–1609, iSSN: 1063-6919.

[47] H. Tang, M. Schrimpf, W. Lotter, C. Moerman, A. Paredes, J. O. Caro, W. Hardesty, D. Cox, and G. Kreiman, "Recurrent computations for visual pattern completion," *Proceedings of the National Academy of Sciences*, vol. 115, no. 35, pp. 8835–8840, Aug. 2018, publisher: National Academy of Sciences Section: Biological Sciences.

[48] M. Zhang, J. Feng, K. T. Ma, J. H. Lim, Q. Zhao, and G. Kreiman, "Finding any Waldo with zero-shot invariant and efficient visual search," *Nature Communications*, vol. 9, no. 1, p. 3730, Sep. 2018, number: 1 Publisher: Nature Publishing Group.

[49] H. Adeli and G. Zelinsky, "Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2013–201 310.

[50] G. Kreiman and T. Serre, "Beyond the feedforward sweep: feedback computations in the visual cortex," *Annals of the New York Academy of Sciences*, vol. 1464, no. 1, pp. 222–241, 2020.

[51] H. Liu, Y. Agam, J. R. Madsen, and G. Kreiman, "Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex," *Neuron*, vol. 62, no. 2, pp. 281–290, Apr. 2009.

[52] Y. Agam, H. Liu, A. Papanastassiou, C. Buia, A. J. Golby, J. R. Madsen, and G. Kreiman, "Robust selectivity to two-object images in human visual cortex," *Current biology: CB*, vol. 20, no. 9, pp. 872–879, May 2010.

[53] J. Jonas, C. Jacques, J. Liu-Shuang, H. Brissart, S. Colnat-Coulbois, L. Maillard, and B. Rossion, "A face-selective ventral occipito-temporal map of the human brain with intracerebral potentials," *Proceedings of the National Academy of Sciences*, vol. 113, no. 28, pp. E4088–E4097, 2016.

[54] S. Khuvis, E. M. Yeagle, Y. Norman, S. Grossman, R. Malach, and A. D. Mehta, "Face-selective units in human ventral temporal cortex reactivate during free recall," *bioRxiv*, 2019.

[55] A. Campbell and C. Wu, "Chronically Implanted Intracranial Electrodes: Tissue Reaction and Electrical Changes," *Micromachines*, vol. 9, no. 9, p. 430, Sep. 2018, number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[56] M. L. Seghier, "The Angular Gyrus: Multiple Functions and Multiple Subdivisions," *The Neuroscientist*, vol. 19, no. 1, pp. 43–61, Feb. 2013, publisher: SAGE Publications Inc STM.

[57] M. Ahissar and S. Hochstein, "The reverse hierarchy theory of visual perceptual learning," *Trends in Cognitive Sciences*, vol. 8, no. 10, pp. 457–464, Oct. 2004.

[58] A. Nayebi, J. Sagastuy-Brena, D. M. Bear, K. Kar, J. Kubilius, S. Ganguli, D. Sussillo, J. J. DiCarlo, and D. L. K. Yamins, "Goal-Driven Recurrent Neural Network Models of the Ventral Visual Stream," *bioRxiv*, p. 2021.02.17.431717, Feb. 2021, publisher: Cold Spring Harbor Laboratory Section: New Results.

[59] A. Nayebi, D. Bear, J. Kubilius, K. Kar, S. Ganguli, D. Sussillo, J. J. DiCarlo, and D. L. K. Yamins, "Task-Driven Convolutional Recurrent Models of the Visual System," *arXiv:1807.00053 [cs, q-bio]*, Oct. 2018, arXiv: 1807.00053.