From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI

Jalees Jahanzaib (2018101001)
Avani Gupta (2019121004)
Mayank Goyal (2018101002)

May 7, 2022

Abstract

The main goal of this project is to reconstruct observed images from fMRI brain recordings. Let's consider a human subject sitting inside a brain scanner and they are presented with a bunch of images and their brain readings are recorded throughout the session. In the time being they are presented to bunch of images and their corresponding fMRI. The task is given a fMRI of never-before seen image, to be able to reconstruct image closer to the original/real one. If we achieve this task then it will be a milestone in brain machine interfaces for example In the field of healthcare alone, Brain Machine Interface can be employed to prevent, detect and diagnose, and rehabilitate and restore from a sickness. This will also helps in developing ability to reconstruct dreams or serves as a scientific tool for studying consciousness.

Figure 1:

Introduction

In this task, developing a method for high-quality reconstruction of seen images from the corresponding brain activity, one attempts to solve for the mapping between fMRI recordings and their corresponding natural images, using many "labeled"Image, fMRIpairs (i.e., images and their corresponding fMRI responses). A good fMRI-to-image decoder is one that will generalize well to reconstruction of new never-before-seen images from new fMRI recordings (we refer to these as "test-data" or "test-fMRI"). However, there are several challenges one must accomplish to achieve a good fMRI-to-image decoder.

Challenges:

- First one is **lack in "labelled" training data** which limits the generalization power of today's fMRI decoders. Due to the limited time a human can spend in the scanner, it is not possible to generate a large number of labelled pairs (Image, fMRI).
- The **poor spatio-temporal resolution** of fMRI signals, as well as their **low Signal-to-Noise Ratio (SNR)**, reduce the reliability of the already scarce labeled training data.
- The train-set and test-set of the fMRI data often differ in their statistical properties, specifically in their SNR. This SNR discrepancy is due to averaging a different number of repeated recordings per image (typical of many fMRI data sets). It therefore introduces an additional challenge of **'domain transfer/adaptation'**, which makes generalization even harder, and affects the performance of current decoding methods.

Prior Work

The task of reconstructing a visual stimulus from fMRI has been approached by a number of methods which can broadly be classified in the following families:

- Linear regression between fMRI data and handcrafted image-features.
- Linear regression between fMRI data and Deep image-features.
- End-to-end Deep Learning

The first two regression-based methods compute a linear model that relates fMRI voxels to image feature representation. This can be done by either linearly predicting each voxel's responses from the image features, or by linearly mapping voxel responses to image features from which the image can be easily recovered. Second model improved the first one by utilizing the data-driven learned features from leading CNN models trained for natural image classification. The last one refers to recent attempts to train high-complexity deep models which directly decode an fMRI recording into its corresponding image stimulus

Figure 2: Prior Work

All these methods are supervised hence suffer from very limited training data and they are also prone to poor generalization to new test data.

Self-Supervision

Now one can overcome the challenges such as inherent lack of training data and the discrepancy between the training and testing set SNR(signal-to-noise ratio), by introducing **self supervision** using unlabelled data.

Figure 3: Self-Supervision

We train two types of networks: a Decoder denoted by D, to map fMRI recordings to their corresponding images and an Encoder E, to map natural images to their corresponding fMRI response. Now in self-supervision, those two networks are concatenated back-to-back, E-D, yields a combined network whose input and output are the same image(top-right). This allows for unsupervised training on unlabeled images (i.e., images without fMRI recordings, e.g., 50,000 randomly sampled natural images in our experiments). Such self-supervision adapts the network to the statistics of never before-seen images. Moreover, concatenating the two networks the other way around, D-E, yields a combined network with the same shared weights

as E-D, but whose input and output are now an fMRI signal (bottom right). This allows unsupervised training on unlabeled fMRI. Specifically, those unlabeled fMRI samples can be legitimately drawn from the test-fMRI set, while their corresponding images ("test-images") are excluded from training.

Effect of Self-Supervision

Figure 4: Effect of Self-Supervision

In this figure we can see the power of adding self-supervision approach i.e unsupervised training on unlabelled data. Notably it was found the unsupervised training on the unlabeled test-fMRI to provide the greatest boost in the performance. Learning the statistics of the test-fMRI directly addresses the lack in labeled training data and the train/test set SNR discrepancy.

Training on these unlabeled test-fMRI (without their images) is the key feature. Note that here "training on the test data" may seem "illegal" at first sight, but in fact is valid as it refers only to training on unlabeled samples from the decoder's input space whereas the test images(labels) are never used at any stage of the training.

Methodology

Before discussing the methodology, let's first discuss the two important models which are building block for our model, namely, encoder and decoder.

Two Deep Models: Encoder and Decoder

The encoder is extracting pre-trained features from Alexnet then there are bunch of convolutions and sub-samplings in the end it takes an image blob to fully connected layer that outputs a vectorized fMRI form.

Similarly there's the decoder which starts with a fully connected layer and then there is bunch of convolutions with up sampling until you get the image in the end.

Figure 5: Encoder and Decoder

Training Encoder and Decoder

Our training consists of two phases.

Phase 1

In the first phase, we apply supervised training of the Encoder E alone. We train it to predict the fMRI responses of input images using the image-fMRI training pairs.

Figure 6: Encoder Training

Let r' denote the encoded fMRI response from s by Encoder E. The fMRI loss is defined by a convex function of mean square error and cosine proximity with respect to the ground truth fMRI, r. The **fMRI loss** is define as:

$$L_r(r^{'}, r) = \alpha||r^{'} - r||_2 - (1-\alpha)\cos(\angle(r^{'}, r)) \tag{1}$$

where α is a hyper-parameter set empirically.

We will use this loss to train our encoder E.

Phase 2

In the second phase, we use the pre-trained Encoder (from the first phase) and train the Decoder D, keeping the weights of E fixed. D is trained jointly using both the labeled and the unlabeled data, simultaneously. Each training batch consists of three types of training data: (i) labeled image-fMRI pairs from the training set (Fig 7a), (ii) unlabeled natural images (Fig 7b), and (iii) unlabeled fMRI (Fig 7c). The motivation for using two training phases is to allow the Encoder to converge at the first phase, and then serve as strong guidance for the more severely ill-posed decoding task, which is the focus of the second phase. The weights of the Encoder are kept fixed during the Decoder training, to ensure that the Encoder's output representation does not diverge from predicting fMRI responses by the unsupervised training objectives.

Figure 7: Decoder Training

The training loss of our decoder consists of three main losses

$$L^D + L^{ED} + L^{DE} \tag{2}$$

L^D is a supervised loss on training pairs of image-fMRI.L^{ED} and L^{DE} are unsupervised losses on unlabeled images (without fMRIs) and unlabeled fMRIs (without images).

Decoder Supervised Training L^D

Given training pairs(r,s), the supervised loss L^D is applied on the decoded stimulus,s'=D(r), and is defined via the image reconstruction objective, L_s, as

$$L^D = L_s(s', s) \tag{3}$$

L_s consists of losses on image RGB values,L_{RGB}, and its features, $L_{features}$. We denote the features extracted from an image,s, by $\phi(s)$, and chose ϕ to be pre-trained a feature-extractor. Specifically we used activations from the first and the second convolutional layers of VGG19[3].The Image loss for a reconstructed image s' reads:

Figure 8: Decoder Supervised Training

$$L_s(s',s) = L_{RGB}(s',s) + L_{features}(s',s) + R(s') \tag{4}$$

$$L_{RGB}(s',s) \propto ||s' - s||_1$$

$$L_{features}(s',s) \propto ||\phi(s') - \phi(s)||_2$$

$$R(s') \propto TV(s')$$

The last term corresponds to total variation (TV) regularization of the reconstructed image, s' = D(r). It is introduced for ensuring spatial continuity and smoothness in the generated image by avoiding noisy and overly pixelated results.

Encoder Decoder Training on unlabeled Natural Images L^{ED}

Figure 9: Encoder Decoder Training

This objective enables to train on any desired unlabeled image, well beyond the 1200 images included in the main fMRI dataset. To train on images without corresponding fMRI responses, we map images to themselves through Encoder-Decoder transformation,

$$s \hookrightarrow s'_{ED} = D(E(s))$$

The unsupervised component L^{ED} of the loss in Eq. 2 on unlabeled images, s, reads:

$$L^{ED} = L_s(s'_{ED}, s)$$

Figure 10: Decoder Encoder Training

where L_s is the loss defined in Eq. 4.

Decoder-Encoder Training on unlabeled test fMRI L^{DE}

Adding this objective greatly improved our reconstruction quality compared to training on paired samples only. To train on fMRI data without corresponding images, we map an fMRI response to itself through Decoder-Encoder transformation:

$$r \hookrightarrow r'_{ED} = E(D(r))$$

The unsupervised component L^{DE} of the loss in Eq. 2 on unlabeled fMRI, r, reads:

$$L^{DE} = L_r(r'_{DE}, r)$$

where L_r is the loss defined in Eq. 1.

Implementation Details

Experimental Datasets

We experimented with a publicly available benchmark fMRI dataset **fMRI on ImageNet**[2] This dataset provide fMRI recordings paired with their corresponding underlying images. Subjects were instructed to fixate at a cross located at center of the presented images.'fMRI on ImageNet'comprises 1250 distinct ImageNet images drawn from 200 selected categories. The train- and test-fMRI data consist of 1 and 35 (repeated recordings) per presented stimulus, respectively.50 image categories provided the 50 test images, one from each category. The remaining 1200 were defined as train set (with only one fMRI recording).

We screened approximately 8500 out of the 50K recorded voxels by their SNR. We used additional 50K unlabeled natural images from ImageNet[1] validation data for the unsupervised training on unlabeled images (Encoder-Decoder objective, Fig. 1d). We verified that the images in our additional unlabeled external dataset, are distinct from those in the "fMRI on ImageNet".

Performance Evaluation

Similarity measure was based on correlating pixel values between the reconstructed image and an original image

Figure 11: 2-way

Identification accuracy: objective image-reconstruction quality measure by computing its correct-identification rate in a multi-image identification task: For each reconstructed image, the task is to identify its ground truth image among n candidate images (n= 2, 5 or 10), one being the true ground truth, while the rest were randomly selected. This identification was based on the same correlation measure (between the reconstructed image and each candidate image). The candidate image which scored the max Pearson correlation is determined to be the identified ground truth.

Training Details

- We train encoder(image to fMRI) for 80 epochs, freeze it then train decoder(fMRI to image) for 150 epochs.
- Decoder is trained jointly on a mix of labelled and unlabelled data with 48 paired and 16 unpaired samples.
- We have a training time of 30-45 mins for encoder(on 1 1080 RTX GPU) and nearly 3-5 hours for decoder on 2 1080 RTX GPUs)
- Encoder is trained using SGD optimizer with an initial learning rate of 0.1, with a predefined learning rate scheduler.
- Decoder is trained for 150 epochs using Adam optimizer with an initial learning rate of 1e-3, and 80% learning rate drop after every 30 epochs

Experimental Results

Figure 12: Training

Figure 13: (a) swan (b) deer (c) fish (d) sitar

Comparison with Paper

Figure 14: Ours

Figure 15: Paper

Some key observations between the 2 shown results:

- swan(row=1,column=5) has better shape in our case
- crab(row=2,column=2) has better shape in our case but original color in paper's result
- glass(row=4,column=3) has better shape in our case
- boat with people(row=5,column=2) has better shape in paper's case

while in other cases results are almost similar.

Evaluation Metric

Correct Identification Accuracy(CIA)

For each reconstructed image, the task is to identify its ground truth image among n candidate images (n= 2, 5 or 10).

Correct Identification accuracy is calculated as follows:

- Given Prediction image, choose 10 random images
- Now find pearson correlation(corr) between Prediction and random images.
- The one with highest corr is identified as matching image.
- Check if matching image is GT(ground truth), if yes, correct, else wrong.
- Calculate Accuracy

CIA comparison

	2-way	5-way	10-way
Ours	79.8	60.3	48.6
Paper	83.1	66.4	51.2

We get more or less similar results to paper.

Conclusions

We implemented the paper "From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI" and reproduced its results. We did additional experimentation by playing with losses, model architecture, training only with supervision(paired examples) and reducing data for Self-supervised training. We found the paper's hyper-parameters and architecture to perform the best.

References

[1] R. Socher L.-J. Li Kai Li Deng, W. Dong and Li Fei-Fei. *"ImageNet: A large-scale hierarchical imagedatabase," in2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, IEEE, 62009*.

[2] T. Horikawa and Y. Kamitani. *Generic decoding of seen and imagined objects using hierarchical visual features, Nature Communications, vol.8 p. 15037*. October 2015.

[3] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. September 2014.