RESEARCH ARTICLE

A zero-shot learning approach to the development of brain-computer interfaces for image retrieval

Ben McCartney[1]*, **Jesus Martinez-del-Rincon**[1], **Barry Devereux**[1], **Brian Murphy**[1,2]

1 Queen's University Belfast, United Kingdom, **2** BrainWaveBank Ltd. Belfast, United Kingdom

* bmccartney06@qub.ac.uk

Abstract

Brain decoding—the process of inferring a person's momentary cognitive state from their brain activity—has enormous potential in the field of human-computer interaction. In this study we propose a zero-shot EEG-to-image brain decoding approach which makes use of state-of-the-art EEG preprocessing and feature selection methods, and which maps EEG activity to biologically inspired computer vision and linguistic models. We apply this approach to solve the problem of identifying viewed images from recorded brain activity in a reliable and scalable way. We demonstrate competitive decoding accuracies across two EEG datasets, using a zero-shot learning framework more applicable to real-world image retrieval than traditional classification techniques.

Introduction

Research in the field of Brain-Computer Interfaces (BCI) began in the 1970s [1] with the aim of providing a new, intuitive, and rich method of communication between computer systems and their users. Typically, these methods involve measuring some aspect of neural activity and inferring or decoding an intended action or particular characteristic of the user's cognitive state. Although BCI is still in its infancy, there are already practical applications in assistive technology as well as in disease diagnosis [2, 3]. Brain-controlled prosthetics [4] and spellers [5] have shown their potential to enable natural interaction in comparison with more traditional methods, such as mechanical prosthetics or eye-movement-based spellers. Other relevant applications include identifying the image that a user is viewing, usually referred to as image retrieval, which is of particular interest in the fields of visual attention applied to advertising and marketing, in searching and organising large collections of images, and in reducing distractions during driving, to name a few.

Although brain decoding technology has immense potential in diverse applications, it faces multiple challenges related to speed and accuracy that must be overcome before it emerges as a disruptive technology. The complexity of BCI stems from the naturally low signal-to-noise ratio (SNR) and high dimensionality of raw brain data, which often complicates automated analysis and can force researchers to manually analyse previously recorded neural activation data. In the context of data collected from electroencephalography (EEG) devices, this is

Citation: McCartney B, Martinez-del-Rincon J, Devereux B, Murphy B (2019) A zero-shot learning approach to the development of brain-computer interfaces for image retrieval. PLoS ONE 14(9): e0214342. https://doi.org/10.1371/journal.pone.0214342

Editor: Simone Borsci, Universiteit Twente, NETHERLANDS

Received: February 27, 2019

Accepted: August 29, 2019

Published: September 16, 2019

Data Availability Statement: The first dataset underlying the results presented in the study is third party data, not owned or collected by the authors, and is available by contacting brian.murphy@qub.ac.uk. The authors had no special access or access privileges that others would not have. The second dataset underlying the results presented in the study is available from https://exhibits.stanford.edu/data/catalog/tc919dd5388.

Funding: The authors received no specific funding for this work. However, during the course of this study, one of the authors (Brian Murphy) has been under the employ of BrainWaveBank Ltd. as CSO. The funder provided support in the form of salaries for author B. Murphy, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the author contributions section.

Competing interests: During the course of this study one of the authors (Brian Murphy) has been under the employ of BrainWaveBank Ltd. as CSO. Although Brian is serving as the point of contact for access to one of the EEG datasets we analyse in this study, this dataset was created prior to the founding of BrainWaveBank and as such the funder has no association with the dataset. At this time Brian was under the employ of the University of Trento. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

typically done either by examining the frequency domain or by plotting Event-Related Potentials (ERPs). In an ERP experiment the participant is presented several times with a given stimulus or stimulus category and their neural response each time can be recorded and averaged. These ERPs can be analysed against well-known response patterns or, alternatively, characteristics such as the strength and timing of signal peaks can be quantified and analysed automatically. ERP analysis is well established and has important applications in medical diagnosis [6] and in cognitive neuroscience research [7, 8]; however, the broad characterisation of brain response used in traditional ERP methods is not richly informative enough to decode the level of detail required to make predictions about a participant's cognitive state, as required for BCI image identification.

Given the complexities of decoding the nature of an arbitrary visual stimulus from a person's brain activity, cognitive neuroscientists and BCI researchers have traditionally tackled the simpler task of determining which of some finite set of category labels corresponds to a particular pattern of brain activity. In one of the first such studies, Haxby and colleagues [9] collected functional Magnetic Resonance Imaging (fMRI) data as participants viewed a series of images from the categories of *human faces*, *cats*, *houses*, *chairs*, *scissors*, *shoes* and *bottles*, along with images of *random noise*. The researchers were able to determine with 83% accuracy which category of object the participant was viewing.

However, fMRI is impractical for general BCI applications. Murphy et al. [10] used EEG rather than fMRI and achieved 72% accuracy in classification across the two classes of *mammals* and *tools*. While this study addressed a much simpler problem with only two possible classes, it demonstrated category decoding using relatively inexpensive and less intrusive EEG data collection methods (fMRI and EEG technologies are discussed in more detail in Section 'Brain Data').

In the studies mentioned above the classifiers would not determine specifically which stimulus image was displayed (as required for image retrieval), instead they only determine the category which the stimulus image belongs to. Moreover, as a classification approach, this is not scalable to new classes and, although it may yield a high accuracy, it becomes less accurate with increasing number of classes. An alternative approach to BCI image classification makes use of rapid serial visual presentation (RSVP) [11, 12]. The participant is presented with a rapid stream of images (approximately 10 each second) and is instructed to count the number of times a particular target image or object appears. A classifier can then reliably decode whether for a given segment of brain data, the participant had been presented with a target or non-target image. This RSVP approach could be more directly applied to our problem by showing a participant a target image from a gallery, and then presenting all of the images in a gallery one by one with the expectation that when the target image should elicit neural activity sufficiently different from the non-target images to identify it. However, as the number of images in the gallery grows, it becomes impractical to present them in a real-world searching scenario.

Traditional machine learning approaches can achieve high accuracy in classification tasks when there is sufficient training samples for each class for a specified set of classes. However, these approaches are not suitable when samples can belong to a novel class which has not been seen before in training, or as the number of classes tends to infinity. This characteristic can be too restrictive for many real-world applications where gathering extensive training data is impractical. Zero-shot learning presents a more scalable approach to brain decoding, where the performance of the system is sustained as the number of new classes and instances scales up. Rather than learning a mapping directly to pre-defined classes or labels, we focus on creating an embedding space which could describe any valid class, and then learning a mapping between neural activation data and this embedding space. Such a mapping can be defined with a subset of the full set of classes and/or instances, and tested using withheld classes/instances. With this approach, the system could in theory decode arbitrary stimulus images it has not yet been exposed to.

Introducing a feature-based model comes at a cost however, as it also impacts the overall accuracy of the system. At present most zero-shot systems in this area [13, 14] exhibit performance insufficient for real-world applications.

Following the work in Haxby et al. [9], Mitchell et al. [13] used fMRI to decode the meanings of nouns corresponding to concrete objects, using as features each noun's textual co-occurrence frequency with a set of 25 verbs. This study was one of the first to make use of zero-shot learning, allowing them to decode classes (i.e. nouns) from outside the training set. Others have used visual image features rather than semantic features to decode cognitive states associated with viewing stimulus images in a zero-shot framework [15]. Using a Gabor-based voxel decoder, Kay et al. [15] achieved accuracy of 51% and 32% among 1620 images in a single-trial identification task, for two distinct subjects. These and other studies, while using fMRI rather than EEG data, demonstrate the relevance of both semantic and visual information in image decoding.

In related work, Palatucci et al. [14] used a similar procedure to Mitchell et al. [13] but with EEG data. In a study by Carlson et al. [16] using visual images, Linear Discriminant Analysis (LDA) was used to determine the categories of objects presented to participants. The aim of this study was to map out the stages of object recognition by comparing the decoding accuracies across different levels of object representation and different time windows. They found that the peak decoding rate for distinguishing between images of the human body was 120ms after the image appeared, whereas the higher-level semantic distinction between animate and inanimate images was best determined after 240ms. Using a combination of low-level visual features and semantic features, Clarke et al. [17] demonstrated that decoding accuracy was significantly improved by the incorporation of the semantic features from around 200ms post-stimulus-onset. Similarly, Sudre et al. [18] also obtained high decoding accuracy using brain data using both visual and semantic feature sets. These studies both use magnetoencephalography (MEG), which is impractical for real-world BCI technology; however, their conclusions suggest that decoding of neural activation data using visual and semantic models can be a feasible approach to image decoding in a real-world BCI framework.

This project is motivated by the goal of designing a BCI system which can retrieve any arbitrary image specified by a neural activation generated by a user viewing that stimulus image. To this aim, this paper proposes an EEG zero-shot learning framework for individual image retrieval which makes full use of both advanced visual and semantic image features [16–18], rather than a single feature type [13–15]. We also apply the correlation-based feature extraction method presented in Mitchell et al. [13]. We evaluate our framework with the datasets recorded in two studies [10, 19], but also apply the challenging zero-shot learning restriction. Given our image retrieval application, we also aim to perform a more difficult task: where Carlson et al. [16] utilises zero-shot learning only to determine membership of the stimulus to a particular object category, our approach will aim to determine which actual image was viewed. Among the previously described studies, only Palatucci et al. [14] uses EEG data to decode images within a zero-shot framework but limited their evaluation to simplistic line drawings rather than photographic images as in this study.

The main contributions of this paper are:

- First time visual and semantic features are used together for EEG zero-shot learning, which translates to potential for a real-world BCI image retrieval system.

- State-of-the-art performance for the particular task of EEG-driven image retrieval in a zero-shot framework.
- Evaluation across two datasets from different sources including a large open dataset for future comparative studies.
- Analysis of how well the feature sets chosen reflect the expected brain activity.

General methodology

Our framework comprises of three main components. First the brain data must be cleaned and a subset of the EEG features extracted to represent the underlying cognitive states. Then we apply our chosen computer vision and semantic models to the stimuli, to create a representation of each image in this visuo-semantic feature space. Finally we use a linear regression algorithm to find a mapping between the brain and stimulus spaces which makes the brain decoding possible. A high-level overview of this architecture can be found in Fig 1.

Brain data

Two of the most widely used approaches to recording brain activity are functional Magnetic Resonance Imaging (fMRI) and EEG. The former can localize the physical source of brain activity with high spatial accuracy. However, the temporal resolution of fMRI is limited to a sampling rate of 1-2 seconds. Moreover, fMRI requires an MRI scanner, a large and expensive piece of equipment using powerful magnetic fields and liquid helium coolant, making it unsuitable for BCI systems outside of laboratory or clinical settings.

As a cheaper and more convenient alternative, EEG can be used to measure the electrical activity produced in the brain. Given that we are interested in eliciting cognitive states associated with particular images, the experimental paradigms used for the EEG data in this study

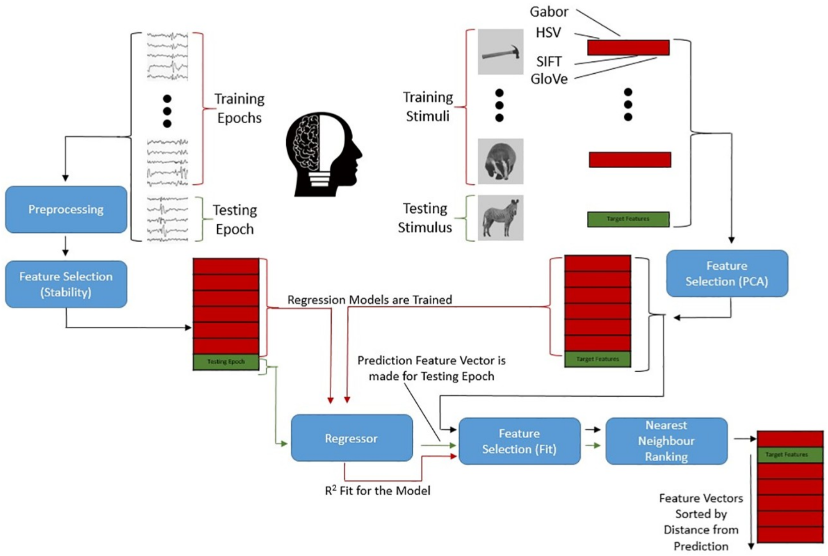


Fig 1. Information flow in the image retrieval architecture. Overview of the flow of information and processing during a single fold of cross-validation ('Zero-shot Prediction' Section). Model performance is determined by fit of the predicted feature vectors: in the example above, the true target features are in the second position of a sorted list of neighbours. In this case with a total of seven possible images, this results in a rank of 2, and a CMC AUC of 78.57% (Cumulative Match Curve, Area Under Curve; see 'Measure of Accuracy' Section for details).

https://doi.org/10.1371/journal.pone.0214342.g001

involve repeated presentations of images on a computer screen ('Datasets' Section). For each image presentation, an "epoch" consisting of 1000ms of EEG data starting at the onset of the stimulus presentation is extracted, which comfortably encompasses the informative brain activity associated with the image [16]. We refer to an epoch of EEG data and its associated stimulus image collectively as a "trial". Our goal is to map epoch data to image features in order to determine which image was presented at the time the epoch was recorded.

EEG data preprocessing. Preprocessing is a necessary stage of EEG data analysis that involves aligning, normalising and otherwise cleaning the raw data in order to make it more suitable for downstream analyses. The main goal of preprocessing the EEG data in our framework is to remove sources of noise in order to minimise obfuscation of underlying useful patterns in the data. Recordings are first filtered to remove ambient interference. One of the strongest noise sources in EEG is ambient electrical activity near the recording equipment, such as personal computers, large lights, or improperly insulated wiring. These signals are relatively easy to separate from brain activity based on their frequency, typically 60hz or 50hz (in America and Europe respectively). A lower frequency cut-off must also be established to remove slower sources of noise—these are generally slow changes in the electrical profile of the scalp or sensors such as a gradual increase or decrease of perspiration leading to a change in conductivity. A band-pass filter was used to remove any signals in our data with a frequency outside the range 1-40hz as in other studies [10, 11, 14, 19–22].

Channels with poor contact with the scalp were then identified using the variation, mean correlation and Hurst exponent [23]. Each metric was calculated and any channel with a value three standard deviations outside the norm across all channels was marked as poor. These channels were then removed and replaced by values interpolated from the remaining clean channels. These interpolated values were generated via spherical spline [24] which made use of the values from clean channels along with their location in 3D-space to approximate a value at the location of each removed channel.

The 1000ms epochs were baselined [20, 22] by subtracting from each channel the average value of that channel from the 500ms prior to the image presentation. Epochs which fell outside the threshold for amplitude range, variance or channel deviation were removed as in other studies [11, 20, 21, 25]. Again these metrics were computed and compared with a threshold for standard deviation of three.

Following this, Independent Component Analysis (ICA) [26] was performed primarily to identify artefacts related to eye movement as in other studies [10, 19, 20]. In this step the input signal is decomposed into an approximation of its sources, and each source (or component) can be examined to determine if it has encapsulated the influence of some artefact or pattern of brain activity on the EEG recording. The process of converting a multi-channel signal into components is reversible in that the original input signal can be reconstructed given a complete set of the corresponding components. When we apply ICA to EEG for the purposes of artefact rejection, components which isolate artefacts such as eye movement can be withheld and the EEG reconstructed with the remaining components. So long as the components removed correctly isolate the ocular artefacts, the reconstructed EEG recording should show no more evidence of said artefacts. These components can be analysed by hand or automatically to determine whether they should be removed.

In this study, components were compared against a channel of the input signal which showed the strongest samples of ocular artefacts, components with a higher correlation to this channel were more likely to have isolated the artefacts. A threshold is set as above to mark components for removal if their correlation coefficients exceeded a z-score of three. In some datasets, ocular activity is recorded by a sensor placed very close to the eyes which should provide a clean example of ocular artefacts, however the datasets we analyse here did not have a dedicated sensor for this purpose. We instead used whatever sensor was closest to the eyes and while this is where ocular activity will be strongest, there is a risk that informative brain activity could be lost. While we are primarily interested in using ICA to remove eye movement, there are a number of other less common artefact sources which can be isolated in components. Thresholds were also set as above for spatial kurtosis, Hurst exponent and mean gradient to detect single electrode short duration offset, non-biological electrical signals [27], and high-frequency content respectively. Components which reached these thresholds were removed and the EEG signal reconstructed from the remaining components.

Next, within each epoch, channels were examined for short term artefacts using variance, median gradient, amplitude range and channel deviation. Channels identified as noisy within the bounds of the epoch were replaced by an interpolation from other nearby channels within that epoch. Epochs were also downsampled to a rate of 120hz as in other studies [11, 14, 19] to reduce dimensionality before machine learning is applied.

All of the above preprocessing steps were implemented using the EEG preprocessing toolkit FASTER (Fully Automated Statistical Thresholding for EEG artefact Rejection) [20].

As a final preprocessing step before the EEG data are used in our regression model, the data are z-scored (standardised). We primarily perform this step to ensure that the mean of the data is zero as this can simplify the parametrisation of our machine learning model. Z-scoring is done separately for each iteration of the cross validation, with the mean values for the transformation calculated using only the training samples and the transformation then applied to the training and testing samples to avoid any influence of the latter in the former.

EEG feature selection. After preprocessing, an EEG feature extraction process is used to continue reducing the dimensionality of the data by extracting the most discriminatory features from the preprocessed data, and further removing uninformative and noisy dimensions of the data. This facilitates the successful mapping of EEG data to our image feature space by extracting only those aspects of the EEG signal which are likely to be informative about the visual and semantic feature sets. Following the approaches used in Mitchell et al. [13] and evaluated in Caceres et al. [28], we ignore all but the features with the highest collinearity across presentations of the same stimulus on the screen. Concretely, the EEG data for a particular participant following preprocessing is a 3D-matrix of size nE × nC × nT, where nE is the number of epochs (i.e. the number of stimulus presentation events), nC is the number of channels (or sensors) in the EEG headset, and nT is the number of timepoints in an epoch (the number of times during an epoch that sensor values were recorded). In this work, we use an epoch length of one second and downsample the data to 120Hz, giving nT = 120. We treat the data from each timepoint and each sensor as a separate feature, giving a total of nC × nT candidate features for each epoch. In order to calculate feature collinearity, we reshape the nE × nC × nT data matrix to a 2D-matrix of size nE × (nC × nT), or, equivalently, (nS × nP) × nF where nS is the number of stimuli, nP is the number of times each stimulus was presented in a recording, and nF is the number of EEG features. We then transform this back into a 3D-matrix of shape nF × nP × nS and term this matrix *D*. *D* is therefore composed of a nP × nS feature matrix for each EEG feature *f*. To calculate a stability score for a feature, we measure the consistency of the feature across different presentations of the same stimulus—we calculate the Pearson correlation for each pair of rows in *D* and use the mean of these correlations as the stability score for that EEG feature *f*:

$$\text{Stability}(f) = \frac{1}{\text{nCom}} \sum_{i=1.}^{\text{nP}} \sum_{j=1, i!=j.}^{\text{nCom}} \cdot \frac{\text{cov}(D_{f,i,:}, D_{f,j,:})}{\sigma_{D_{f,i,:}} \sigma_{D_{f,j,:}}}$$

where σ_x is the standard deviation of x and

$$\mathrm{nCom} = \frac{\mathrm{nP}(\mathrm{nP} - 1)}{2}$$

In each iteration of our cross-validation, we calculate the stability of each EEG feature using the training set and select the most stable features for fitting the regression model. More detail of the cross-validation architecture can be found in the 'Zero-shot Prediction' Section.

Image feature space

Computer vision is the field of study devoted to designing algorithms to interpret digital images, so it is a natural place to look for an appropriate feature space. Different computer vision models extract features at different levels of abstraction, ranging from recognising simple lines or colours through to recognising objects. Previous research [16, 17, 29–32] shows that these levels of abstraction are evaluated sequentially in the human ventral visual processing stream. In light of these findings, we expect earlier EEG features to contain predominantly low-level visual information, with higher-level visual features being increasingly present in later EEG features. For maximal decoding performance, it is therefore essential to find a set of computer vision models which cover each level of abstraction that will be represented in the EEG features. Furthermore, we chose feature sets which are grounded in similar mechanics to human visual processing, under the rationale that these feature sets have the potential to best match with human brain activity.

Gabor filters. In order to model human edge and texture detection we chose to use Gabor Filter Banks as this well-established computer vision technique identifies visual edges in a very similar way to the lowest-level of human visual processing in cortical areas V1 and V2 [33–36]. A bank is comprised of a set of filters which each represent an edge with a particular orientation and spatial frequency, these filters can be used to identify where in an image there is a matching edge. The filter bank used here contains eight evenly spaced orientations (θ) and four standard deviation values (σ) ranging from two to five, resulting in a bank of 32 filters. The rest of the parameters were fixed at default with ksize = (31, 31), wavelength of the sinusoidal factor (λ) = 6.0, spatial aspect ratio (γ) = 0.5 and phase offset (ψ) = 0.

Each pixel co-ordinate in an image x, y is convolved with a Gabor filter described by the parameters above:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right)$$

where

$$x' = x \times \cos\theta + y \times \sin\theta \qquad y' = -x \times \sin\theta + y \times \cos\theta$$

Let Lϑ and Lσ denote the sets of parameter values defining the filter bank:

$$\mathrm{L}\vartheta = \left\{0, \frac{1}{7}\pi, \frac{2}{7}\pi, \ldots\right\} \qquad \mathrm{L}\sigma = \{2, 3, 4, 5\}$$

Each image in our feature set was convolved with every filter, and the result summed to generate a histogram of 32 dimensions v_{gabor} for each image:

$$v_{\text{gabor}}(\text{image}) = \begin{bmatrix} \sum_{x,y\in\text{grid}} g(x, y, \lambda, \mathrm{L}\vartheta_1, \psi, \mathrm{L}\sigma_1, \gamma) \\ \vdots \\ \sum_{x,y\in\text{grid}} g(x_1, y_1, \lambda, \mathrm{L}\vartheta_8, \psi, \mathrm{L}\sigma_1, \gamma) \\ \vdots \\ \sum_{x,y\in\text{grid}} g(x_1, y_1, \lambda, \mathrm{L}\vartheta_1, \psi, \mathrm{L}\sigma_2, \gamma) \\ \vdots \\ \sum_{x,y\in\text{grid}} g(x_1, y_1, \lambda, \mathrm{L}\vartheta_8, \psi, \mathrm{L}\sigma_4, \gamma) \end{bmatrix}^T$$

We stack the v_{gabor} vectors to create the final matrix of Gabor features for our image set:

$$x_{\text{gabor}} = \begin{pmatrix} v_{\text{gabor}}(\text{image}_1) \\ v_{\text{gabor}}(\text{image}_2) \\ \vdots \\ v_{\text{gabor}}(\text{image}_{\text{nS}}) \end{pmatrix}$$

Scale invariant feature transform. The brain is also sensitive to higher-level visual information which is not adequately captured by simple and spatially local Gabor Filters. In order to make use of higher-level visual processing in our system we chose to apply a prominent computer vision model which detects and describes keypoints in an image such as corners, spots and other simple geometric entities. Keypoint locations in the image are important in the model as they offer a degree of abstraction from the raw pixel values, as well as a degree of transformation invariance (i.e. a particular set of features in an image selected as keypoints will continue to be the keypoints even if the image is subjected to low-level transformations such as scaling and rotation). In a similar way, human visual processing involves constructing representations of the visual input which are invariant with respect to low-level transformations. Moreover, in the computer vision model the keypoints tend to reflect interesting or important components of the images, and such locations within the image may also be reflected in humans' visual attention during object processing (e.g. where a keypoint represents the edge or boundary between an object outline and the background).

Scale Invariant Feature Transform (SIFT) identifies these keypoints using difference of Gaussians and generates a descriptor for the pixel neighbourhood [37]. We then applied Visual Bag Of Words (VBOW) [38] to the extracted SIFT features. Bag of words is a well-known technique in information retrieval and natural language processing which allows for a more discriminative and consistent representation of the feature representation. To achieve this, the different feature components which are most common across different exemplars in a large calibration set are identified and collated to form a "codebook" or dictionary, a conceptual bag of words to be used to describe the input (in the context of image processing, the "words" are not literally words, but are a set of image feature types that can be used to represent images, analogous to how a document can be represented as a collection of word types). Those feature components that are uncommon, and therefore likely to be of little discriminative value or associated with noise in individual images, are discarded. Using this codebook, a histogram is generated detailing how often each feature component of the codebook appears in the image features. This histogram is what can be used to represent images in a much lower dimensional space, which is also invariant to spatial transformations. Since the histogram has the same length as the number of elements in the codebook, every exemplar is represented with a vector of the same length, which simplifies the subsequent stages of the machine learning pipeline.

Following this VBOW process, we generated SIFT keypoint descriptors for a large corpus of images and then selected the most informative descriptors to compile the codebook. As our goal is to create an approach applicable to real-world BCI systems, we require our image feature space to have the capacity to describe and discriminate arbitrary representative images. To this end, we used a SIFT codebook trained with a large, diverse corpus of images taken from ImageNet [39].

Each SIFT Descriptor represents a 4 × 4 grid around some keypoint in the image. Difference of Gaussian (DoG) is run on each segment of this grid and the results compiled into a histogram with eight bins. A SIFT descriptor is the resulting 8 × 4× 4 = 128 dimensional vector, indexed by *x*-*y*. Each element of the codebook is a SIFT Descriptor. K-means clustering was performed with a random subset of 10 million SIFT descriptors generated over the ImageNet corpus to produce 1000 clusters. The centroid of each cluster was then taken to produce a codebook of 1000 dimensions to categorise future SIFT descriptors.

Using VBOW has the benefit of finding features that generalise well across multiple different objects and as such have the best chance of extending to new classes. Moreover, it removes spatial data making the feature vector invariant to spatial transformations such as rotation, translation and scale which is less relevant to intermediate-level visual information. A list of imageDescriptors were generated for an image, and used to produce a histogram v_{sift} of how often each 'visual word' encoded in the codebook appeared in the stimulus image.

$$v_{\text{sift}}(\text{image}) = [\text{hs}_1, \text{hs}_2, \ldots, \text{hs}_{1000}]$$

where

$$\text{hs}_i = \sum_{k \in \text{imageDescriptors}} I(k = \text{codebook}_i)$$

and I is an indicator function evaluating to 1 if the argument is true and 0 otherwise.

This implementation made use of Dense SIFT, meaning the keypoints correspond to a regularly sampled grid, rather than a set of natural keypoints estimated for an image. A histogram v_{sift} was generated for each image, and collated into a matrix representing our stimulus image SIFT features x_{sift}.

$$x_{\text{sift}} = \begin{bmatrix} v_{\text{sift}}(\text{image}_1) \\ v_{\text{sift}}(\text{image}_2) \\ \vdots \\ v_{\text{sift}}(\text{image}_{nS}) \end{bmatrix}$$

Colour histogram. Finally, as none of the previous visual features encapsulates colour information, we chose a global Hue, Saturation, and Value (HSV) histogram to model colour in our approach, since there is some evidence that a HSV colour space comes closer to reflecting human vision than Red, Green, and Blue (RGB) [40]. A HSV histogram v_{hsv} is generated for each image using a quantisation of four bits per pixel and channel:

$$v_{\text{hsv}}(\text{image}) = [h_1, h_2, \ldots, h_{16}, s_1, s_2, \ldots, s_{16}, v_1, v_2, \ldots, v_{16}]$$

$$h_i = \sum_{j=1}^{16} \sum_{k \in \text{iP}} I(k_{\text{hue}} = j) \qquad s_i = \sum_{j=1}^{16} \sum_{k \in \text{iP}} I(k_{\text{sat}} = j) \qquad v_i = \sum_{j=1}^{16} \sum_{k \in \text{iP}} I(k_{\text{value}} = j)$$

Where iP is the list of pixels in the image, and k_{hue}, k_{sat} and k_{value} are the hue, saturation, and value of the pixel k respectively. This gives each HSV channel 16 bins to produce a histogram of 48 features. The histograms are then collated into a matrix representing our HSV feature space x_{hsv}.

$$x_{\text{hsv}} = \begin{bmatrix} v_{\text{hsv}}(\text{image}_1) \\ v_{\text{hsv}}(\text{image}_2) \\ \vdots \\ v_{\text{hsv}}(\text{image}_{\text{nS}}) \end{bmatrix}$$

Global vectors for word representation. While visual features allow us to describe images at low and intermediate levels, higher-level semantic processing requires us to characterise the image in terms of the object it contains. To model object-level information, we included a general set of features describing the semantic differences between concepts. We chose a set derived with the Global Vectors for Word Representation (GloVe) algorithm as it is well established [41–43] and state-of-the-art vector datasets are readily available. The learning objective of GloVe is to generate for each word an M-dimensional vector such that the dot product of two of the vectors equals the logarithm of the probability of the associated words co-occurring in text. As M increases, a larger number of words can be more accurately described; however this increases the computation time both in training GloVe and in any downstream analyses which use GloVe vectors as input. In this project, we make use of a pretrained matrix gMat of 1.9 million words with 300 dimensions indexed by the word [44].

Firstly a name is assigned to each stimulus image to describe the subject of the image. A number of our stimulus images were labeled with a Multi-Word Expression (MWE) which did not have a corresponding feature vector in *gMat*. In these cases we used the mean of its composite words, following [45]. For example, the stimulus "plaster trowel" was set to the mean of the vector for "plaster" and the vector for "trowel".

For each of our images we chose a single word or MWE to represent the content (i.e. the depicted object), and take the row of the GloVe matrix which corresponds to that word as the feature vector for the image in our high-level semantic feature space.

$$x_{\text{sem}} = \begin{bmatrix} gMat_{\text{names}_1} \\ gMat_{\text{names}_2} \\ \vdots \\ gMat_{\text{names}_{nS}} \end{bmatrix}$$

where

$$\text{names} = \{\text{armadillo, axe, badger, beaver}, \ldots, \text{zebra}\}$$

Combining the feature sets. The complete visuo-semantic feature set is then composed by combining x_{gabor}, x_{sift}, x_{hsv} and x_{sem}. Concatenating the raw feature sets together would result in a poor and imbalanced feature space due to the differences in dimensionality and value scaling across the different constituent feature sets. We therefore normalise each feature set to ensure that the values in each row range from zero to one and perform Principal Component Analysis (PCA) to reduce the dimensionality of the concatenated feature space.

With $y \in x_{\text{gabor}}$ and z be the length of y, we normalise the feature vector by its range and stack the results to form x_{gaborR}:

$$v_{\text{gabor}}(\text{image}) = \begin{bmatrix} \dfrac{y_1 - \min(y)}{\max(y) - \min(y)} \\ \dfrac{y_2 - \min(y)}{\max(y) - \min(y)} \\ \vdots \\ \dfrac{y_z - \min(y)}{\max(y) - \min(y)} \end{bmatrix}^T$$

$$x_{\text{gaborR}} = \begin{bmatrix} v_{\text{gaborR}}(x_{\text{gabor}_1}) \\ v_{\text{gaborR}}(x_{\text{gabor}_2}) \\ \vdots \\ v_{\text{gaborR}}(x_{\text{gabor}_{nS}}) \end{bmatrix} \quad (1)$$

The final complete set of features is the concatenation of the features from each of the component visual and semantic models:

$$\text{features}_{vs} = [x_{\text{gab}}, x_{\text{hsv}}, x_{\text{sift}}, x_{\text{sem}}]$$

Finally, before using these features in our classification model, we apply one further feature selection based on a measure of fit from the regression model (as described in Section 'Zero-shot Prediction').

EEG mapping

The mapping of our EEG data to our visuo-semantic feature space is essentially a problem of fitting a regression model f_i for each image feature i such that $f(EEG_y) = [f_1, \ldots, f_i] \approx \mathrm{features}_y$ where EEG_y is the preprocessed nC × nT-dimensional brain activity vector associated with stimulus image y.

Assuming a linear relationship exists between these two components, multiple linear regression can be applied to find some set of weights w1 such that $f_1(\mathrm{EEG}_y) = v_{\mathrm{EEG}_1} * \mathrm{w1}_0 + v_{\mathrm{EEG}_2} * \mathrm{w1}_1 + \ldots$ will produce a value as close as possible to $\mathrm{features}_{y_1}$, some vector of weights w2 such that $f_2(\mathrm{EEG}_y) = v_{\mathrm{EEG}_1} * \mathrm{w2}_0 + v_{\mathrm{EEG}_2} * \mathrm{w2}_1 + \ldots$ will produce a value as close as possible to $\mathrm{features}_{y_2}$, and so on until a vector can be stacked which is as close as possible to $\mathrm{features}_y$.

Prior studies [10, 14] have shown success using a linear regression model with brain data when they are regularised. This coupled with its speed and simplicity made it a natural choice for a baseline approach. L2 regularisation is used to reduce overfitting and improve the generalisation properties of the model. This choice is preferred over L1 regularisation given the expected high collinearity of our samples, i.e. signals recorded from nearby locations in very similar temporal instants should register very similar sources in brain activity. A good model will be able to generalise the relationship rather than being limited to projecting the particular samples and/or classes used in training. If this is achieved, the mapping mechanism and the representative feature spaces could be used within a zero-shot learning architecture.

Zero-shot prediction. Once a mapping between EEG data and the image feature space has been learned from training, a prediction of image features can be made for an EEG epoch withheld from the training set. To ensure a zero-shot framework, we use leave-one-class-out cross-validation to iteratively withhold all epochs associated with a particular stimulus/image for testing in each iteration. Concretely, this means we withhold the data for trials related to Stimulus 1, and train a regression model using the trials for the rest of the stimuli. We then pass the withheld testing trials into our regression model to produce a predicted image feature vector for each trial. We then return the trials for Stimulus 1 to the training set and instead withhold the trials for Stimulus 2. A separate regression model is trained from scratch for this new training set, and then predicted image feature vectors are produced for the Stimulus 2 trials. This pattern is repeated for each stimulus image in the recording.

Following the regression, there is one final step of feature selection over the predicted image features before moving to the feature matching for image retrieval. We do not make use of all image features in the predicted image feature vector, but instead select just those which are best represented in the EEG data. To make the distinction between useful and under-represented features, we approximate each feature's informativeness by calculating the measure of fit of our regression model. When predictions are fed to the classifier, we ignore the columns of the feature space and the predicted feature vectors with the lowest measure of fit.

For each iteration of train/test split, after the regression model has been fit, an R^2 measure of fit is calculated for each image feature column in features. For each epoch in a recording we produce a predicted image feature vector and collate these vectors into the matrix p. Each epoch is associated with a particular stimulus image and each stimulus image is associated with a feature vector in features, so we generate t such that t_i is the feature vector associated with the stimulus image used in epoch i.

These fit values are then averaged across iterations to produce an estimate of which image features are best represented in the EEG data. This estimation is reached entirely without influence from the withheld epochs. The last step of the brain decoding mechanism is implemented using a nearest neighbour classifier between the predicted image feature vector p_j from the

EEG and the target image feature vector t_j. This allows us to order all the images in our database (including the target stimuli) in the image feature space by their distance from the predicted feature vector. We can then convert this ordered list of stimuli into a rank by counting how far down the list the target image is, where a perfect prediction results in rank one and where the expected rank assuming chance performance is nS/2, where nS is the number of stimuli images.

Results

Datasets

Two different datasets are used to evaluate our zero-shot prediction architecture, in order to reduce the risk of overfitting to a particular dataset and limiting the generality of our conclusions. To facilitate comparison with previous approaches, two datasets with similar tasks, "Trento" and "Stanford", are used.

Trento data. The first collection of EEG data analysed in this study is the Trento set [10] which uses 60 grayscale photographs as stimuli. Since this dataset was initially designed for classification, images are grouped in 30 pictures of 30 different land mammals and 30 pictures of 30 different hand tools. However, due to our image retrieval goal this category-level information is discarded and the stimuli are treated as 60 individual images. In an EEG experimental session, these images were each presented six times to each participant, for a total of 360 trials (i.e. 360 epochs). There were seven participants; five males and two females (ages 25-33), all of whom were native speakers of Italian. Each participant completed a single experimental setting.

Participants were instructed to silently name the image with whatever term occurs naturally whilst EEG data was collected with a 64-channel EEG headset sampling at 500Hz. More details of the paradigm and recording of the data can be found in Murphy et al. [10]. The epoched data for each session therefore consists of a matrix of shape $nE \times nC \times nT$, where $nE = 360$, $nC = 64$ and $nT = 500$. Through the preprocessing steps outlined in the 'EEG Data Preprocessing' Section (including removal of noisy epochs), the resulting cleaned set was a matrix of size 340×7680 on average per recording. The number of epochs is approximate as for each experimental session, a different number of low quality epochs for each participant are removed during preprocessing. Similarly, each recording had a different number of components subtracted after ICA (two on average). In the original study, the aim was to train a linear binary classifier to distinguish between epochs associated with *mammal* or *tool* stimuli, which differs from our goal of matching epochs to particular images. As such, the Trento materials use a narrow selection of stimuli from just two semantic categories, and each object image will be visually and semantically similar to many other images in the set. This provides a challenging test of our method's ability to predict the correct image from a set of possible and very close alternatives.

Stanford data. The second EEG dataset we used to test our approach is an openly available dataset compiled at Stanford University [19]. Participants were presented with a series of colour photographs, drawn from the categories *human body*, *human face*, *animal body*, *animal face*, *fruit/vegetable*, or *man-made* (inanimate object). There were 12 images in each category and each image was presented 12 times in random order for a total of 864 trials per recording. Again, we discard category information and the experiment is treated as an image retrieval task with 72 individual images. There were 10 participants (ages 21-57), three of whom were female and one who was left-handed, all reported normal or corrected-to-normal vision. Every participant completed two sessions which each comprised of three separate EEG recordings for a total of 60 recordings. The EEG was recorded using a 128-channel headset sampling at 1kHz. Each recording therefore contained 864 epochs, each with 128,000 features in its raw form. The resulting cleaned set after preprocessing measured approximately 792 epochs × 128 channels × 120 timepoints, giving a EEG feature matrix of average size 792 × 15,360 per recording. Across the recordings in the Stanford dataset, preprocessing resulted in the interpolation of approximately five channels, the subtraction of four independent components, and the removal of 0-2 trials of each stimulus. Following removal of noisy trials during preprocessing, four of these recordings were left with no trials for one of their stimuli; these recordings were excluded from further analysis.

This dataset is the largest of those used in the related image decoding studies above. For comparison, three of the studies [13, 14, 46] involved nine participants who each attended a single recording, though in the case of Palatucci et al. [14] three of these participants were simultaneously recorded with an EEG headset. There is also established work in the field making use of even fewer participants, such as Haxby et al. [9] which had six, or Kay et al. [47] which had two. A slightly larger cohort of 14 was used by Clarke et al. [17], and 20 by Carlson et al. [16], however again each of these participants only attended a single recording task.

Measure of accuracy

Given the difficulty of the zero-shot prediction task, we used an accuracy metric more sensitive to small improvements in prediction power based on the Cumulative Match Curve (CMC). Once a set of predicted visuo-semantic image features is produced for the EEG associated to a particular image presentation, all the images were ranked by their Euclidean distance from the predicted feature vector. A CMC was then generated by counting how often the true target appears in the top X ranked images as shown in Fig 1. For example, the first value on the x-axis represents the percent of cases where the target image is the nearest to the predicted features in the feature space, the second value on the x-axis represents the percent of cases in which the target image was one of the two closest images to the predicted features, and so on. The Area Under Curve (AUC) is calculated as the normalised volume below the curve for use as the final metric.

Parameter optimisation

A short gridsearch was performed to empirically optimise the parameters. A random recording from each dataset was chosen and used to perform this gridsearch for each experiment below. We then used the highest performing parameter set to perform the decoding for the rest of the recordings with the same dataset and image feature set. We do expect that different recordings will perform best under different parameter settings, and as such accuracy could be maximised with a more rigorous approach to gridsearching. That said we have chosen to determine parameters from a single recording in order to better reflect training in a real-world BCI system. Gridsearching for the Trento dataset was performed with the recording from Subject 6, and for the Stanford dataset with the first recording of Subject 10. In Tables 1 and 2 the recording used for each dataset has been marked (GS) and excluded from the mean column.

Table 1. Trento decoding accuracies (CMC AUC % (Cumulative Match Curve Area Under Curve)).

Feature Set	S1	S2	S3	S4	S5	S7	GS	Mean
Visual	57.2	65.12	58.72	55.75	64.58	67.18	62.03	62.56
Semantic	57.74	64.74	56.26	61.66	65.98	63.9	64.28	61.71
Visuo-semantic	57.63	63.32	58.46	62.8	62	63.72	66.06	61.32

https://doi.org/10.1371/journal.pone.0214342.t001

Table 2. Stanford decoding accuracies (CMC AUC % (Cumulative Match Curve Area Under Curve)).

Feature Set	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	GS	Mean
Visual	59.27	56.91	57.28	57.51	62.57	60.96	56.59	54.5	54.11	61.08	61.97	58.08
Semantic	60.12	56.89	61.38	60.35	65.49	63.93	62.36	55.93	58.07	64.28	64.82	60.88
Visuo-semantic	61.86	58.45	60.8	61.85	67.9	66.08	63.07	56.59	57.84	65.77	66.01	62.02

https://doi.org/10.1371/journal.pone.0214342.t002

Alpha values {1e-2, 1e-1, 1e0, 1e1, 5e1, 1e2} were tested for the ridge regressor. The number of EEG features retained during feature selection ('EEG Feature Selection' Section) was tested over the values {25, 50, 75, 100, 125, 150, 175, 200, 250, 500, 750, 1000, 1500, 2000, 2500, 3000}.

Decoding accuracy

In order to compare the effectiveness of our chosen image feature models and confirm our expectation that combining the models would provide more predictive power than using them in isolation, the CMC AUC for both datasets was calculated when using all visuo-semantic features (features$_{vs}$) and compared against using only visual feature set (features$_{v}$) or the semantic feature set (features$_{s}$) individually.

$$\text{features}_{vs} = [x_{\text{gab}}, x_{\text{hsv}}, x_{\text{sift}}, x_{\text{sem}}]$$

$$\text{features}_{v} = [x_{\text{gab}}, x_{\text{hsv}}, x_{\text{sift}}]$$

$$\text{features}_{s} = x_{\text{sem}}$$

Results are shown in Table 1 for the Trento dataset and Table 2 for the Stanford dataset.

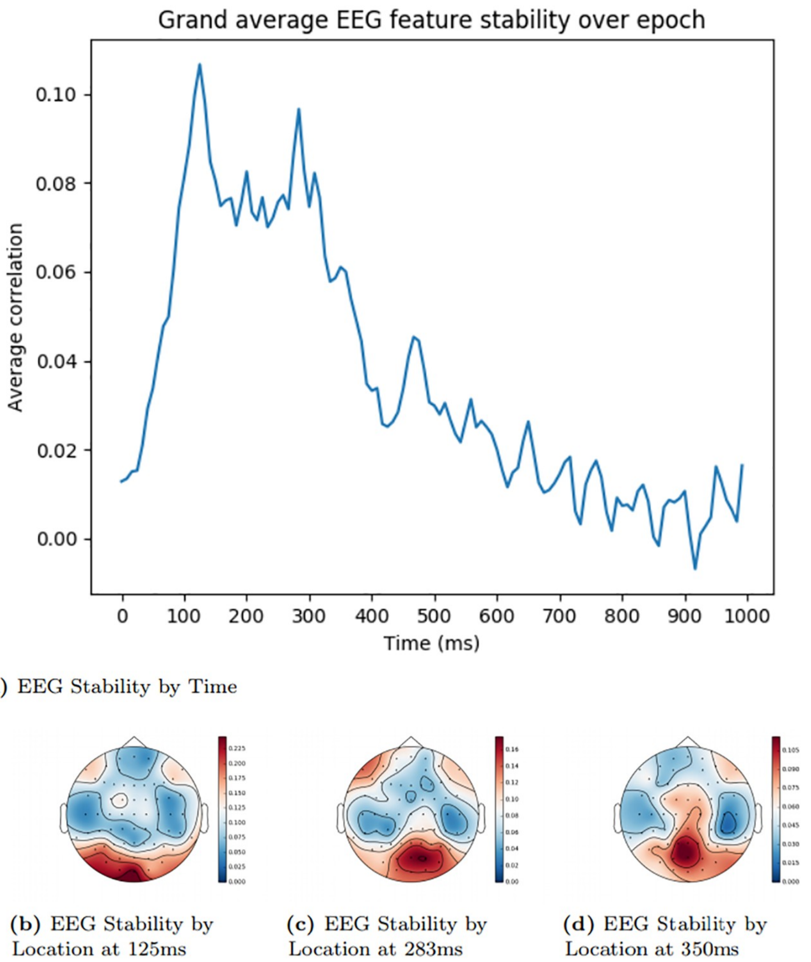
All the exemplar decoding results we present are significantly above chance (50%), indicating a mapping between EEG activity and the image feature sets we have chosen that can be used for zero-shot brain decoding. For the Stanford dataset, which uses a larger and more diverse set of object images, the semantic feature set gives better accuracy than the visual feature set, and best performance is obtained with the combined visuo-semantic feature set.

In the results for the Trento dataset, these trends are less clear, but it can be seen that the combination of all features is a robust approach overall.

EEG feature selection visualisation

In order to demonstrate that our EEG feature selection was performing as expected and was properly selecting activity from channels and timepoints known to relate to meaningful visual processing, we analysed which EEG features were assigned the highest score by the stability measure outlined in the 'EEG Feature Selection' Section. Figs 2a and 3a show a grand average of the stability scores at different times during an epoch. These values were generated by taking the mean of the scores across each channel for the time offset in question. Figs 2b–2d and 3b–3d show snapshots of the stability values at particular timepoints from the temporal plots, distributed over the EEG sensor locations.

Stability peaks within the expected time window [16], indicating that the feature selection method is properly determining the most informative features within the EEG activity. Based on previous research [16, 17, 30], we would expect the posterior sensors to be more useful early in an epoch when there is more visual processing, and that informative areas later on would be more spatially diffuse, when activity reflects more widely-distributed semantic



(a) EEG Stability by Time

(b) EEG Stability by Location at 125ms

(c) EEG Stability by Location at 283ms

(d) EEG Stability by Location at 350ms

Fig 2. Trento EEG feature stability by time and location. The areas shaded in red signify the locations with highest electroencephalography (EEG) stability, while areas shaded in blue signify the lowest.

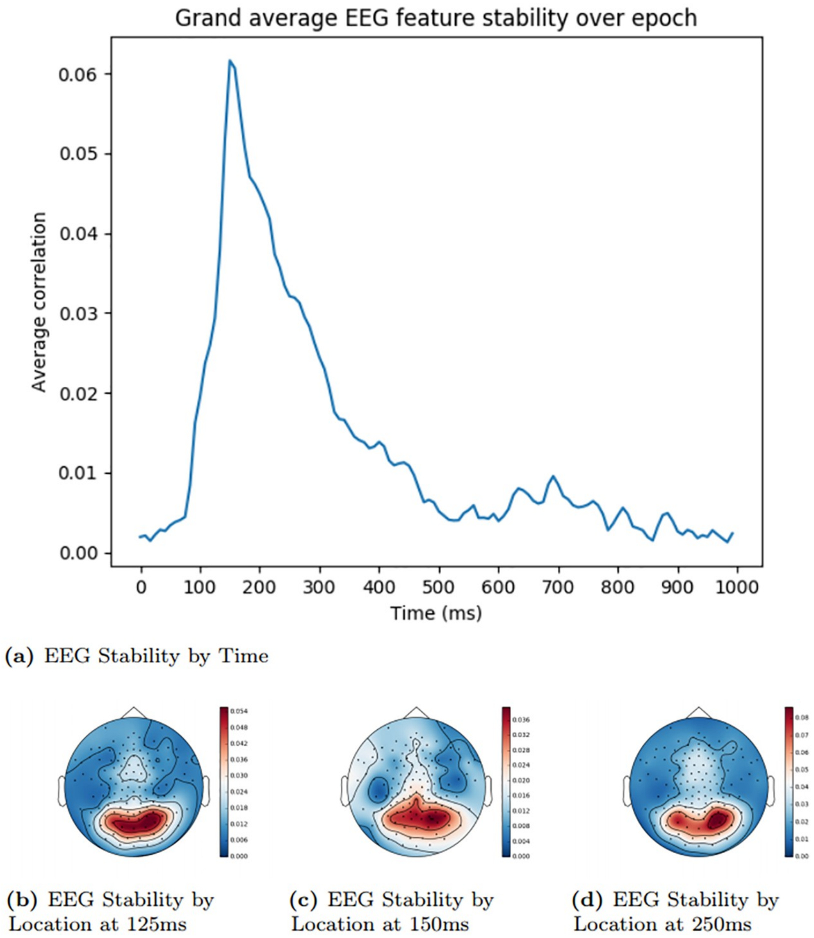
https://doi.org/10.1371/journal.pone.0214342.g002

processing of the stimulus. The stability analysis reflects this pattern, more clearly in the Trento dataset than the Stanford dataset.

Comparison with state-of-the-art

Because of our zero-shot analysis framework, a study with directly comparable results could not be identified in a review of relevant EEG literature. However, the studies mentioned in the background section can provide a frame of reference. While Palatucci et al. [14] used image stimuli and decoded the image from brain activity, the focus was on decoding semantic information about the object in the image rather than retrieving the stimulus image based on the brain data. The datasets we have access to in this study involve much more visually complex image stimuli. Moreover, where Palatucci et al. [14] made use of minimilistic line drawings, the photographs used in both datasets analysed in this study are much more visually complex. In order to best leverage this extra visual information, we added several visual feature sets to our analysis.

The leave-one-class-out task performed by Palatucci et al. [14] is similar enough to the task in this study to give context to our results, though given that the two studies use different datasets a direct comparison with our approach is not possible. The paradigm used in the Palatucci



(a) EEG Stability by Time

(b) EEG Stability by Location at 125ms

(c) EEG Stability by Location at 150ms

(d) EEG Stability by Location at 250ms

Fig 3. Stanford EEG feature stability by time and location. The areas shaded in red signify the locations with highest electroencephalography (EEG) stability, while areas shaded in blue signify the lowest.

https://doi.org/10.1371/journal.pone.0214342.g003

et al. study was very similar to those used in the Trento and Stanford experiments, with participants being presented with a series of images and asked to silently name them. Compared with the Palatucci et al. [14] study, we obtain slightly better results (Table 3).

Conclusion

In this paper we proposed an approach to zero-shot image retrieval in EEG data using a novel combination of feature sets, feature selection, and regression modeling. We have shown that a combination of visual and semantic feature sets performs better than using either of those feature sets in isolation. We also analysed the performance of each image feature model used in our approach individually to help identify where future improvements could be made.

Table 3. Leave-one-class-out task percent rank accuracy.

Dataset Name	Rank Accuracy
Palatucci et al. [14] (their approach)	56%
Trento dataset (our approach)	61.1%
Stanford dataset (our approach)	61.65%

https://doi.org/10.1371/journal.pone.0214342.t003

We hope that future work can improve upon this approach using the same open dataset for comparison as it is difficult to accurately predict how well our approach would perform on the other datasets mentioned in Sections 'Introduction' and 'Comparison with State-of-the-art'. While this study considers a relatively large amount of EEG data in comparison to related studies (as shown in the 'Stanford Data' Section), a larger and more diverse dataset may be beneficial in evaluating the performance of our approach.

We have demonstrated that the features we extracted for the EEG data and images are justified and perform significantly above chance. However it is possible that our image features do not accurately reflect all stages of human visual processing, and that a different set of features would better facilitate a regression model. For example, large neural networks that recognise images have a hierarchical architecture which reflects some aspects of human visual processing [48, 49] and which could provide an effective model for our feature space. Alternatively we could replace our semantic features with a set derived from a distributional word embedding model such as word2vec [50] or fastText [51].

Moreover, our EEG feature selection may correctly quantify the usefulness of each particular timepoint in each channel, but it is likely that features which are close in time and location will have very similar information and thus similar scores, and so a feature selection method may select a set of good quality but redundant features. In future work, we will explore feature selection methods that produce a small set of maximally informative EEG features. Nevertheless, our approach has demonstrated a marked improvement over current state-of-the-art for EEG zero-shot image decoding and is a step towards the application of EEG to real-world BCI technologies.

Author Contributions

Conceptualization: Ben McCartney, Brian Murphy.

Data curation: Ben McCartney.

Formal analysis: Ben McCartney.

Investigation: Ben McCartney.

Methodology: Ben McCartney, Jesus Martinez-del-Rincon, Barry Devereux, Brian Murphy.

Project administration: Ben McCartney, Jesus Martinez-del-Rincon, Barry Devereux, Brian Murphy.

Resources: Jesus Martinez-del-Rincon, Barry Devereux, Brian Murphy.

Software: Ben McCartney.

Supervision: Jesus Martinez-del-Rincon, Barry Devereux, Brian Murphy.

Validation: Ben McCartney.

Visualization: Ben McCartney.

Writing – original draft: Ben McCartney.

Writing – review & editing: Ben McCartney, Jesus Martinez-del-Rincon, Barry Devereux.

References

1. Vidal J. Toward Direct Brain-computer Communication. Annual Review of Biophysics and Bioengineering. 1973.

2. Rajendra Acharya U, Fujita U. Application of entropies for automated diagnosis of epilepsy using EEG signals: A review. Knowledge-Based Systems. 2015; 88:85–96. https://doi.org/10.1016/j.knosys.2015.08.004

3. Bhat S, Rajendra Acharya U, Dadmehr N, Adeli H. Clinical neurophysiological and automated EEG-based diagnosis of the Alzheimer's disease. European Neurology. 2015; 74:202–210. https://doi.org/10.1159/000441447 PMID: 26588015

4. Schwartz AB, Cui XT, Weber DJ, Moran DW. Brain-Controlled Interfaces: Movement Restoration with Neural Prosthetics. Neuron. 2006; 52(1):205–220. https://doi.org/10.1016/j.neuron.2006.09.019 PMID: 17015237

5. Li Y, Guan C, Li H, Chin Z. A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system. Pattern Recognition Letters. 2008; 29(9):1285–1294. https://doi.org/10.1016/j.patrec.2008.01.030

6. Pedroso RV, Fraga FJ, Corazza DI, Andreatto CAA, Coelho FGdM, Costa JLR, et al. P300 latency and amplitude in Alzheimer's disease: A systematic review. Brazilian Journal of Otorhinolaryngology. 2012; 78(4):126–132. https://doi.org/10.1590/S1808-86942012000400023 PMID: 22936149

7. Sarnthein J, Andersson M, Zimmermann MB, Zumsteg D. High test-retest reliability of checkerboard reversal visual evoked potentials (VEP) over 8 months. Clinical Neurophysiology. 2009; 120(10):1835–1840. https://doi.org/10.1016/j.clinph.2009.08.014 PMID: 19762276

8. Larson MJ, Clayson PE, Clawson A. Making sense of all the conflict: A theoretical review and critique of conflict-related ERPs. International Journal of Psychophysiology. 2014; 93(3):283–297. https://doi.org/10.1016/j.ijpsycho.2014.06.007 PMID: 24950132

9. Haxby JV. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science. 2001; 293(5539):2425–2430. https://doi.org/10.1126/science.1063736 PMID: 11577229

10. Murphy B. EEG responds to conceptual stimuli and corpus semantics. In: Conference on Empirical Methods in Natural Language Processing; 2009. p. 619–627.

11. Matran-Fernandez A, Poli R. Collaborative brain-computer interfaces for target localisation in rapid serial visual presentation. In: 2014 6th Computer Science and Electronic Engineering Conference (CEEC). IEEE; 2014. p. 127–132.

12. Sajda P, Pohlmeyer E, Wang J, Hanna B, Parra LC, Chang Sf. Cortically-Coupled Computer Vision. Brain-Computer Interfaces. 2010; p. 133–148.

13. Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason Ra, et al. Predicting human brain activity associated with the meanings of nouns. Science (New York, NY). 2008; 320(5880):1191–1195. https://doi.org/10.1126/science.1152876

14. Palatucci MM. Thought recognition: predicting and decoding brain activity using the zero-shot learning model. Citeseer; 2011.

15. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452(7185):352–5. https://doi.org/10.1038/nature06713 PMID: 18322462

16. Carlson T, Tovar DA, Kriegeskorte N. Representational dynamics of object vision: The first 1000 ms. Journal of Vision. 2013; 13:1–19. https://doi.org/10.1167/13.10.1 PMID: 23908380

17. Clarke A, Devereux BJ, Randall B, Tyler LK. Predicting the time course of individual objects with MEG. Cerebral Cortex. 2015; 25(10):3602–3612. https://doi.org/10.1093/cercor/bhu203 PMID: 25209607

18. Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, et al. Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage. 2012; 62(1):451–463. https://doi.org/10.1016/j.neuroimage.2012.04.048 PMID: 22565201

19. Kaneshiro B, Perreau Guimaraes M, Kim HS, Norcia AM, Suppes P. A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. Plos One. 2015; 10(8):e0135697. https://doi.org/10.1371/journal.pone.0135697 PMID: 26295970

20. Nolan H, Whelan R, Reilly RB. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. Journal of neuroscience methods. 2010; p. 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015 PMID: 20654646

21. Cecchi M, Moore DK, Sadowsky CH, Solomon PR, Doraiswamy PM, Smith CD, et al. A clinical trial to validate event-related potential markers of Alzheimer's disease in outpatient settings. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring. 2015; 1(4):387–394. https://doi.org/10.1016/j.dadm.2015.08.004 PMID: 27239520

22. Ramos-Murguialday A, Birbaumer N. Brain oscillatory signatures of motor tasks. Journal of neurophysiology. 2015; 7:jn.00467.2013.

23. HURST HE. Long-Term Storage Capacity of Reservoirs. Trans Amer Soc Civil Eng. 1951; 116:770–799.

24. Perrin F, Pernier J, Bertrand O, Echallier J. Spherical splines for scalp potential and current density mapping. Electroencephalography and clinical neurophysiology. 1989; 72(2):184–187. https://doi.org/10.1016/0013-4694(89)90180-6 PMID: 2464490

25. Jas M, Engemann D, Raimondo F, Bekhti Y, Gramfort A. Automated rejection and repair of bad trials in MEG/EEG. In: 2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE; 2016. p. 1–4.

26. Makeig S, Ca SD, Bell AJ, Sejnowski TJ. Independent component analysis of electroencephalographic data. Advances in neural information processing systems. 1996; p. 145–151.

27. Vorobyov S, Cichocki A. Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis. Biological Cybernetics. 2002; 86(4):293–303. https://doi.org/10.1007/s00422-001-0298-6 PMID: 11956810

28. Caceres CA, Roos MJ, Rupp KM, Milsap G, Crone NE, Wolmetz ME, et al. Feature Selection Methods for Zero-Shot Learning of Neural Activity. Frontiers in Neuroinformatics. 2017; 11(June):1–12.

29. Carlson TA, Ritchie JB, Kriegeskorte N. RT for Object Categorisation Is Predicted by Representational Distance. Journel of Cognitive Neuroscience. 2007; p. 1–11.

30. Clarke A, Taylor KI, Tyler LK. The evolution of meaning: spatio-temporal dynamics of visual object recognition. Journal of cognitive neuroscience. 2011; 23(8):1887–1899. https://doi.org/10.1162/jocn.2010.21544 PMID: 20617883

31. Clarke A, Taylor KI, Devereux B, Randall B, Tyler LK. From perception to conception: how meaningful objects are processed over time. Cerebral Cortex. 2012; 23(1):187–197. https://doi.org/10.1093/cercor/bhs002 PMID: 22275484

32. Clarke A, Tyler LK. Understanding what we see: how we derive meaning from vision. Trends in cognitive sciences. 2015; 19(11):677–687. https://doi.org/10.1016/j.tics.2015.08.008 PMID: 26440124

33. Hamilton W. Biologically Inspired Object Recognition using Gabor Filters; 2013.

34. Leeds DD, Seibert DA, Pyles JA, Tarr MJ. Comparing visual representations across human fMRI and computational vision. Journal of Vision. 2013; 13(13):25–25. https://doi.org/10.1167/13.13.25 PMID: 24273227

35. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology. 2011; 21(19):1641–1646. https://doi.org/10.1016/j.cub.2011.08.031 PMID: 21945275

36. Jones JP, Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of neurophysiology. 1987; 58(6):1233–1258. https://doi.org/10.1152/jn.1987.58.6.1233 PMID: 3437332

37. Lindeberg T. Scale Invariant Feature Transform. Scholarpedia. 2012; 7(5):10491. https://doi.org/10.4249/scholarpedia.10491

38. Yang J, Jiang YG, Hauptmann AG, Ngo CW. Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM; 2007. p. 197–206.

39. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. 2009.

40. Sural S, Gang Qian, Pramanik S. Segmentation and histogram generation using the HSV color space for image retrieval. Proceedings International Conference on Image Processing. 2002;2:II–589–II–592.

41. Güçlü U, van Gerven MAJ. Semantic vector space models predict neural responses to complex visual stimuli. arXiv preprint. 2015.

42. Trask A, Gilmore D, Russell M. Modeling order in neural word embeddings at scale. arXiv preprint arXiv:150602338. 2015.

43. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.

44. GloVe: Global Vectors for Word Representation;. https://nlp.stanford.edu/projects/glove/.

45. Mitchell J, Lapata M. Composition in distributional models of semantics. Cognitive science. 2010; 34 (8):1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x PMID: 21564253

46. Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, et al. Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage. 2012; 62(1):451–463. https://doi.org/10.1016/j.neuroimage.2012.04.048 PMID: 22565201

47. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452(7185):352. https://doi.org/10.1038/nature06713 PMID: 18322462

48. Devereux BJ, Clarke A, Tyler LK. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Scientific reports. 2018; 8(1):10636. https://doi.org/10.1038/s41598-018-28865-1 PMID: 30006530

49. Kriegeskorte N. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual review of vision science. 2015; 1:417–446. https://doi.org/10.1146/annurev-vision-082114-035447 PMID: 28532370

50. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. arXiv preprint arXiv:160704606. 2016.

51. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T, et al. Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems; 2013. p. 2121–2129.