Visual and Semantic Feature Spaces for Zero-shot Image Decoding


Published in:

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2018 The Authors and The Irish Pattern Recognition & Classification Society.
This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Abstract

Large image galleries are notoriously difficult to categorise and navigate without a great deal of manual organisation. We present a new methodology that could enable a Brain-Computer Interface (BCI) system for image retrieval, as an alternative to the automated tagging approaches recently emerging. We propose a regression framework that allows us to effectively map EEG signals into a visuo-semantic feature space, where the stimulus image can be matched. We use a set of intermediate features ranging from low-level edge detection, to the semantic properties of an image’s depicted content, reflecting theoretical and empirical knowledge of how the brain processes visual input and extracts meaning from it. A real-world system would need to generalise to images not seen during training, so we use zero-shot learning to decode image features from brain activity. Using our approach under this challenging setting, we were able to decode left-out individual stimuli at a rate significantly above chance.

1 Introduction

This project is based in the field of BCI, a system which directly measures brain activity associated with the user’s intent and translates the recorded brain activity into corresponding control signals. There are numerous and varied BCI applications such as in research [Rajendra Acharya and Fujita, 2015, Bhat et al., 2015] and assistive technology [Schwartz et al., 2006, Yin et al., 2015]. We intend to produce a system enabling us to decode features about the stimulus image a user is looking at from their brain activity for the purpose of image retrieval.

The approach outlined in this paper is built upon research showing that it is possible to decode information about current semantic [Mitchell et al., 2008, Murphy et al., 2011, Murphy et al., 2012] and visual [Sajda et al., 2010, Leeds et al., 2013] processing from neural activation. We will employ data-driven techniques such as signal processing and machine learning to find a relationship between a user’s neural response and the semantic/visual features we can compute about the stimulus image. We use an intermediary feature space [Mitchell et al., 2008, Palatucci and Pomerleau, 2011] rather than classifying directly over a set of images; to train a system to directly categorise an image never seen before, we would need a huge representative image set. This would be impractical for us to collate and train over, so finding a relationship to a latent space defined by a set of feature models which could be used to describe any given image provides us an alternative.

We expect that the different levels of information in our feature set will reflect different stages of neural processing. Meaningful object vision has been shown to be a hierarchical, incremental process [Carlson et al., 2007, Carlson et al., 2013], so we expect that the most useful time for detecting low level vision be sooner after stimulus onset than high-level semantic processing. We will run this analysis across our feature sets to assess how informative each set is and how much redundancy lies between them. Given the former assumption, we would expect that the lower level visual features (e.g. Gabor texture features) would be more informative early in each trial and higher-level features related to the image content (e.g. GloVe semantic vectors) would be more informative later (see Section 2.1 for more detail on image feature models used). We will run a sliding window analysis to test this theory.

The two main approaches for recording brain activity involve either measuring changes in blood oxygenation (such as in functional Magnetic Resonance Imaging (fMRI)), or measuring the electrical activity (such as in
Electroencephalography (EEG)). Typically fMRI will have better spatial resolution and will be less susceptible
to noise, and EEG will have better temporal resolution and will be more logistically accessible. The recorded
brain data we will be using to test our approach was initially gathered by the authors of [Murphy, 2009] using
EEG, as the the paper targeted a similar problem and behavioural paradigm is suitable. We require a paradigm
which encourages the participant to consider a series of single object images with minimal distractions, at a
rate no faster than 2Hz [Carlson et al., 2013]. More detail on the brain data can be found in Section 2.1.3.

The inspiration for our project is the plan to create a system allowing a user to open their large, unorganised
photo gallery and think about a particular image there in order to retrieve and display it. In order to investigate
the feasibility of such a system, the work detailed here attempts to solve a more realistic version of the problem,
where images were explicitly viewed by participants, rather than imagined.

1.1 State of the Art

A common approach in EEG analysis involves several relatively clearly defined patterns known as Event Re-
lated Potentials (ERPs). One such pattern is the Steady State Visually Evoked Potential (SSVEP) in which
brain signals will fluctuate at the same frequency as a visual stimulus. Spelling systems based on SSVEPs
[Yin et al., 2015] present several options on a screen which each flicker at a different rate, and whichever op-
tion is fixated by the participant will be reflected in their brain activity.

An alternative approach in a similar vein is using Rapid Serial Visual Presentation (RSVP)
[Sajda et al., 2010]. This typically leverages a different ERP in the P300 defined as the third positive peak
following the onset of some stimulus. Stimuli are displayed one after another at a rate of around 10Hz. Partic-
ipants attend these images searching for a particular target image or characteristic. When this target is shown,
changes in the timing and strength of the P300 can be detected.

The most simple form of image decoding from brain data involves determining which of a finite set of
categories the stimulus image falls within rather than determining the particular image exemplar displayed.
Studies such as [Haxby, 2001] and [Murphy, 2009] have achieved strong results by applying a linear classifier
to brain data.

The two ERP-based approaches above could be modified to fit our problem of image exemplar decoding,
however it would involve presenting the entire gallery of images until the target is shown. While this may be
feasible in a research paradigm, it would be impractical for a real-world application.

There are relatively few studies which attempt zero-shot decoding of image stimuli from brain data. The
most appropriate of these are the aforementioned [Mitchell et al., 2008, Palatucci and Pomerleau, 2011]. The
main difference in our approach from these studies is that we include visual features in our feature space
rather than just semantic features in an effort to reflect a more complete model of neural activation following
presentation of a stimulus image. It is also worth noting that direct accuracy comparison would be difficult
given our study makes use of a different dataset of brain recordings with a slightly different recording paradigm
and equipment.

2 Methodology

In this paper, we propose to model the relationship between a EEG feature space and a visuo-semantic feature
space. Once those feature spaces are created by automatically deriving discriminative features from either the
EEG sensors or the image file, a linear regression model is trained to reflect the relationship from the former to
the latter space. Finally, EEG features from a test trial withheld from the set used to train the regression model
is projected into the visual-semantic space. A nearest neighbour search is performed using the projected image
features along with the features computed for all images in the gallery to predict which image is most likely the
test stimulus.
2.1 Image Feature Sets

The feature sets used in this study were chosen to best reflect both brain activity and state of the art in their field in order to give our algorithm the best chance of discovering a stable mapping and to generalise to images not seen during training. The components of human vision we chose to represent cover low-level edge/texture detection, colour, and keypoint description. The higher level processing of visual input such as categorising between mammal or tool is less completely understood, so we model the rest with semantic features. It is likely that each time an image is presented to the participant approximately the same neural processes will occur, however there will be small differences in timing and order that will change which brain features are informative. With that in mind, in the implementation of all these image feature sets we attempted to generalise as much as possible away from embedding spatial details into our model by using histogram representations. In order to minimise the impact of overfitting, we used default or general values for the parameterisation of our feature extraction models (details in following subsections). Future work will aim to optimise these settings.

Once we generate a feature vector from each feature set, we must normalise them to ensure sets typically of difference scale are not ignored when calculating similarity. We chose to perform row-wise normalisation so the values in each feature vector would range from zero to one, as we felt the scale of features within the feature set would be informative and should be maintained. We repeat this process of generating feature vectors for each image in our gallery, these vectors are then collated to form a matrix which serves as our image feature space. In order to control the large resulting dimensionality and its imbalance between feature sets, we applied PCA to reduce each feature set to 32 features. This PCA transformation was fit using the training data each iteration of cross-validation and also applied to the test data to avoid influence of the withheld epochs. More detail on the cross-validation used can be found in Section 2.2.

2.1.1 Visual Feature Sets

We chose a number of feature sets that are calculated from the low-level perceptual properties of images. The Gabor Filter Bank is a well established computer vision technique for modelling edge and texture detection, and functions in a very similar way to the lowest level of human visual processing [Hamilton, 2013]. Each of these filters represents an edge with a particular orientation and spatial frequency, and applying these to an image highlights the areas where there is a match. We generate a bank with 8 evenly spaced orientations and 4 spatial frequencies ranging from 2 to 5. The rest of the parameters were fixed at ksize=(31, 31), lambda=6.0, gamma=0.5 and psi=0 to result in a bank of 32 filters. Each image in our feature set was convolved with each filter, and the result summed to generate a histogram of 32 dimensions for each image.

We next create a feature set modelling colour information in images. We used a global HSV histogram, as there is some evidence that a HSV colour space comes closer to reflecting human vision than an RGB colour space [Hutchison and Mitchell, 1973]. Each HSV channel is coded using 16 bins to result in a histogram of 48 features.

The highest level visual processing was modelled using Scale-Invariant Feature Transform (SIFT) [Lindeberg, 2012]. SIFT is a well established and commonly used general computer vision model for describing keypoints in an image. It locates these keypoints using difference of Gaussian over different scales and describes the local pixels with general orientation. In order to again abstract spatial data away from our feature set, a Visual Bag Of Words (VBOW) implementation [Yang et al., 2007] is used to generate a histogram over a codebook of precomputed SIFT descriptors. This approach involves generating SIFT keypoint descriptors for a large corpus of images then selecting the top X most informative descriptors to compile a codebook. In order to accurately generalise to any given image this codebook must be trained across a large representative set of images, so we made use of a codebook generated by the well-known image database Imagenet [Deng et al., 2009]. We fed our stimulus set into their SIFT feature kit and produced a histogram indicating how often each ‘visual word’ appeared in the image. This feature kit made use of Dense SIFT meaning that instead of searching for natural key points in an image, it uses a regularly sampled in a grid instead.
2.1.2 Semantic Feature Set

We next generated a feature set that described semantic information about the objects in the images, but which did not make use of low-level visual properties. We chose a set derived with the Global Vectors for Word Representation (GloVe) algorithm as it is well established [Güçlü and van Gerven, 2015, Trask et al., 2015] and large well trained vector sets are easily available. In the cases where our stimulus was a Multi-Word Expression (MWE) and did not appear in the set, we took the mean of its composite words in a similar way to the approaches in [Mitchell and Lapata, 2010]. For example the stimulus ‘plaster trowel’ was set to the mean of the vector for ‘plaster’ and the vector for ‘trowel’. This will become more of an issue when we move on to more naturalistic, unlabelled images in contrast to our highly controlled, single object images set on a gray background. In an image of a busy street, we may need to use some form of automatic annotation such as [Vinyals and Toshev, 2015] in tandem with a semantic feature space. In addition while GloVe is sufficient for describing single concepts, it is unlikely that applying the same feature-mean approach would be effective in describing a scene involving a great many and varied subjects. It may be that a BOW histogram of GloVe descriptors would be informative, however use in a zero-shot system may require it to be impractically large.

2.1.3 EEG Feature Set

The EEG data analysed here was recorded for the study [Murphy, 2009] which used an EEG headset to record the brain activity of participants as they were presented with a series of stimulus images. The images were comprised of 60 grayscale photographs, 30 of land mammals and 30 of hand tools, on a neutral gray background. These images were presented to the participant six times each for a total of 360 epochs per recording. The participants were instructed to silently name the image with whatever term occurs naturally while their brainwaves were recorded with a 64-channel EEG headset at 500Hz. The recordings were band pass filtered at 1-50Hz in order to remove noise unrelated to brain activity and downsampled to 120Hz before applying Independent Component Analysis (ICA) to attempt to identify and remove noise caused by eye-movement. More details of the paradigm and preprocessing can be found in [Murphy, 2009].

Following preprocessing, each EEG recording is a 3D-matrix measuring $n_E \times n_S \times s_R$ where $n_E$ is the number of epochs in the recording, $n_S$ is the number of sensors in the recording headset, and $s_R$ is the sampling rate. In our case this results in a matrix of $360 \times 64 \times 120$. EEG data can be categorised as very noisy given the fact that there will always be brain activity present unrelated to the paradigm and imperfections introduced by the recording equipment. It is also highly collinear as features recorded at spatially close sensors and similar
times will reflect very similar electrical sources. Finally, it is difficult for a participant to maintain concentration for long paradigms meaning there are often few epochs, complicating training of machine learning techniques.

### 2.2 Regression Mapping

The continuous EEG data recordings were cut into 360 one-second epochs corresponding to the presentation of each of the images in the EEG recording session. Each one-second epoch begins at the presentation onset of the image on the screen. We then trained an L2 regularised linear regressor from scikit-learn [Pedregosa et al., 2011] to predict the image features outlined above from the epochs. L2 regularisation is suited to EEG data with its high channel collinearity and high feature count versus trial count. We chose a linear regression model because it is relatively fast and there is precedent for its use in this kind of intermediary feature-space learning [Mitchell et al., 2008]. In this first analysis with this method we used a simple exhaustive search of regularisation parameter settings.

We performed leave-one-class-out cross-validation to avoid accuracy inflation from overfitting and to create a zero-shot learning scenario. In each iteration of the train/test split we leave epochs associated with one stimulus out of the training set, and use this training set to train a separate linear regression model to predict a value for each image feature. The withheld epochs are then used for testing. We make a prediction into the visuo-semantic space with each of these test epochs using the linear regressor and compare the projected feature vectors with the vector we have derived in advance for each image in our gallery and rank them by Manhattan distance to generate a sorted list of candidates.

![Figure 2: Mapping EEG Epochs to Image Features.](image)

### 3 Results

We performed two main experiments on our data, exemplar decoding with a sliding 50ms window over the neural activations of an epoch and exemplar decoding using the whole epoch. Experiments were performed for each feature sets in isolation as well as their combination to evaluate the influence of each feature set. As evaluation metric, we use a Cumulative Match Characteristic curve [Moon and Phillips, 2001] over the resulting sorted list of candidates and calculate the Area Under the Curve (CMC AUC). We then normalise this AUC value by converting it to a percentage of the best possible result with chance accuracy at 50%. Given the difficulty of this zero-shot scenario and the low likelihood of getting the stimulus image as first candidate, the CMC AUC should be more sensitive to small improvements than an accuracy measure.
3.1 Experiment 1: Sliding Window Analysis

In this first experiment, the zero-shot exemplar decoding is performed using a short segment of the total epoch and using a sliding window strategy. We used 50ms windows with 25ms overlap between 0ms and 500ms. In each window we used raw EEG data from all channels to make predictions and calculate an accuracy. This setting aims to explore the temporal evolution of our chosen feature sets where low level features should be more informative earlier, while high level feature will have a stronger relationship with later brain activity. It also aims to validate the feature sets by comparing their behaviour with our current knowledge of the human visual and semantic processing.

Results for this experiment are depicted in Figure 3. Looking a temporal development across the epoch, we can see a number of expected patterns according to neuroscience. The initial large peak present in all feature sets occurs around 100ms which is in line with the findings in [Carlson et al., 2013]. We can also see that while the semantic feature set does also peak before 100ms, it has a second later peak around 250ms as the visual sets tend toward chance. These findings are in accordance to our understanding of brain image processing, which indicate our proposed image features are representative of the brain activity and therefore useful for BCI.

Finally, the total visuo-semantic combination, while apparently dominated by the semantic feature, slightly improves the decoding in comparison to the semantic set even after the visual sets begin to fall off.

![Figure 3: AUC Results for Window Decoding.](image)

3.2 Experiment 2: Full Epoch Analysis

In this first experiment, the zero-shot exemplar decoding is performed using the full 1000ms epoch. The goal is to evaluate the maximum performance achievable for our approach for brain image decoding.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Total</th>
<th>Semantic</th>
<th>Visual</th>
<th>SIFT</th>
<th>HSV</th>
<th>Gabor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average CMC AUC(%)</td>
<td>59.79</td>
<td>59.25</td>
<td>59.54</td>
<td>56.75</td>
<td>59.03</td>
<td>59.5</td>
</tr>
</tbody>
</table>

Table 1: Decoding Results Across Different Feature Combinations.

Decoding accuracies presented in Table 1 indicates how, by using the full epoch, the final performance improves regarding experiment 1, as it could have been expected. In addition, results confirm most findings in the previous experiment. The combination of visual and semantic feature sets outperform any set in isolation, although both semantic and Gabor features in isolation are representative of the brain decoding mechanism.
4 Discussion

The main motivation behind our choice of a variety of feature sets was to make use of the different levels of information processing involved in human vision. From our results it can be observed that low level visual features exhibit very similar peak decoding times, suggesting that there is shared information between the feature sets, but also that they correlate with human brain activity as they peak at the expected 100ms timestamp. The timing of the second later peak of the semantic curve in Figure 3 would suggest that it does reflect processing conceptually more abstract than visual characteristics. Semantic features also exhibit a strong early peak, and there are a number of possible explanations. It may be that this set contains some of the same information as the visual sets, perhaps derived from the fact that semantically similar objects also tend to share some visual characteristics (e.g. 'car' and 'truck' versus 'apple' and 'pear'). Finally, we might expect earlier parts of the EEG timecourse to be more informative as processing is more automatic, less subject to top-down control, and so more consistent in its timing.

The spread of results in Table 1 is exactly as expected with the exception of the generally poor performance of SIFT. However, the decoding deltas are so small that it would be difficult to draw any concrete conclusions from this. We can see at least in Figure 2 that the differing curves for each feature set suggest that there is differing useful information in these different sources.

5 Conclusion

We have shown with this initial approach that we can decode exemplars significantly above chance using an intermediate feature space. Our results demonstrate that this model of zero-shot learning has viability for the task, and that our assumptions about a sequence of visual processing stages mapping to a range of feature sets have an empirical basis. That said there is still much room to improve predictions. Future work could involve pushing the responsibility of selecting useful features elsewhere, such as to the distance metric. We would expect this to improve the poor performance of our semantic set and combination sets. We may also gain some performance in optimising some of the parameterisation of our image feature extraction, or trying other feature sets more closely related to neural processes. It may also be worth searching for non-linear relationships.

References


