Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features

Simone Palazzo, Concetto Spampinato, *Member, IEEE,* Isaak Kavasidis,
Daniela Giordano, *Member, IEEE,* and Mubarak Shah, *Fellow, IEEE,*

Abstract—This paper tackles the problem of learning brain-visual representations for understanding neural processes behind human visual perception, with a view towards replicating these processes into machines. The core idea is to learn plausible presentations through the use of human neural activity evoked by natural images as a supervision mechanism for deep learning models. We propose a multimodal approach that uses deep encoders for images and EEGs, trained in a siamese configuration for learning a joint manifold that maximizes a compatibility measure between visual features and brain representations. We carry out image classification and saliency detection on the learned manifold, and shed light on the possible representations generated by the human brain when perceiving the visual world. Performance analysis shows that neural signals can be used to effectively supervise the training of deep learning models, as demonstrated by the achieved performance in both image classification and saliency detection. Furthermore, the learned brain-visual manifold is consistent with cognitive neuroscience literature about visual perception and, most importantly, highlights new associations between brain areas, image patches and computational kernels. In particular, we are able to approximate brain responses to visual stimuli by training an artificial model with image features correlated to neural activity.

1 Introduction

Human visual capabilities are coming within reach of artificial systems, mainly thanks to the recent advances in deep learning. Indeed, deep feedforward and recurrent neural networks, loosely inspired by the primate visual architecture, have led to a significant boost in performance of computer vision, natural language processing, speech recognition and game playing. Beside the significant performance gain in such tasks, the representations learned by deep computational models appear to be highly correlated to brain representations; for example, correlations can be found between brain representations in the visual pathway and the hierarchical structures of layers in deep neural networks (DNNs) [1], [2]. These findings have paved the way to a recent multidisciplinary effort, involving cognitive neuroscientists and artificial intelligence researchers, aiming at reverse-engineering the human mind and its adaptive capabilities [3], [4], [5], [6]. Nevertheless, this multidisciplinary field is still in its infancy. Indeed, the existing computational neural models loosely emulate computations and connections of biological neurons and often ignore feedforward and feedback neural interactions. For example, visual recognition in humans appears to be mitigated by a multi-level aggregation of information being processed forward and backward across cortical brain regions [7], [8], [9], [10]. Recent approaches [11], inspired by the hierarchical predictive coding in neuroscience [12], [13], have attempted to encode such kind of additional information into computational models by proposing recurrent neural networks with feedforward, feedback, and recurrent connections. These models have shown promising performance in visual classification tasks and demonstrate that understanding the human brain in more detail and transferring this knowledge to engineering models is the key for better machines.

We believe that a promising direction towards uncovering the workings of the human visual system lies in combining and correlating neural activity data recorded from human subjects while performing specific tasks, and computational models developed for the same exact tasks. By investigating the learned computational representations and how they correlate with neural activity over time, it is possible to infer and analyze complex brain processes. In this paper we propose a multimodal approach based on deep encoders trained in a siamese configuration that, given EEG brain activity data from several human subjects performing visual categorization of still images, learns a joint brain-visual embedding by finding similarities between brain representations and visual features. This embedding is used to perform image classification, saliency detection, and to shed light on the possible representations generated in the human brain for visual scene analysis. In particular, this paper demonstrates that a) neural activity data can used as an alternative and richer way to supervise effectively the development of deep learning models, as demonstrated by the first saliency detection method leveraging neural signals, b) neural processes involved in human vision system can be uncovered, with sufficient approximation, by maximizing similarities with deep models; indeed we were able to demonstrate that EEG encodes saliency information, as well as to localize neural generators, i.e. low- and middle-level visual stimuli responsible for activations of specific cortex areas.

The paper is organized as follows. In the next section we review the state of the art of recent methods combining brain data and computational models, as well as related

- *S. Palazzo, C. Spampinato, I. Kavasidis and D. Giordano are with the Department of Electrical, Electronic and Computer Engineering, University of Catania, Viale Andrea Doria, 6, Catania, 95125, Italy.*

  E-mail: palazzosim, cspampin, kavasidis, dgiordan@dieei.unict.it
- *M. Shah and C. Spampinato are with the Center of Research in Computer Vision, University of Central Florida. E-mail: shah@crcv.ucf.edu*

approaches on multimodal learning. In Sect. 3 we describe the core of our approach, specifically the framework to learn a joint brain-visual embedding from data, then proceed with the methods for extracting the most relevant image and brain activity patterns and their interpretation (Sect. 4–6). Sect. 7 reports the achieved experimental results on image classification and saliency detection, and shows the learned representations that mostly affect visual categorization. In the last section, conclusion and future directions are presented.

2 Related work

Our work relates mainly to the fields of computational neuroscience for brain decoding, machine learning guided by brain activity and multimodal learning. The recent state of the art of these areas are briefly reviewed in this section.

Computational neuroscience for decoding brain representations. Decoding brain representations has been a long sought objective and it still is a great challenge of our times. In particular, cognitive neuroscience works have made great progress in understanding neural representations originated in the primary visual cortex (V1). Indeed, it is known that the primary visual cortex is a retinotopically organized series of oriented edge and color detectors [14] that feed-forward into neural regions focused on more complex shapes and feature dimensions, which operate over larger receptive fields in areas V4 [15], before finally arriving at object and category representations in the inferior temporal (IT) cortex [16]. Neuroimaging methods, such as fMRI, MEG, and EEG, have been crucial for these findings. However, the level of neural activity detail (spatial or temporal) provided by these techniques is insufficient to fully decode visual processes, although they clearly contain enough information for accurate reconstruction of visual experiences [17]. To overcome technology limitations, brain representation decoding has been recently tackled by investigating the correlation between neural activity data and computational models [1], [2]. However, these approaches mainly perform simple correlations between deep learned representations and neuroimaging data and, according to the obtained outcomes, draw conclusions on brain representations, which is too simplistic from our point of view. Indeed, the core point of our idea is that understanding the human visual system will come as a result of training automated models to maximize signal correlation between brain activity and evoking stimuli, not as a pure analysis of brain activity data. In addition, while most of the methods attempting to decode brain representations use brain images at high spatial resolution (recorded by fMRI), our work is the first one employing EEG data that, despite being at lower spatial resolution, has higher temporal resolution, making it more suitable to decode fast brain processes like those involved in the visual pathway.

Machine learning guided by brain activity. The intersection and overlap between machine learning and cognitive neuroscience has increased significantly in recent years.

Deep learning methods are used, for instance, for neural response prediction [18], [19], [20], and, in turn, biologically-inspired mechanisms such as coding theory [11], working memory [21] and attention [22], [23] are increasingly being adopted. However, to date, human cognitive abilities still seem too complex to understand computationally, and a data-driven approach for "reverse engineering" the human mind might be the best way to inform and advance artificial intelligence [24]. Under this scenario, recent studies have employed neural activity data to constrain model training. For example, in our recent work [3], we mapped visual features learned by a deep feed-forward model to brain-features learned directly from EEG data for performing automated visual classification. The authors of [25] employed fMRI data to bias the output of a machine learning algorithm and push it to exploit representations found in visual cortex. This work resembles one of the first methods relying on brain activity data to perform visual categorization [26], with the difference that the former, i.e., [25], explicitly utilizes neural activity to weigh the training process (similarly to [27]), while the latter, i.e. [26], proposes a kernel alignment algorithm to fuse the decision of a visual classifier with brain data.
In this paper, we propose a deeper interconnection between the two fields: instead of using neural data as a signal to weigh computationally-learned representations, we learn a mapping between images and corresponding neural activity, so that visual patterns are related one-to-one to neural processes. This mapping, as we demonstrate in the experimental results, may reveal much more information on brain representations and be able to guide the training process in a more intrinsic and comprehensive way. Thus, our approach is not just a hybrid machine learning method inspired or constrained by brain knowledge, but a method that implicitly finds similarities between computational and brain representations and uses them to perform visual tasks.

Multimodal learning. Another line of research related to ours is multimodal learning, which relies on the fact that real-world information comes in several modalities, each carrying different — yet equally useful — content for building intelligent systems. Multimodal learning methods [28], [29], [30], in particular, attempt to learn embeddings by finding a joint representation of the different modalities that encodes the real-world features of the common concept corresponding to the input data. An effective joint representation must preserve both intra-modality similarity (e.g., two similar images should have close vector representation in the joint space; likewise, two equivalent text descriptions should have similar representations as well) and inter-modality similarity (e.g., an image and a piece of text describing the content of that image should be closer in the joint space than an image and an unrelated piece of text). Following this property, most methods find correspondences between visual data and text [30], [31], [32], [33] or audio [34], [35], [36], [37] to support either discriminative tasks (e.g., classification) or prediction of one modality conditioned on another (e.g., image synthesis or retrieval). For the former type of methods, captions and

tags have been used to improve accuracy of both shallow and deep classifiers [33], [38]. Analogously, [35] used audio to supervise visual representations; [36], [37] used vision to supervise audio representations; [39] used sound and vision to jointly supervise each other; and [34] investigated how to separate and localize multiple sounds in videos by analyzing motion and semantic cues. Some other works, instead, have focused on predicting missing data in one modality conditioned on another, for example, generating text description from images and vice versa [40], [41], [42], [43], [44]. Reed et al. in [43] propose a joint representation space to condition generative adversarial networks (GANs) for synthesizing images from text descriptions. Similarly, Mansimov et al. [44] synthesized images from text captions using a variational autoencoder. In our recent paper [45], we used an embedding learned from brain signals to synthesize images both using GANs and variational autoencoders in a brain-to-image effort.

In this paper, our approach is inspired by the methods that learn a shared multimodal representation, with several crucial differences. First of all, one of the modalities we employ is brain activity data (EEG), whose informative content — differently from text/audio — is largely unknown and much noisier. This makes it much harder to discover relationships between the visual and brain modalities. In this sense, our approach is intended not only to improve prediction accuracy, but as a knowledge discovery tool to uncover brain processes. Thus, our main objective is to learn a reliable joint representation and explore the learned space to find correspondences between visual and brain features that can uncover brain representations; these, in turn, can be employed for building better deep learning models.
In addition, the proposed deep multimodal network, consisting of two encoders (one per modality), is trained in a siamese configuration and employs a loss function enforcing the learned embedding to be representative of intra-class differences between samples, and not just of the inter-class discriminative features (as done, for instance, in [43]).

3 Multimodal Learning of Visual-Brain Features

Neural activity (recorded by EEG) and visual data have very different structures, and finding a common representation may not be trivial. Previous approaches [3] have attempted to find such representations by training each side of the problem individually: for example, by first learning brain representations by training a recurrent classifier on EEG signals, and then training a CNN to regress the visual features to brain features for corresponding EEG/image pairs. While this provides useful representations, the utility of the learned features is strongly tied to the proxy task employed to compute the initial representation (e.g., image classification), and focuses more on learning class-discriminative features than on finding relations between EEG and visual patterns. Hence, we argue that any transformations from human neural signals and images to a common space should be learned jointly by maximizing the similarity between two embeddings of each input representation. To this aim, we define a siamese network for learning a structured joint embedding between EEG signals and images using deep encoders, and maximize a measure of similarity between the two modalities. The architecture of our model is shown in Fig. 1.

More formally, let $\mathcal{D} = \{e_i, v_i\}_{i=1}^{N}$ be a dataset of neural signal samples and images, such that each neural (EEG) sample e_i is recorded on a human subject in response to viewing image v_i. Ideally, latent information content should be shared by e_i and v_i. Also, let $\mathcal{E}$ be the space of EEG signal samples and $\mathcal{V}$ the space of images. The objective of our method is to train two encoders that respectively map neural responses and images to a common space $\mathcal{J}$, namely $\varphi : \mathcal{E} \rightarrow \mathcal{J}$ and $\theta : \mathcal{V} \rightarrow \mathcal{J}$.

In other approaches for structured learning (e.g. [43]), the training of the encoders is proxied as a classification problem based on the definition of a *compatibility function* $F : \mathcal{E} \times \mathcal{V} \rightarrow \mathbb{R}$, that computes a similarity measure as the dot product between the respective embeddings of an EEG/image pair:

$$F(e, v) = \varphi(e)^T \theta(v) \ . \tag{1}$$

While we employ the same modeling framework, we formulate the problem as an embedding task whose only objective is to maximize similarity between corresponding pairs, without implicitly performing classification, as this would take us back to the limitation of [3], i.e., learning representations tied to the classification task.
In order to abstract the learning process from any specific task, we train our siamese network with a triplet loss aimed at mapping the representations of matching EEGs and images to nearby points in the joint space, while pushing apart mismatched representations. We can then stick to the structured formulation of the compatibility function in Eq. 1 by employing F directly for triplet loss computation. Thus, given two pairs of EEG/image (e_1, v_1) and (e_2, v_2), we consider e_1 as the *anchor* item, v_1 as the *positive* item and v_2 as the *negative* item. Using compatibility F (which is a similarity measure rather than a distance metric, as is more commonly used in triplet loss formulations), the loss function employed to train the encoders becomes:

$$L(e_1, v_1, v_2) = \max\{0, F(e_1, v_2) - F(e_1, v_1)\} \ . \tag{2}$$

This equations assigns a zero loss only when compatibility is larger for (e_1, v_1) than for (e_1, v_2). Note that class labels are not used anywhere in the equation. This makes sure that the resulting embedding does not just associate class-discriminative vectors to EEG and images, but tries to extract more comprehensive patterns that explain the relations between the two data modalities. Also, there is no margin term in Eq. 2, as would be typical in hinge loss formulations of a triplet loss. This is due to (e_1, v_1) and (e_2, v_2) possibly being members of the same visual class, and forcing a minimum distance between the same-class items is not strictly needed: as long as the learned representation assigns a larger compatibility to matching EEG/image pairs and learns general and meaningful patterns, class separability would still be implicitly achieved.

EEG encoder $\varphi(\cdot)$, which maps neural activity signals to the joint space $\mathcal{J}$, is based on convolutional layers for short-term temporal feature extraction and a recurrent module for long-term analysis, as shown in Figure 2. The input signal to the encoder, normalized channel-wise to zero mean and

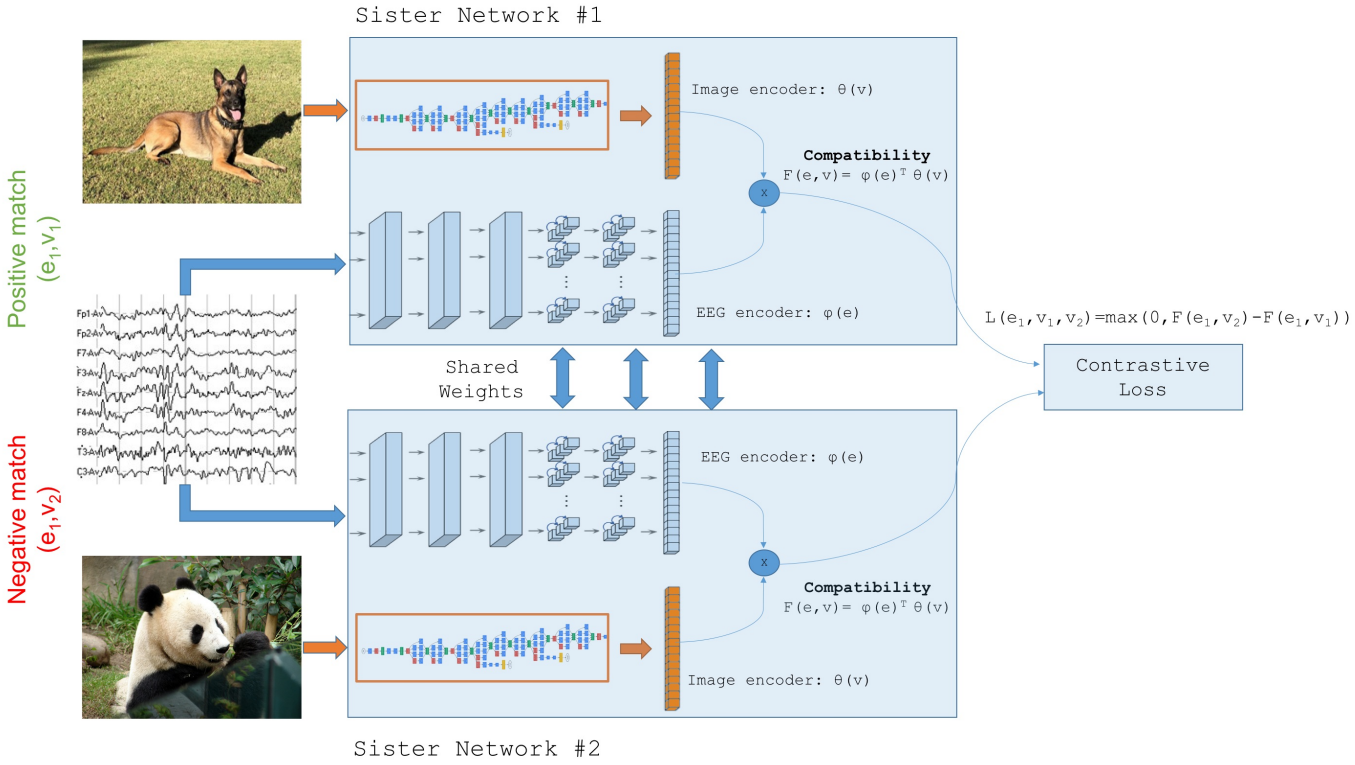

Fig. 1: **Siamese network for learning a joint brain-image representation**. The idea is to learn a space by maximizing a compatibility function between two embeddings of each input representation. Given a positive match between an image and the related EEG from one subject, and a negative match between the same EEG and a different image, the network is trained to ensure a closer similarity (higher compatibility) between related EEG/image pairs than unrelated ones.

unitary standard deviation, goes through an initial cascade of convolutional layers employing kernels of size 3 and dilated convolutions [46] to capture small/medium scale dynamics in the temporal dimension of the EEG signal, such as high-frequency or more specific local self-learned patterns. The features extracted by the convolutional layers are then fed to a recurrent module (e.g., LSTM or GRU), in order to analyze longer term temporal dynamics in the signal. The hidden state of the recurrent module at the last time step is used as a summary of the whole signal and then passed to a linear layer that projects it to the joint embedding space. Using a recurrent layer makes the encoder independent of the input sequence length, thus making it easy to adapt to different data sizes and sampling frequencies. The specific design configurations employed are evaluated and discussed in Section 7.2.

Visual encoder, $\theta(\cdot)$ maps, instead, images to the joint space $\mathcal{J}$ through convolutional neural networks. We use a pre-trained CNN to extract visual features and feed them to a linear layer for mapping to the joint embedding space. Differently from [43], we learn the compatibility function in an end-to-end fashion, also by fine-tuning the image encoder, in order to better identify low- and middle- level visual-brain representations, which — suitably decoded — may provide hints on what information is used by humans when analyzing visual scenes.

4 Image Classification and Saliency Detection

Our siamese network learns visual and EEG embeddings in order to maximize the similarities between images and related neural activities. We can leverage the learned manifold for performing visual tasks. In cognitive neuroscience there is converging evidence that: a) brain activity recordings contain information about visual object categories (as also demonstrated in [3]) and b) attention influences the processing of visual information even in the earliest areas of the primate visual cortex [47]. In particular, bottom-up sensory information and top-down attention mechanisms seem to fuse in an integrated saliency map, which in turn, distributes across the visual cortex. Thus, EEG recordings in response to visual stimuli should encode both visual class and saliency information. However, while for image classification we can simply use the trained encoders as feature extractors for a subsequent classification layer (see performance evaluation in Sect. 7), for saliency detection we designed a *multiscale suppression-based* approach, inspired by the methods identifying pixels relevant to CNN neuron activations (e.g., [48]), that analyzes fluctuations in the compatibility measure F. The idea is based on measuring how brain-visual compatibility varies as image patches are suppressed at multiple scales. Indeed, the most important features in an image are those that, when inhibited in an image, lead to the largest drop in compatibility score (computed by feeding an EEG/image pair to the siamese network proposed in the

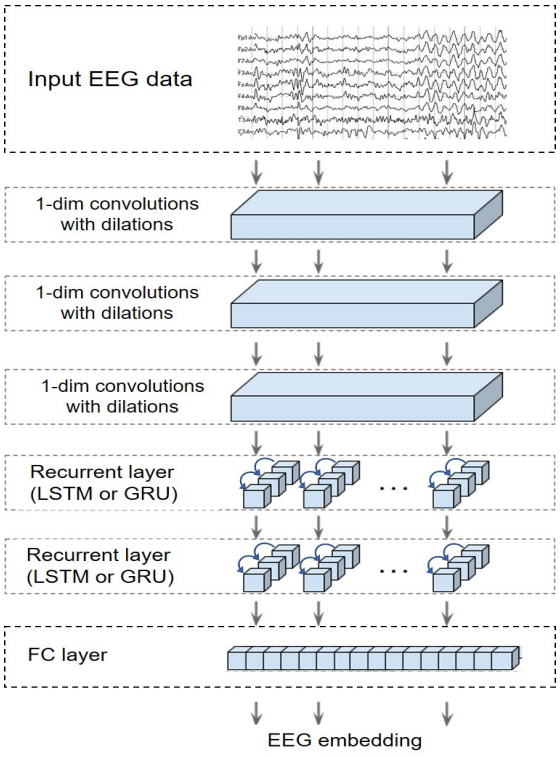

Fig. 2: **Architecture of the EEG encoder**. The EEG signal is first processed by banks of 1D convolutions operating on each channel individually. The resulting features, encoding short-term temporal features, are then fed into a cascade of recurrent layers for analyzing long-term dynamic. A final fully-connected layer projects the last recurrent hidden state to the joint embedding space.

previous section) with respect to the corresponding neural activity signal. Thus, we employ compatibility variations at multiple scale for *saliency detection*. Note that, for this approach to work, the EEG encoder must have learned to identify patterns related to specific visual features in the observed image, so that the absence of those featueres reflects on smaller similarity scores on the joint embedding space.

The saliency detection method is illustrated in Fig. 3 and can be formalized as follows. Let (e, v) be an EEG/image pair, with compatibility $F(e, v)$. The saliency value $S(x, y, \sigma, e, v)$ at pixel (x, y) and scale σ is obtained by removing the $\sigma \times \sigma$ image region around (x, y) and computing the difference between the original compatibility score and the one after suppressing that patch. More formally, if $m_\sigma(x, y)$ is a binary mask where all pixels within the $\sigma \times \sigma$ mask around (x, y) are set to zero, we have:

$$S(x, y, \sigma, e, v) = F(e, v) - F(e, m_\sigma(x, y) \odot v) \ , \qquad (3)$$

where $\odot$ denotes element-wise multiplication (Hadamard product). For multiple scale values, we set the overall saliency value for pixel (x, y) to the normalized sum of (per scale) saliency scores:

$$S(x, y, e, v) = \sum_\sigma S(x, y, \sigma, e, v) \ . \qquad (4)$$

Normalization is then performed on an image-by-image basis for visualization.

5 Visual-related brain processes

While the saliency detection approach studies how alterations in images reflect on compatibility scores, it is even more interesting to analyze how neural patterns act on the learned representations. Indeed, following the principle that large variations in compatibility can be found when the most important visual features are masked, we may similarly expect compatibility to drop when we remove "important" (from a visual feature–matching point of view) components from neural activity signals. Performing this analysis traditionally requires a combination of *a priori* knowledge on brain signal patterns and manual analysis: for example, it is common to investigate the effect of provided stimuli while monitoring the emergence of event-related potentials (ERPs) known to be associated to specific brain processes.

Of course, posing the problem in this way still requires that the processes under observation be at least partially known, which makes it complicated to automatically detect previously-unknown signal patterns.

Instead, the joint representation makes it easy to correlate brain signals with visual stimuli by analyzing how compatibility varies in response to targeted modifications of the inputs. Thus, similar to saliency detection, we can identify the spatial components in brain activity that convey visual information.

As mentioned in Sect. 2, object recognition in humans is performed by a multi-level aggregation of shape and feature information across cortical regions, resulting in a distributed representation that can easily adapt to a wide variety of tasks on the received stimuli. For these reasons, understanding how this distributed representation is spatially localized over the brain cortex is a fundamental step towards a successful emulation of the human visual system. In order to evaluate the importance of each EEG channel (and corresponding brain area), we employ the learned joint embedding space to "filter" (the exact procedure is defined below) that channel from the EEG signal and measure the corresponding change in compatibility between images and filtered signals.

The importance of each channel for a single EEG/image pair can be measured by computing the difference between the pair's compatibility score and the compatibility obtained when suppressing that channel from the EEG signal. Given an EEG/image pair (e, v), and indicating with e_{-c} a transformation of e such that information on channel c has been suppressed, we can define the *importance* of channel c as:

$$I_c(e, v) = \mathbb{E}_{(e,v)} \left[F(e, v) - F(e_{-c}, v) \right] \ , \qquad (5)$$

where the expectation is computed over all dataset samples. The intuition behind this formulation is that the suppression of a channel that conveys unnecessary information (at least, from the point of view of the representation learned by the EEG encoder) should result in a small difference in the compatibility score; analogously, if a channel contains important information that match brain activity data to

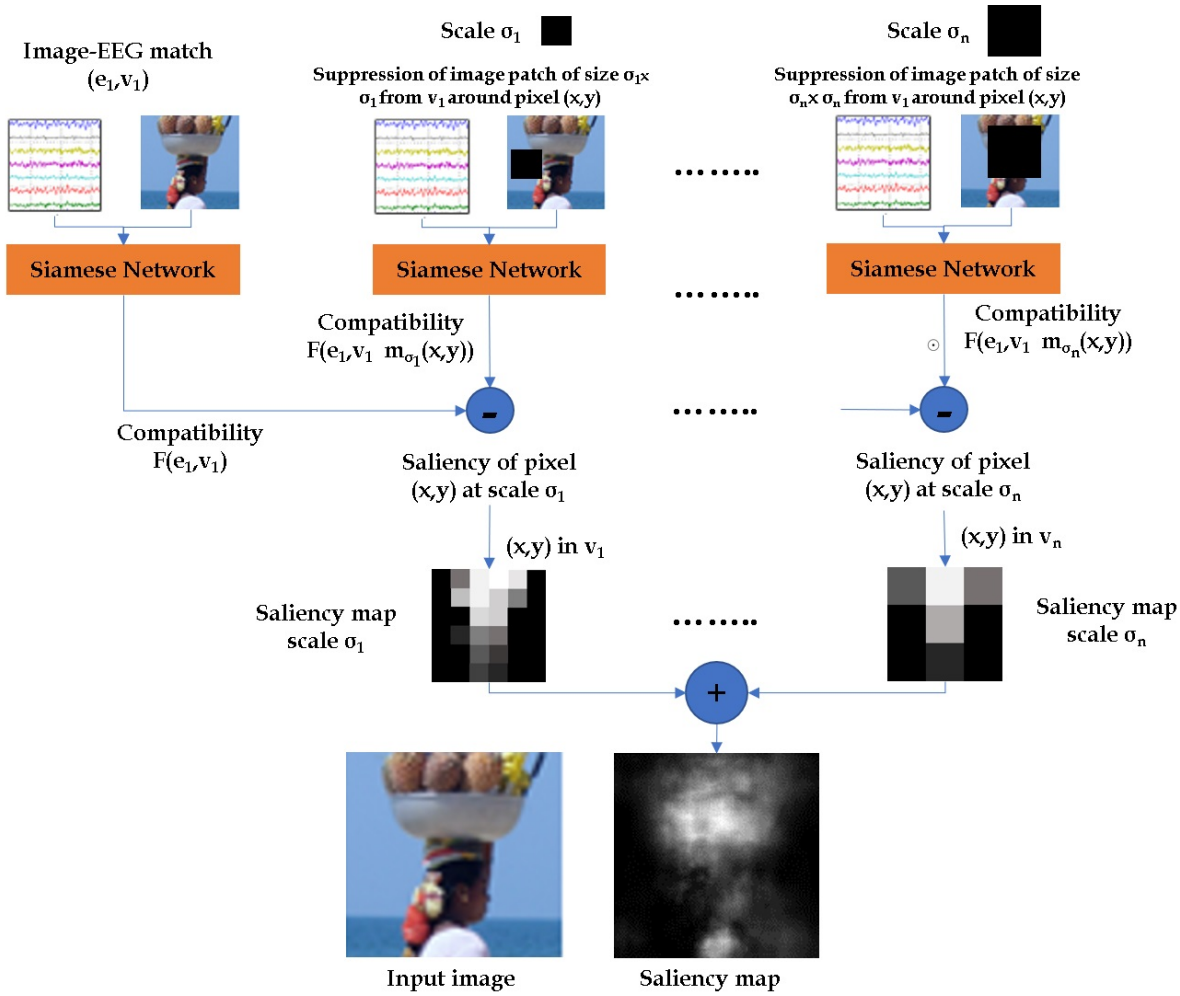

Fig. 3: **Our multiscale suppression-based saliency detection**. Given an EEG/image pair, we estimate the saliency of an image patch by masking it and computing the corresponding variation in compatibility. Performing the analysis at multiple scales and for all image pixels results in a saliency map of the whole image. Note that, although the example scale-specific saliency maps appear pixellated, that is only a graphical artifact to give the effect of scale: in practice, scale-specific maps are still computed pixel by pixel.

visual data, compatibility should drop when that channel is suppressed.

To compute e_{-c}, channel c is suppressed by replacing its values with a sequence of random Gaussian samples, low-pass filtered at 100 Hz and distributed according to its estimated statistics (mean and variance).

More formally, if EEG signal e is represented as a matrix with one channel per row:

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_c \\ \cdots \\ e_n \end{pmatrix}, \tag{6}$$

we compute $I_c(e, v)$ as:

$$I_c(e,v) = F(e,v) - \mathbb{E}\left[F\left(\begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ H\left(\mathcal{N}(\mu_c, \sigma_c^2)_{L\times 1}\right) \\ \cdots \\ e_n \end{pmatrix}, v \right) \right]. \tag{7}$$

where μ_c and σ_c^2 are the sample mean and variance for channel c, L is the EEG temporal length, $\mathcal{N}(\mu, \sigma^2)_{N\times M}$ is an $N \times M$ matrix sampled from the specified distribution, and H is a low-pass filter at 100 Hz.

6 Decoding Brain Representations

Each of the previous approaches investigates the effect of altering either the brain activity signals or the image content, but they are limited in that the differential analysis they provide is carried out on only one modality: we can identify the visual features that mostly impact the similarity between

two corresponding encodings, or we can identify the spatial patterns in brain activity that likewise are more relevant to the learned representation. However, we still do not know *which* visual features give rise to *which* brain responses, i.e. *neural generators*. To fill this gap, we propose an additional modality for interpreting compatibility differences, by employing the learned manifold to carry out an analysis of the EEG channels — and, therefore, the corresponding brain regions — that are most solicited in the detection of visual characteristics at different scales, from edges to textures to objects and visual concepts. To carry out this analysis, we evaluate the differences in compatibility scores computed when specific feature maps in the image encoder are removed, and map the corresponding features to the EEG channels that appear to be least active (compatibility-wise) when those features were removed. In practice, given EEG/image pair (e, v), let us define $F(e, v_{-l,f})$ as the value of the compatibility function computed by suppressing the f-th feature map at the l-th layer of the image encoder. According to Eq. 5, the importance of channel c computed when a certain layer's feature is removed is:

$$I_c(e, v_{-l,f}) = \\ F(e, v_{-l,f}) - \mathbb{E}\left[F\left(\left(\begin{array}{c} e_1 \\ e_2 \\ \cdots \\ H\left(\mathcal{N}(\mu_c, \sigma_c^2)_{L\times 1}\right) \\ \cdots \\ e_n \end{array}\right), v_{-l,f}\right)\right] \quad (8)$$

We then define the *association* between feature (l, f) and channel c for a pair (e, v) as follows:

$$A_{c,l,f}(e, v) = \mathbb{E}_{(e,v)}\left[I_c(e, v_{-l,f}) - I_c(e, v)\right] . \quad (9)$$

We consider channel c and feature (l, f) "associated" if, after removing the intrinsic importance score for that channel for a given (e, v) pair, the variation in compatibility for channel c does not vary when that feature is removed, which would mean no visual component in the encoded representation is left unmatched.

We can estimate the association between channel c and layer l by averaging over all features in that layer:

$$A_{c,l}(e, v) = \mathbb{E}_{(e,v),f}\left[A_{c,l,f}(e, v)\right] . \quad (10)$$

The resulting score provides an interesting indication of how much the features computed at a certain layer in a computational model resemble the features processed by the brain in specific cortical areas.

7 Experiments and Applications

We evaluated the quality and meaningfulness of the joint encoding learned by our model in several applications with the main objective to assess the correspondence of visual and neural contents in the shared representation:

- *Brain signal/image classification*: we evaluate if and how the learned neurovisual manifold is representative of the two input modalities by assessing their capabilities to support classification tasks, namely, brain signal classification and image classification.
- *Visual saliency detection from neural activity/image compatibility variations*: the similarity of the mapped neural signals and images should be based on the most significant features of each image, as described in Sect. 4. This evaluation, therefore, aims at assessing the performance of our brain-based saliency detection and comparing it to state of the art methods.
- *Localizing neural processes related to visual content*: this experiment aims at identifying neural locations related to specific image patches by using the method described in Sect. 5. By combining the results of this analysis and saliency detection, we obtain the first ever retinotopic saliency map created by training an artificial model with salient visual features correlated with neural activity.
- *Correslating deep learned representations with brain activity*: by analyzing the learned visual and neural patterns, we identify what the most influencing learned visual features (kernels) are and how they correlate with neural activity. The outcome of this evaluation indicates roughly what visual representations correlate the most with the ones learned/used by the human brain, thus representing a first, important, step forward for developing a methodology to better uncover and emulate brain mechanisms.

7.1 Brain-Visual Dataset

The employed neural activity dataset, published in [3], contains 11,965 EEG sequences recorded while a set of 6 participant subjects looked at images displayed on a computer screen. The images were taken from a subset of 40 classes from ImageNet [49], with 50 images per class. Each sample in the dataset then corresponds to a triplet of EEG data, image reference and subject.

In particular, each EEG sample has 128 channels, recorded for 0.5 seconds at 1 KHz sampling rate, represented as a 128×L matrix, with $L \approx 500$ being the temporal length of the channel. The exact duration of each signal may vary, so we discard the first 40 samples (40 ms) to reduce interferences from the previous image and then cut the signal to a common length of 440 samples (to account for signals with $L < 500$), when supposedly all image-related visual and cognitive processes will have been completed. EEG placement and a rough matching with brain cortices is shown in Fig. 4, where we also show the neural activity visualization scale employed in this paper. Activity heatmaps for an image/EEG pair are generated by applying Eq. 5 to estimate how much each channel affects the pair's compatibility, then plotting normalized channel importance scores on a 2D map of the scalp (at the positions corresponding to the electrodes of the employed EEG cap), and applying a Gaussian filter for smoothing (using a kernel with standard deviation of 13 pixels, for a 400×400 map).

In order to replicate the training conditions employed in [3] and to make a fair comparison for brain signal classification, we use the same training, validation and test splits of the EEG dataset, consisting respectively of 1600 (80%), 200 (10%), 200 (10%) images with associated EEG signals, ensuring that all signals related to a given image belong to the same split.

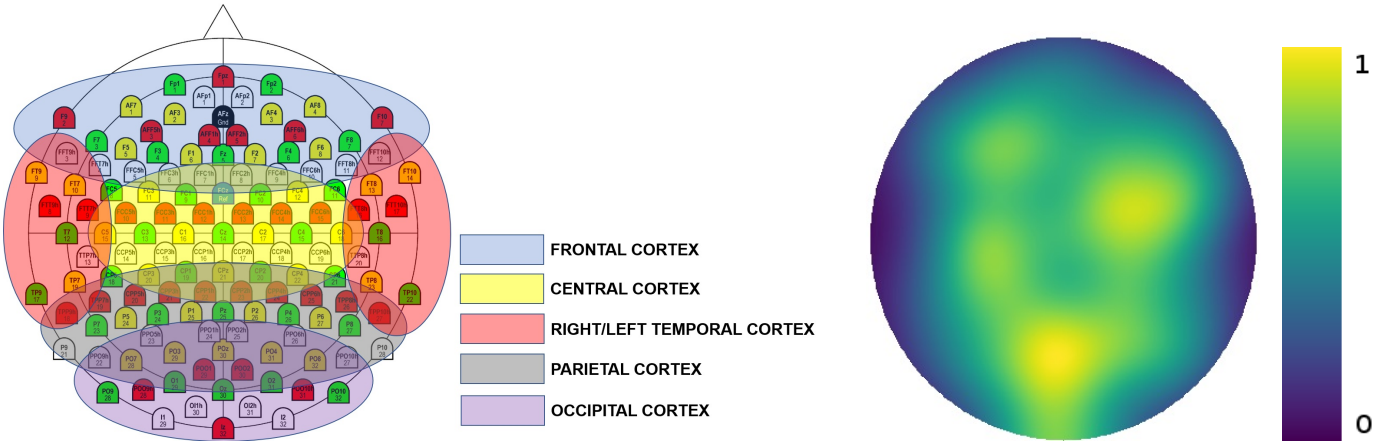

Fig. 4: **Mapping between EEG channels and brain cortices.** (Left) EEG channel placement and corresponding brain cortices (background image source : Brain Products GmbH, Gilching, Germany). We used a 128-channel EEG, each channel in the figure is identified by a prefix letter referring to brain cortex (Fp: frontal, T: temporal, C: central, P: parietal, O: occipital) and a number indicating the electrode. (Right) Neural activation visualization — top view of the scalp — employed in this paper. A detailed mapping between EEG channels and brain cortices can be found in [50].

7.2 Siamese network training for classification

In this section, we describe the training procedure of our siamese network and evaluate the quality of the learned joint embedding for visual and brain signal classification tasks. In particular, we investigate a) what configurations of the two encoders of our model (defined in Sect. 3) provide the best trade-off between EEG and image classification, b) how conditioning the classifier of one modality with the other one affects classification accuracy, and c) if augmenting the visual representation space with features derived from the brain leads to better performance than state-of-the-art methods that only use visual features.

We train our siamese network consisting of the two EEG and image encoders, sampling a triplet (e_i, v_i, v_j) of one EEG (e_i) and two images (v_i, v_j), representing, respectively, the positive (e_i, v_i) and negative (e_i, v_j) samples for the siamese network. The optimization algorithm for our contrastive loss is Adam with suggested hyperparameters (learning rate: 0.001, β_1: 0.9, β_2: 0.999), a mini-batch size of 16, and number of training epochs set to 100.

We also test different configurations of EEG and image encoders. For the former, we use an initial cascade of four 1D convolutional layers (whose number of filters starts with 32 and increases by 32 in each layer), followed by one recurrent layer, for which we test LSTM and GRU architectures with hidden state size of 256. As image encoder, we test different architectures for the internal feature extractor, namely, ResNet-101, DenseNet-161, Inception-v3, and AlexNet; all of them are pre-trained on ImageNet and fine-tuned during our siamese network training. We also perform data augmentation by generating multiple crops for an image associated to a given EEG sample. In particular, we resize each image by a factor of 1.1 with respect to the image encoder's expected input size (299×299 for Inception-v3, 224×224 for the others), then extract ten crops from the four corners and the center of the input image, with corresponding horizontal flips. The resulting dataset is akin to the one employed in [51], where multiple text descriptions are available for each image. The size of the joint embedding space is set to 128.

Once training is completed, we use the trained EEG and image encoders as feature extractors in the joint embedding space, followed by a softmax layer, for both image and EEG classification. The classification tasks, beside providing a way to assess the quality of our multimodal learning approach, are used to identify the best encoders' layouts, based on the accuracy on the validation set.

The specific values for number of convolutional layer, layer sizes, number of filters, manifold size are the ones giving the best validation performance in our experiments.

Table 1 shows the test classification accuracy for all the tested models. All configurations are able to successfully perform EEG classification, with a slight accuracy increase associated to the use of GRU in the EEG encoder. Image classification accuracy is also high, except for the configurations using AlexNet, which suffers in comparison to more recent models.

Afterwards, we investigate the impact that one modality has for classification in the other modality domain, i.e., the effect of integrating brain activity–derived features into visual features and vice versa. In particular, we first compare the image classification performance obtained by our approach to the performance of pre-trained image encoders alone. Both our model and pre-trained visual encoders are used as feature extractors followed by a softmax layer and performance are computed on the test split of the employed visual dataset. As visual encoders we use the ones giving the best performance according to Tab. 1. The results in Tab. 2 indicate that learning features that maximize EEG-visual correlation (as discussed in Sect. 3) leads to enhanced performance for all models. The largest increase is found when using AlexNet in the image encoder, which is due to the fact that the other models are complex enough to "saturate" the classification capacity, while the lower accuracy by AlexNet allows for improvement by integrating the information coming from the neural activity.

Analogously, we compare the performance in terms of EEG signal classification accuracy between the EEG encoder

Image encoder	EEG	EEG Accuracy	Image Accuracy	Average Accuracy
Inception-v3	LSTM	90.1%	93.6%	91.9
Inception-v3	**GRU**	**90.4%**	**94.7%**	**93.0**
ResNet-101	LSTM	90.7%	91.2%	91.0
ResNet-101	GRU	92.3%	91.5%	91.9
DenseNet-161	LSTM	92.4%	92.3%	92.4
DenseNet-161	GRU	93.7%	91.8%	92.8
AlexNet	LSTM	85.6%	70.1%	77.8
AlexNet	GRU	77.2%	69.9%	73.6

TABLE 1: EEG and image classification accuracies for different layouts of the EEG and image encoders.

Model	Visual Feature Learning	Joint Learning with EEG
AlexNet	65.5 %	70.1 %
Inception-v3	93.1 %	94.7 %
ResNet-101	90.3 %	91.5 %
DenseNet-161	91.4 %	92.3 %

TABLE 2: Comparison of image classification performance when using only visual features and when using joint neural-visual features. For each model, we report the best performance according to Tab. 1

in [3] (using both LSTM and GRU for fairer comparison) and the one obtained when enriching brain-derived features with visual ones. The results are given in Table 3, showing that including visual features in EEG classification improves performance. By comparing Tab. 2 and 3, it can be noted that EEG classification benefits more from the the integration of the two modalities than image classification. This is not surprising, given the noisy and mostly-unexplored nature of neural activity data: in this case, the integration of the more easily-classifiable visual features helps to "guide" the learning of a more discriminative representation.

7.3 Saliency detection

With the previous experiments, we demonstrated that the learned EEG/image embedding is able to encode enough visual information to perform both EEG and image classification. We then investigate if and how the shared visual-brain space relates to visual saliency information using the approach described in Sect. 4. In this particular evaluation, we employ the trained encoders (see Sect. 7.2) to measure how compatibility varies as different image patches are removed. Note that this evaluation does not require any additional training, but is based on the same EEG and image encoders as described in Sect. 7.2. For our evaluation, we use the following set of values for scale σ (defined in Sect. 4): 3, 5, 9, 17, 33, 65. Fig. 5 shows qualitatively the saliency maps obtained by our approach, compared to those achieved by state-of-the-art saliency detectors [52], [53].

We also quantitatively assess the accuracy of the maps generated by our joint-embedding–driven saliency detector. To this aim, we first built a saliency dataset using a 60-Hz Tobii T60 eye-tracker, by having six human subjects participate to a free-viewing experiment (participants were asked to look at the displayed images). The dataset consists of the same set of 2,000 images from the brain activity dataset described in Sect. 7.1, divided into the same training, validation and test splits as used for training the encoders. As baselines for comparison we use pre-trained SALICON [52] and SalNet [53] models, fine-tuned on the dataset's training set. We evaluate the performance on the dataset's test set, employing the metrics defined by [54] — shuffled area under curve (s-AUC), normalized scanpath saliency (NSS) and correlation coefficient (CC) scores. In addition, in order to demonstrate that EEG indeed encode visual saliency information and that the generated maps are not simply driven by the image encoder, we include an additional baseline by implementing a similar approach to the one described in Sect. 4 using a pre-trained visual classifier (specifically, Inception-v3, as it gives better performance according to Table 1): we apply the same multi-scale patch-suppression method, with the difference that the saliency score is not based on compatibility, but on the log-likelihood variation for the image's correct class.

Tab. 4 reports the achieved results on our saliency dataset, showing that our method significantly outperforms the baseline saliency detectors. In addition, the results clearly show the role of brain features in computing visual saliency: employing the joint neural/visual features gives a significant performance boost compared to use of visual cues alone. This is also an important finding for cognitive neuroscience research, since we demonstrate for the first time that EEG recordings do encode visual saliency.

7.4 Decoding Brain Representations

The objective of this analysis is to approximate cortex-level representations of human visual pathway. Indeed, while the hierarchical multi-stage architecture of the human visual pathway is known, the representations generated at each stage are poorly understood. In particular, we perform both a coarse analysis on the global interaction between neural activity and images, and a fine analysis on the interaction between neural activity components and deep-learned visual features, in order to identify which neural areas focus attention and neural generators, i.e., visual stimuli that make specific neurons fire. Beside revealing the importance of some brain mechanisms, the following evaluations aim to further demonstrate that the learned embedding is indeed sensitive to brain activity relevant to visual cues.

EEG encoder	EEG Feature Learning	Joint Learning with visual data
LSTM	80.9%	90.1%
GRU	81.8%	90.4%

TABLE 3: Comparison of EEG classification performance when using only neural features and when using joint neural-visual features. The reported EEG classifcation performance for our approach are those achieved when training the image encoder using Inception-v3, as it gave the highest average accuracy according to Tab. 1.

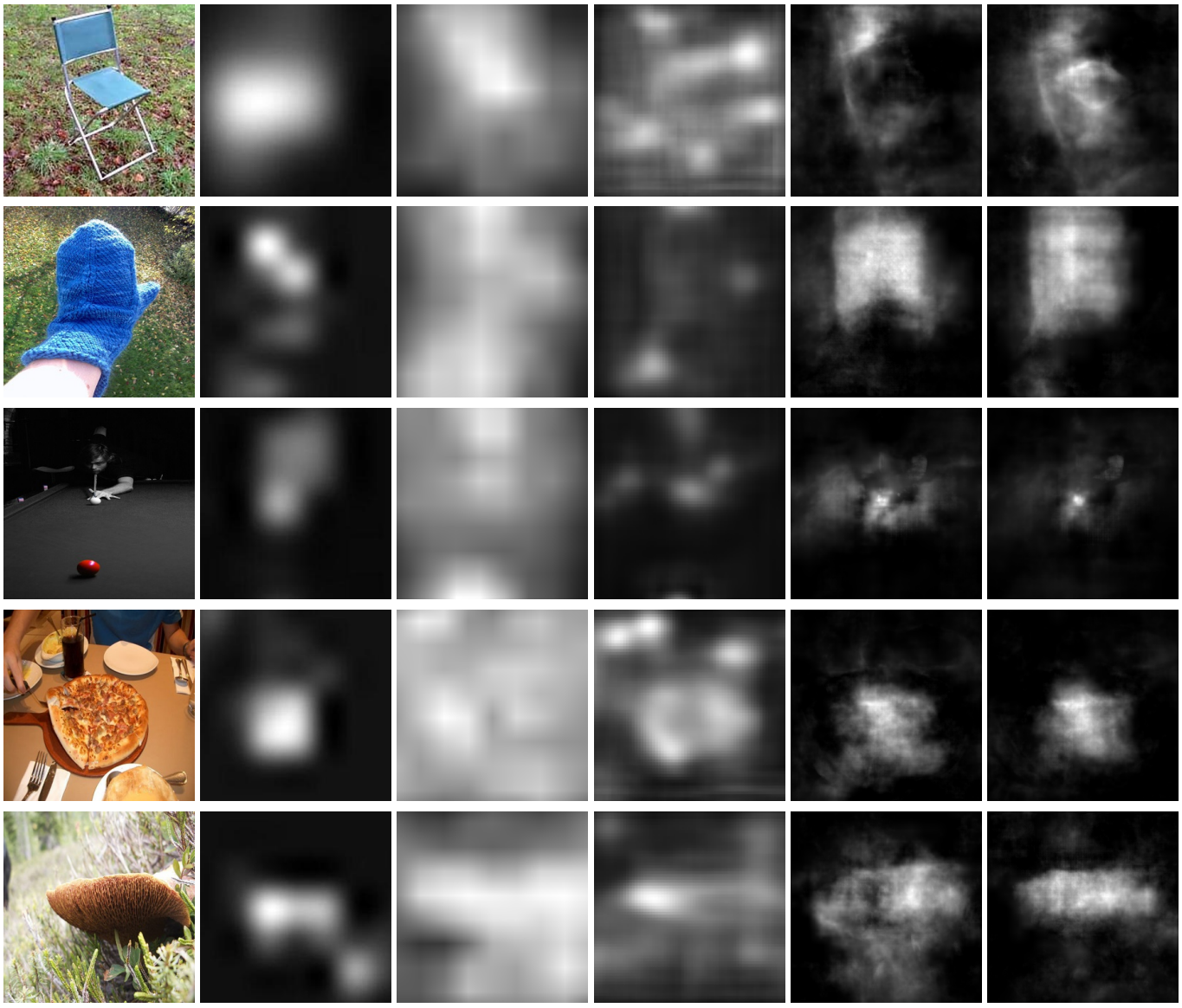

Fig. 5: **Qualitative comparison of generated saliency maps.** From left to right: input image, human gaze data (ground truth), SALICON, SalNet, visual classifier–driven detector, and our visual/EEG–driven detector. It can be noted a) that the maps generated by our method resemble more the ground truth masks than state-of-the-art methods; b) adding brain activity information to visual features results in an improved reconstruction (more details and less noise) of saliency (compare the 5^{th} and 6^{th} columns).

Method	s-AUC	NSS	CC
SalNet	0.637	0.618	0.271
SALICON	0.678	0.728	0.348
Visual classifier–driven detector	0.580	0.505	0.201
Our neural-driven detector	**0.692**	**1.061**	**0.378**
Human Baseline	0.939	3.042	1

TABLE 4: Saliency performance comparison in terms of shuffled area under curve (s-AUC), normalized scanpath saliency (NSS) and correlation coefficient (CC) between our compatibility-driven saliency detector and the baseline models. We also report the human baseline, i.e., the scores computed using the ground truth maps.

7.4.1 Global analysis of cortical-visual representations

The approach to perform this evaluation was presented in Sect. 6 and is based on measuring compatibility changes when important neural activity/visual features are removed.

The first experiment aims at identifying high-level correlations between EEG channels and visual content. To accomplish this, we apply Eq. 5, which assesses average compatibility changes by suppressing specific EEG channels. Fig. 6 shows the activation maps obtained by averaging channel importance scores over all images for each class. In order to show the relation between the temporal dynamics and spatial activation of EEG, Fig. 7 shows the average activation map over all classes and how cortical areas are activated by different parts of the input EEG signals. These maps are computed by exploiting the length-agnostic nature of the EEG encoder: instead of feeding the full-length signal to Eq. 5, we extract parts of EEG signals corresponding to a specific time interval and perform encoding and compatibility measurements on the resulting embedding.

Of course, differently from the saliency experiment, from which it was possible to assess quantitatively the achieved results, in this case there is no established comparison protocol that measures the importance of EEG components in visual analysis tasks. Nevertheless, some considerations can be drawn, which are consistent with cognitive neuroscience literature: 1) for all the visual classes, the most activated area is the V1 cortex (located in the occipital area), which is known to be responsible for early visual processing in the human brain [17]; 2) from the average activation maps in different time ranges, it can be noted that the processing starts in V1 and then flows to the frontal (responsible of higher cognitive functions) and temporal (responsible for visual categorization as suggested in [10]) cortices; 3) other activation areas fire according to the observed visual content; e.g., for the "piano" visual class, areas located in the auditory cortex are activated, and this is inline with evidence that the sensation of sounds is often associated with sight [55].

Finally, by combining the achieved activation maps with the saliency maps previously computed, we can obtain retinotopic saliency maps, i.e., maps associating neural activity to the most salient visual patches. Examples of these maps are given in Fig. 8. For each input image, we show a triple consisting of: input image, saliency map obtained through the approach described in Sect. 4 and neural activation map regions that mostly affect the learned embedding between images and EEGs, averaged over all participants.

7.4.2 Local analysis of cortical-visual data for neural generator identification

The previous evaluation analyzed cortical-visual data at a global scale and led to the identification of correlations between image patches and brain activation areas. While this information is useful for uncovering brain mechanisms, approximations of visual representations through the entire visual pathway would provide a deeper look into the finer processing stages underpinning visual perception. Under this perspective, the following evaluations aim at identifying local neural generators, i.e., low- and middle-level visual stimuli responsible for activations of specific cortex areas. To accomplish this task, we employ the learned compatibility measure to find a mutual correspondence between deep features and specific cortical areas. Using the *association* score defined in Sect. 6 (Eq. 10), we investigate the encoding of visual information in primate brains by deriving neural activation maps that maximally respond to the deep-learned visual features. Fig. 9 shows the activation maps of the association scores related to specific layers of our best-performing image encoder as per Tab. 1, that internally employs a pre-trained Inception network fine-tuned on our brain/image dataset during encoder training. In order to give an idea of the complexity of features learned at each level, we show a few examples obtained by performing activation maximization [56] on a subset of features for each layer. For each feature/neural association, we also measure the relative contribution to brain activity by different temporal portions of the EEG, by feeding each interval to the EEG encoder when applying Eq. 10. In this case, unlike the representations in Fig. 7, we are not interested in the differences in activations between cortical regions, so we compute the average unnormalized association scores over all channels, and use it as a measurement of how much each portion of the EEG affects the association to a layer's features. This multiple information allows us to understand neural generators and their timing. The results suggest that hierarchical representations in DCNNs have a tight correlation with hierarchical processing stages in the human visual pathway. In particular, at the lowest layer simple texture and color features are generated and they have a correspondence with the V1 cortex. Moving to deeper layers in a DCNN, we can notice that the activation propagates from the V1 cortex to the temporal one and back to the early visual processing cortex.

Moreover, more complex features (at higher layers) seem to be more influenced by the later temporal dynamics in brain activity, while simple features are more affected by processes happening just after stimulus onset. These findings are consistent with the cognitive neuroscience literature investigating hierarchical processing and timing in primate visual brains and suggest reliable approximation of brain representations. It is interesting to notice a consistent drop in the contribution to activation related to the time portion between 100-200 ms, and a following increase: though a comprehensive neurological interpretation is outside the scope of this paper, this may be due to a relocation of visual cognitive processes to deeper cortical areas that are

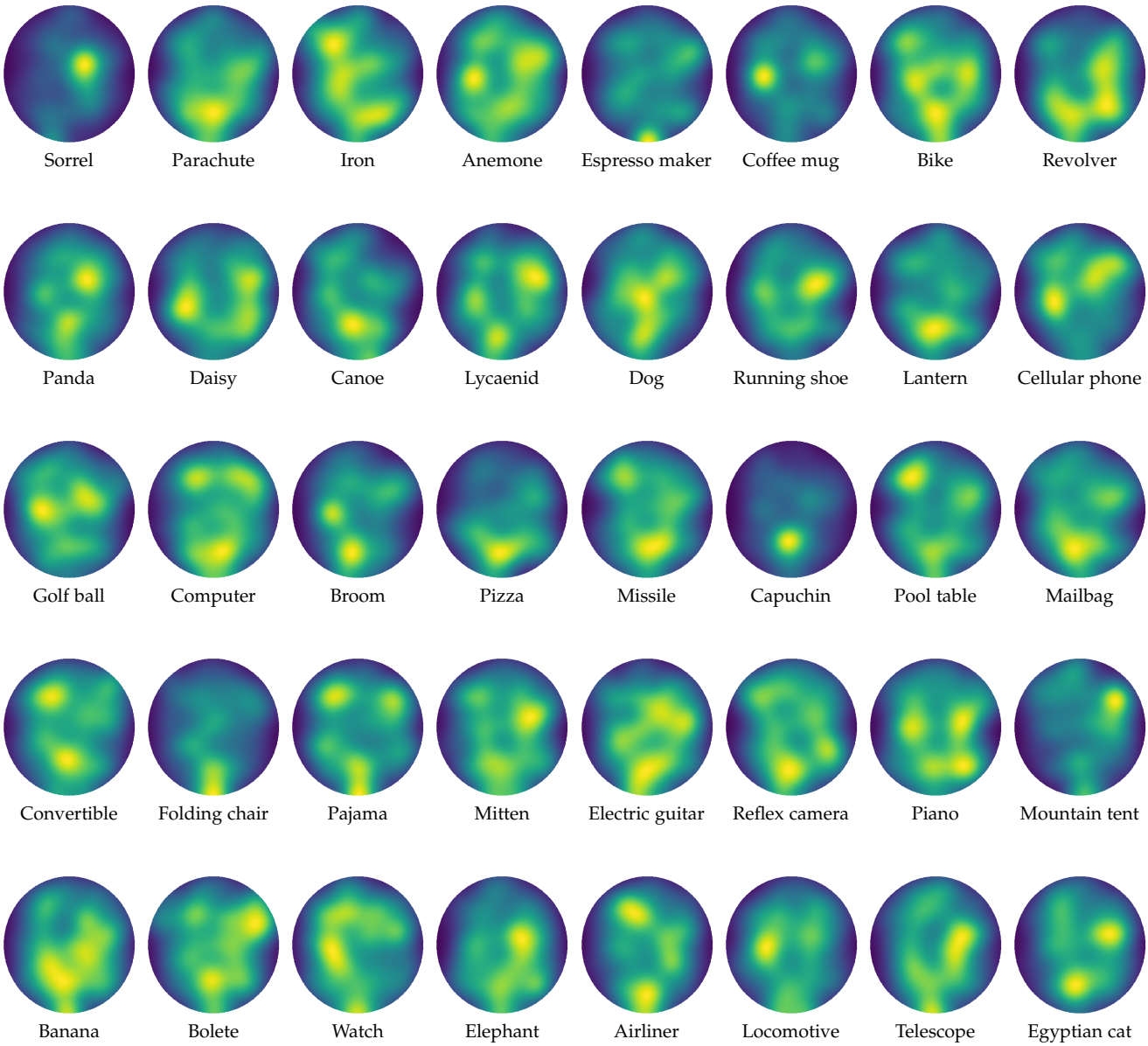

Fig. 6: **Activation maps per visual class**. Average activation maps for each of the 40 visual classes in the dataset.

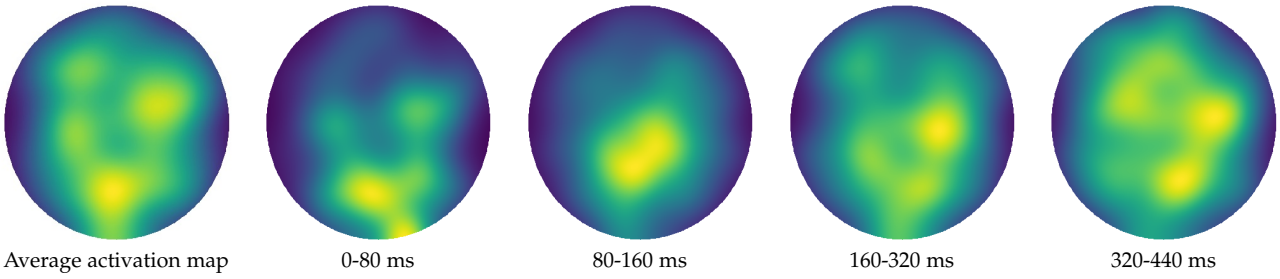

Fig. 7: **Average activation maps**. (Left image). Average activation map across all image classes. (Right images). Average activation in different time ranges.

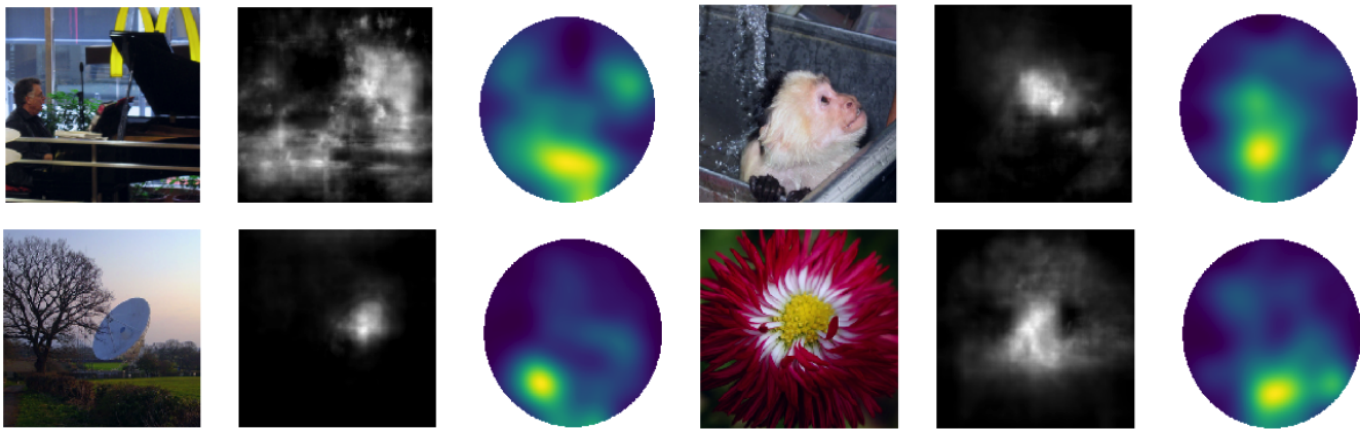

Fig. 8: **Retinotopic saliency maps derived by analyzing the learned embedding on different brain regions**. For each triplet of images: (left) image shown to the subject; (middle) most significative visual features; (right) brain activation areas.

less detectable via EEG, followed by a feedback activity to initial regions in the visual pathway.

8 Conclusion

In this work, we presented an approach for learning a joint feature space for images and EEG signals recorded while users look at pictures on a screen, by training two encoders in a siamese configuration and maximizinz a *compatibility* score between corresponding images and EEGs. The learned embeddings make the representation useful to perform several computer vision tasks supervised by brain activity. Our experiments, in particular, show how reliably neural activity can be used to enhance performance of image classification and saliency detection. Indeed, beside advancing our previous work on brain-guided image classification [3], we developed the first ever saliency detection approach supervised by neural activity, which, also, provides a very useful insight from a neurocognitive perspective, i.e., that EEG recordings encode visual attention information.

We also performed two experiments that verify that the representation does learn to identify the most important visual features and EEG signal components that correspond to the related processes in the brain. The second experiment is the most interesting one from the perspective of understanding the way brain activity signals encode information. Indeed, our approach was able to generate retinotopic maps by combining visual stimuli and brain activity through artificial intelligence methods.

While drawing general conclusions on these findings is not the main goal of this work and would require a deeper and more extensive evaluation, to the best of our knowledge, this is the first work that suggests reliable approximations of brain representations and their localization by jointly learning a model that maximizes the correlation between neural activity and visual cues.

The natural direction for future work leads to further investigation of these associations, with the objective of finding a finer correspondence between EEG signals and visual patterns — e.g., by identifying different responses in brain activity corresponding to specific objects or patterns. We believe that a joint research effort combining artificial intelligence and neuroscience is necessary to advance both fields, by studying how brain processes relate to artificial model structures and, in turn, applying the uncovered neural dynamics to models that move closer to human perceptual and cognitive performance.

Acknowledgments

The authors would like to thank Dr. Martina Platania for supporting the data acquisition phase and NVIDIA for the generous donation of two Titan X GPUs.

References

[1] T. Horikawa and Y. Kamitani, "Generic decoding of seen and imagined objects using hierarchical visual features," *Nat Commun*, vol. 8, p. 15037, May 2017.

[2] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Sci Rep*, vol. 6, p. 27755, 06 2016.

[3] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep Learning Human Mind for Automated Visual Classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jul 2017, pp. 4503–4511.

[4] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Generative adversarial networks conditioned by brain signals," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 3430–3438.

[5] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Curr. Biol.*, vol. 21, no. 19, pp. 1641–1646, Oct 2011.

[6] D. E. Stansbury, T. Naselaris, and J. L. Gallant, "Natural scene statistics account for the representation of scene categories in human visual cortex," *Neuron*, vol. 79, no. 5, pp. 1025–1034, Sep 2013.

[7] J. Bullier, "Integrated model of visual processing," *Brain Res. Brain Res. Rev.*, vol. 36, no. 2-3, pp. 96–107, Oct 2001.

[8] Z. Kourtzi and C. E. Connor, "Neural representations for object perception: structure, category, and adaptive coding," *Annu. Rev. Neurosci.*, vol. 34, pp. 45–67, 2011.

[9] D. J. Kravitz, K. S. Saleem, C. I. Baker, and M. Mishkin, "A new neural framework for visuospatial processing," *Nat. Rev. Neurosci.*, vol. 12, no. 4, pp. 217–230, Apr 2011.

[10] 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.

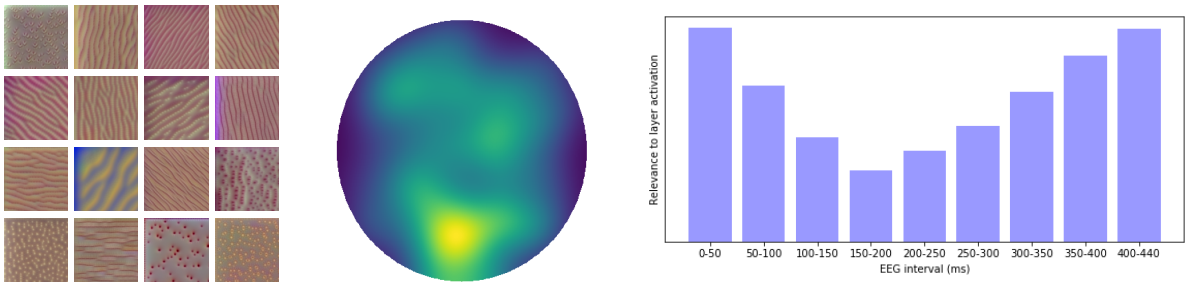

Image encoder, layer 3/20

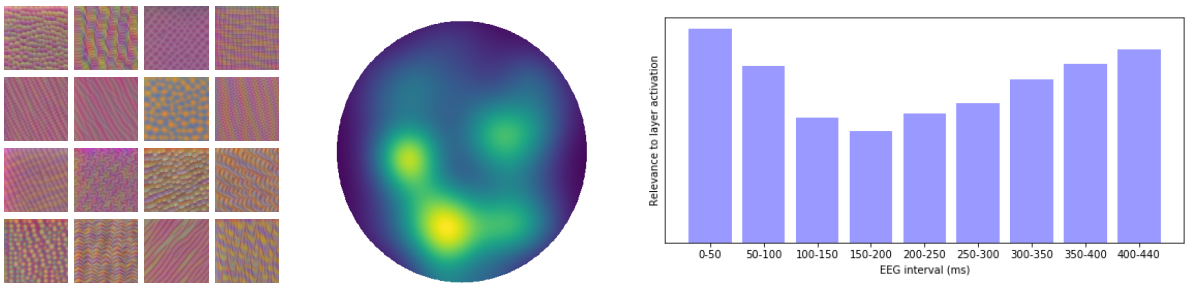

Image encoder, layer 7/20

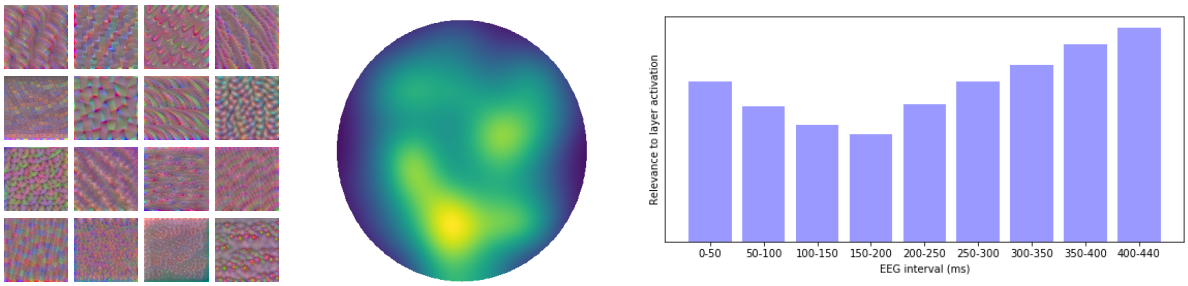

Image encoder, layer 12/20

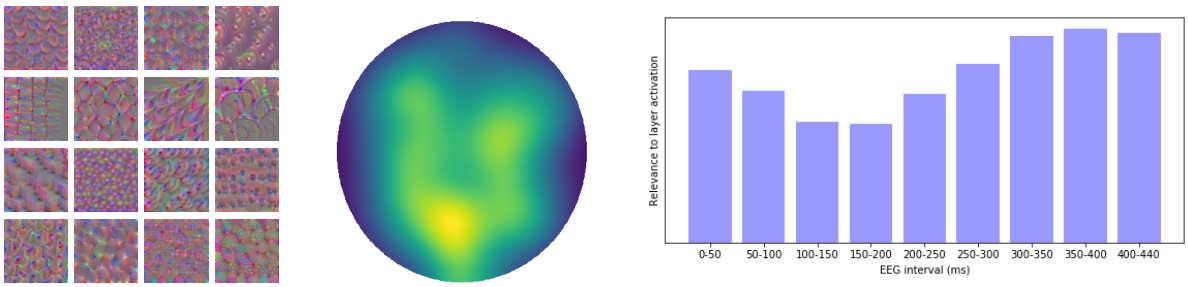

Image encoder, layer 17/20

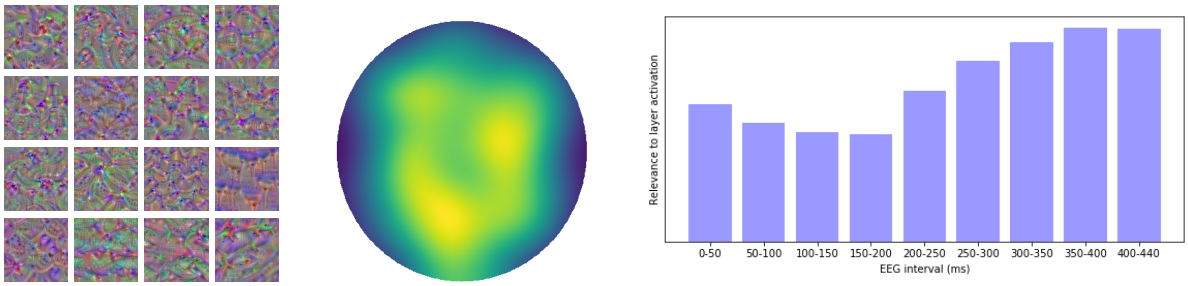

Image encoder, layer 20/20

Fig. 9: **Brain activity localization in association to specific visual representation levels**. Each row shows a set of feature maps (manually picked for interpretability and visualized through activation maximization) from a specific layer in the image encoder, the neural activity areas with the highest association to the layer's features, and the contribution that different time ranges in the EEG signals give to association scores. It can be noted that, as feature complexity increases, the activated brain regions move from the V1 visual cortex (occipital region) to the IT cortex (temporal region); moreover, the initial temporal portions of EEG signals seem to be more related to simpler features, while there is a stronger association between more complex features and later temporal dynamics.

[11] H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, and Z. Liu, "Deep predictive coding network for object recognition," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmssan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 5266–5275. [Online]. Available: http://proceedings.mlr.press/v80/wen18a.html

[12] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behav Brain Sci*, vol. 36, no. 3, pp. 181–204, Jun 2013.

[13] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical microcircuits for predictive coding," *Neuron*, vol. 76, no. 4, pp. 695–711, Nov 2012.

[14] K. J. Seymour, M. A. Williams, and A. N. Rich, "The Representation of Color across the Human Visual Cortex: Distinguishing Chromatic Signals Contributing to Object Form Versus Surface Color," *Cereb. Cortex*, vol. 26, no. 5, pp. 1997–2005, May 2016.

[15] J. W. Peirce, "Understanding mid-level representations in visual processing," *J Vis*, vol. 15, no. 7, p. 5, 2015.

[16] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo, "Fast readout of object identity from macaque inferior temporal cortex," *Science*, vol. 310, no. 5749, pp. 863–866, Nov 2005.

[17] A. K. Robinson, P. Venkatesh, M. J. Boring, M. J. Tarr, P. Grover, and M. Behrmann, "Very high density EEG elucidates spatiotemporal aspects of early visual processing," *Sci Rep*, vol. 7, no. 1, p. 16248, Nov 2017.

[18] D. L. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo, "Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3093–3101. [Online].

[19] D. L. 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," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 111, no. 23, pp. 8619–8624, Jun 2014.

[20] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis - connecting the branches of systems neuroscience," *Front Syst Neurosci*, vol. 2, p. 4, 2008.

[21] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwi?ska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 10 2016.

[22] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1462–1471. [Online]. Available: http://proceedings.mlr.press/v37/gregor15.html

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057. [Online]. Available: http://proceedings.mlr.press/v37/xuc15.html

[24] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav Brain Sci*, vol. 40, p. e253, Jan 2017.

[25] R. C. Fong, W. J. Scheirer, and D. D. Cox, "Using human brain activity to guide machine learning," *Sci Rep*, vol. 8, no. 1, p. 5397, Mar 2018.

[26] A. Kapoor, P. Shenoy, and D. Tan, "Combining brain computer interfaces with vision for object categorization," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[27] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox, "Perceptual Annotation: Measuring Human Vision to Improve Computer Vision," *IEEE Trans Pattern Anal Mach Intell*, vol. 36, no. 8, pp. 1679–1686, Aug 2014.

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning." in *ICML*, L. Getoor and T. Scheffer, Eds. Omnipress, 2011, pp. 689–696.

[29] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2141–2149. [Online]. Available: http://papers.nips.cc/paper/5279-improved-multimodal-deep-learning-with-variation-of-information.pdf

[30] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14b.html

[31] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] I. Ilievski and J. Feng, "Multimodal learning and reasoning for visual question answering," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 551–562. [Online]. Available: http://papers.nips.cc/paper/6658-multimodal-learning-and-reasoning-for-visual-question-answering.pdf

[33] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 902–909.

[34] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," *arXiv preprint arXiv:1804.03160*, 2018.

[35] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 801–816.

[36] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. USA: Curran Associates Inc., 2016, pp. 892–900. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157196

[37] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 131–135.

[38] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative," in *Proceedings of the International Conference on Multimedia Information Retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 527–536. [Online]. Available: http://doi.acm.org/10.1145/1743384.1743475

[39] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition*, 2015. [Online]. Available: http://arxiv.org/abs/1411.4555

[41] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, 2017.

[42] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description." in *CVPR*. IEEE Computer Society, 2015, pp. 2625–2634.

[43] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1060–1069. [Online]. Available: http://proceedings.mlr.press/v48/reed16.html

[44] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov, "Generating images from captions with attention," *ICLR2016*, vol. abs/1511.02793, 2016.

[45] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," in *Proceedings of the 2017 ACM on Multimedia Conference*, ser. MM '17. New York, NY, USA: ACM, 2017, pp. 1809–1817. [Online]. Available: http://doi.acm.org/10.1145/3123266.3127907

[46] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *ICLR2016*, vol. abs/1511.07122, 2015.

[47] S. Treue, "Visual attention: the where, what, how and why of saliency," *Current Opinion in Neurobiology*, vol. 13, no. 4, pp. 428 – 432, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438803001053

[48] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[50] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clin Neurophysiol*, vol. 112, no. 4, pp. 713–719, Apr 2001.

[51] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[52] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *ICCV 2015*, 2015, pp. 262–270.

[53] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *CVPR 2016*, 2016.

[54] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *TIP 2013*, 2013.

[55] A. M. Proverbio, G. E. D'Aniello, R. Adorni, and A. Zani, "When a photograph can be heard: vision activates the auditory cortex within 110 ms," *Sci Rep*, vol. 1, p. 54, 2011.

[56] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.