Image Reconstruction from fMRI Data

Rudransh Agarwal
IIIT Hyderabad

Akshit Sharma
IIIT Hyderabad

***Abstract*—Perceptual experience consists of an enormous number of possible states. Previous fMRI studies have predicted a perceptual state by classifying brain activity into pre-specified categories. Constraint free visual image reconstruction is more challenging, as it is impractical to specify brain activity for all possible images. We start off with implementing and extending the technique of visual image reconstruction from human brain activity introduced in Miyawaki et al. (2008) paper. In another approach, we directly trained a DNN model with fMRI data and the corresponding stimulus images to build an end-to-end reconstruction model. We accomplished this by training a generative adversarial network (GAN) with an additional loss term that was defined in high-level feature space. We have also given an ROI wise analysis, to show which ROIs are responsible for encoding specific features of a visual stimulus.**

I. Interim Submission

A. Introduction

In Miyawaki et al. (2008) paper, they reconstructed visual images by combining local image bases of multiple scales, whose contrasts were independently decoded from fMRI activity by automatically selecting relevant voxels and exploiting their correlated patterns. Binary-contrast, 10x10-patch images (2^{100} possible states) were accurately reconstructed without any image prior on a single trial or volume basis by measuring brain activity only for several hundred random images.

B. Dataset

For the interim submission we have used the Binary Contrast Pattern dataset from Miyawaki et.al. which consists of visual stimuli that are just binary images.

C. Procedure

In this method, the author presents an approach to visual image reconstruction using multivoxel patterns of fMRI signals and multiscale visual representation (Fig. 1). They assume that an image is represented by a linear combination of local image elements of multiple scales (colored rectangles). The stimulus state at each local element $(C_i, C_j, ...)$ is predicted by a decoder using multivoxel patterns (weight set for each decoder, $w_i, w_j, ...$), and then the outputs of all the local decoders are combined in a statistically optimal way (combination coefficient, $\lambda_i, \lambda_j, ...$) to reconstruct the presented image. Only the voxels in the early visual cortex, that is the V1 and V2 regions are taken into account. The weights for the local elements are calculated using a suitable linear regression model with feature selection. The image is then reconstructed by a weighted sum of all the the elements that we have calculated.

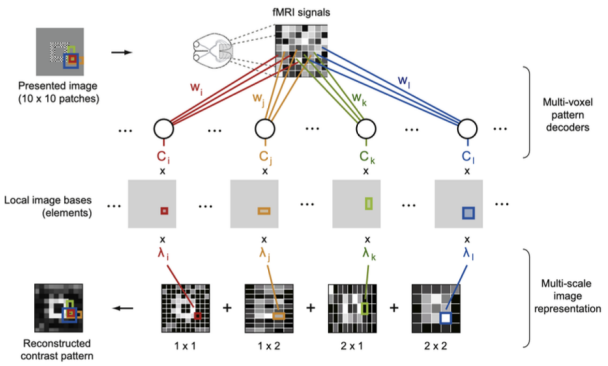

Fig. 1. Reconstruction Procedure

D. Extending this Method

While the nilearn implementation uses Orthogonal Matching Pursuit(OMP) for calculating the weights, we have tried different models with the likes of linear regression and bayesian ridge regression. OMP performs the best becasuse of the feature selection associcated with it (takes only the features with weights which are not zero). While the Nilearn code assigns fixed weights to all the decoders(0.25), we have trained the weights for each decoder using SGD. This is done because while performing MVPA, the contribution of a single voxel pattern might be different from a shared voxel pattern. Furthermore the paper only suggests a multi-scale pattern of (1*1, 1*2, 2*1 and 2*2) whereas we have tried various other patterns as well like (1*1, 1*3, 3*1 and 3*3) and (1*1,1*5, 5*1, 5*5). We have done this because consideration of voxels which are farther apart might also effect the reconstruction. As a baseline, we have also experimented with a univariate analysis.

E. Results and Observations

We present the results for each of these cases in terms of metrics like accuracy of reconstructed image along with precision, recall and F1 scores. The results include a direct side-by-side image comparison, showing the original stimulus image, the reconstructed image we get and the binarised image (obtained by making pixel values 0 or 1 by using cutoff at 0.5).

*1) **Comparing the obtained results**:*

- **Multi-scale reconstruction using only (1*1) vs. (1*1, 1*2, 2*1 and 2*2) images**
  We observe that the accuracy without multi-scale reconstruction i.e. with only (1*1) image gives an accuracy

score of 75.1% whereas the multi-scale reconstruction method using the 4 images i.e. (1*1, 1*2, 2*1 and 2*2) gives 80.2% score, which is a significant improvement (Fig. 2 and 3).

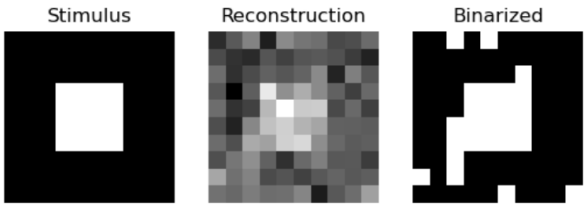

Fig. 2. (1*1), Accuracy: 75.1%

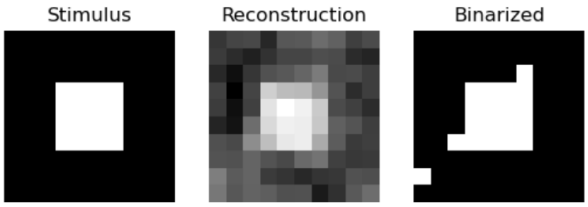

Fig. 3. (1*1, 1*2, 2*1 and 2*2), Accuracy: 80.2%

- **Multi-scale reconstruction using OMP vs. Linear-Regressor vs. BayesianRidge**
  We observe that OMP (Fig. 4) gives an accuracy of 80.2%, Linear Regressor (Fig. 5) gives 72.9% and Bayesian Ridge (Fig. 6) gives 76.1%. We can also see that qualitatively, the reconstruction provided by OMP is much better compared to the other two. The clear low accuracy of linear regression is because it provides for no feature selection, with the fMRI data, we can notice that many features would just be redundant and not carry high weights. Between OMP and Bayesian Ridge, OMP let's us choose the number of significant coefficients that we want to use, thus we can omit the zero weights, hence OMP clearly outperforms the other two.

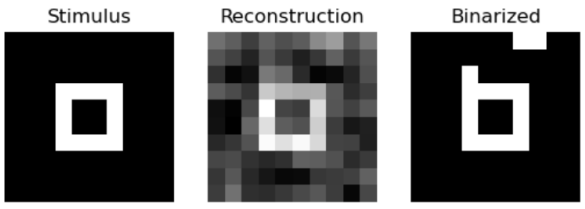

Fig. 4. OMP, Accuracy: 80.2%

- **MSR using (1*1, 1*2, 2*1 and 2*2) vs. (1*1, 1*3, 3*1 and 3*3) vs. (1*1, 1*5, 5*1 and 5*5)**
  We observe that accuracy of voxel pattern with 3(83.2%) and 5(82.1%) are better than with 2(80.2%). Although qualitatively, the reconstruction in Fig. 9 with patterns

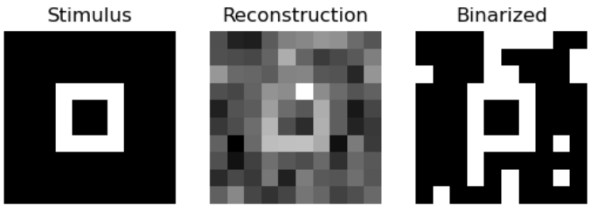

Fig. 5. LinearRegressor, Accuracy: 72.9%

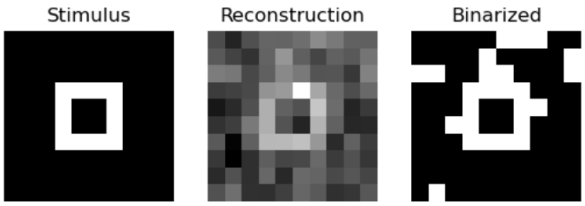

Fig. 6. BayesianRidge, Accuracy: 76.1%

of 5 is better than that of 3 in Fig. 8 and that of 3 is better that of 2 in Fig 7. We can observe the opposite effect in Fig. 10, 11 and 12. That is, the reconstruction with pattern of 2 is better than that of pattern with 3 and 5. Thus, we notice that when it comes to simple patterns with lesser variations in adjacent pixels, a higher window helps while with a complex pattern, with more variation in adjacent pixels, a smaller window performs better.

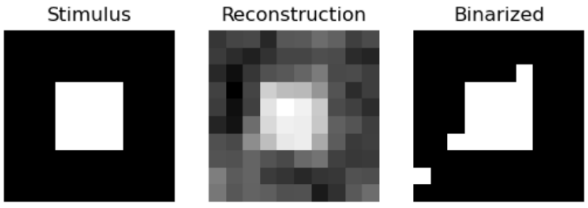

Fig. 7. (1*1, 1*2, 2*1 and 2*2), Accuracy: 80.2%

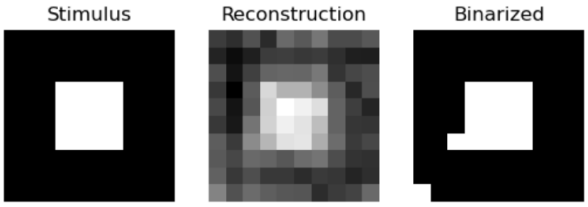

Fig. 8. (1*1, 1*3, 3*1 and 3*3), Accuracy: 83.2%

*2) **Analysing weights assigned to each scale in multi-scale reconstruction**:* As indicated in the Figure 12, we train multiple scales and take weighted sum of these to get our reconstructed image. We have used SGD in order to train weights for each scale. We have reported the corresponding weights in the Table 1. As we can see, as the scale x increases,

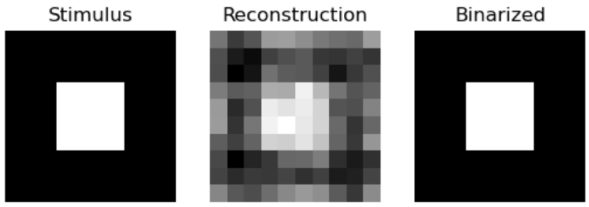

Fig. 9. (1*1, 1*5, 5*1 and 5*5), Accuracy: 82.1%

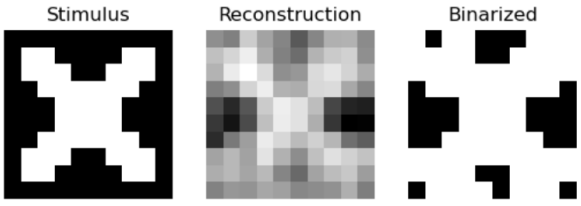

Fig. 10. (1*1, 1*2, 2*1 and 2*2), Accuracy: 80.2%

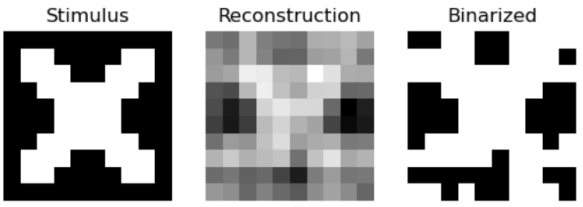

Fig. 11. (1*1, 1*3, 3*1 and 3*3), Accuracy: 83.2%

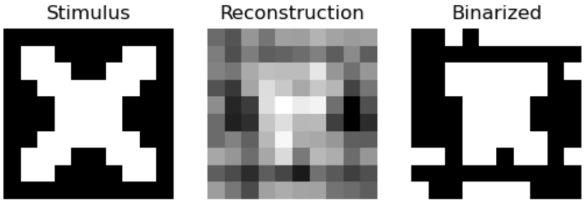

Fig. 12. (1*1, 1*5, 5*1 and 5*5), Accuracy: 82.1%

the importance of the x*x scales decreases significantly, and 1*1 increases significantly.

Scale(x)	1*1	1*x	x*1	x*x
2	0.18	0.34	0.27	0.19
3	0.21	0.33	0.26	0.18
5	0.25	0.32	0.25	0.16

TABLE I
TRAINED λ VALUES (WEIGHTS) FOR EACH SCALE IN THE 3 CASES FOR MULTI-SCALE RECONSTRUCTION

II. Final Submission

A. Dataset

We have used the Deep Image reconstruction dataset from Shen et. al The train images consist of 1200 set of images repeated 5 times, thus creating 6000 fMRI samples in total. The test images here are a set of 50 images each repeated 24 times, thus making a total of 1200 images.

B. Procedure

The idea here is to train an end-to-end reconstruction network which takes fmri as input and outputs the reconstructed image. For this, we train a GAN architecture. Basically, each fmri is fed into the GAN, and it generates an image, the disriminator then tries to predict if this image is real or fake. The discriminator loss is thus dependent on the amount of images it can correctly classify as real or fake. The generator loss we have defined here in 3 different losses. One is the one related to discriminator, that is how much is the generator able to fool the discriminator, next is the image loss, this is the direct MSE loss between the original image and the reconstructed image. Lastly, we have the feature loss. We use a pretrained alexnet, network to generate features in the feature space, and then perform an mse loss in the feature space.

C. Results and Experiments

Initially, we reconstructed the images with the weights weights given in the paper for each loss term. We notice that due to varied data pre-processing, specifically, they treat images in int format and we consider images in float, the final loss is not very representative of all the losses. Thus, we conducted multiple experiments to figure out the best weights, and here we report the one with the best results. Finally, Image loss weight = 100, Feature loss weight = 100, Adverserial loss weight = 10000.

The best model we have trained with the changed weights, gives a Pearson correlation of 96% and SSIM score of 75% on the train set. On comparing the our reconstructions and the results from the paper, we find a consistent effect that although the shapes are preserved, the colors are not preserved. Fig 15 to 17 show the original image, their and our reconstruction.

All the results presented till now were solely done with the entire visual cortex. Next we try to do a ROI wise reconstruction to see if fmri data from certain ROIs could be used for reconstruction.

Reconstructions done with the lower visual cortex regions, V1 and V2 are almost comparable to the whole visual cortex (Fig. 15). This is probably because the lower visual cortex is essentially responsible for processing all the low level features like shapes, colors etc. and that level of information might be useful for reconstruction. In particular, as seen in the Fig 23, the lower visual cortex also performs very well on some natural scenes' color reconstruction as well, as it is responsible for color processing.

Reconstructions done with the higher visual cortex, and its sub regions, particularly, the PPA and FFA, are nowhere close to the previous reconstructions, even in terms of shapes etc. The pearson and SSIM are almost equal to half the ones from before. Fig 22 shows the image reconstructed from HVC, PPA and FFA compared to that from VC. This is probably because

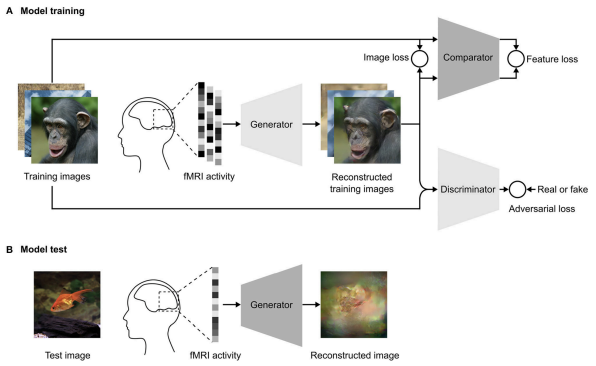

Fig. 13. Schematics of our reconstruction approach

$$L(\boldsymbol{\theta},\boldsymbol{\Phi}) = \lambda_{\text{img}} L_{\text{img}}(\boldsymbol{\theta}) + \lambda_{\text{feat}} L_{\text{feat}}(\boldsymbol{\theta}) + \lambda_{\text{adv}} L_{\text{adv}}(\boldsymbol{\theta},\boldsymbol{\Phi})$$

where

$$L_{\text{img}}(\boldsymbol{\theta}) = \sum_i \left|\left| \mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i) - \mathbf{X}_i \right|\right|_2^2$$

$$L_{\text{feat}}(\boldsymbol{\theta}) = \sum_i \left|\left| \mathbf{C}\left(\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i)\right) - \mathbf{C}\left(\mathbf{X}_i\right) \right|\right|_2^2$$

$$L_{\text{adv}}(\boldsymbol{\theta},\boldsymbol{\Phi}) = -\sum_i \log \mathbf{D}_{\boldsymbol{\Phi}}\left(\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i)\right)$$

Fig. 14. Generator Loss $L(\theta, \phi)$ as weighted sum of 3 different loss terms

HVC areas are specialized to perform only certain tasks, and mainly responsible to integrate information from various regions, like FFA is responsible for face recongition, PPA for natural scene representations, etc. Thus, the information encoded in these regions might not be enough for complete reconstruction of images. But they do perform well in certain cases as we define in the following regions

Also, qualitatively, we can see from image 23, 24 and 25, the best model performs much better on artificial objects, with well defined shapes, and very less details otherwise. Like in fig 23, the entire shape of the chameleon is captured but other details like stem and all are missing. But with fig 24 and 25 we can see that shape is captured very well, which is all there is to the image of the artificial object. While all the reconstructions shown till now were on the test images, we have shown, the Fig 18 shows reconstruction on the train images as well.

Fig. 15. Test set image reconstruction for V1V2 model

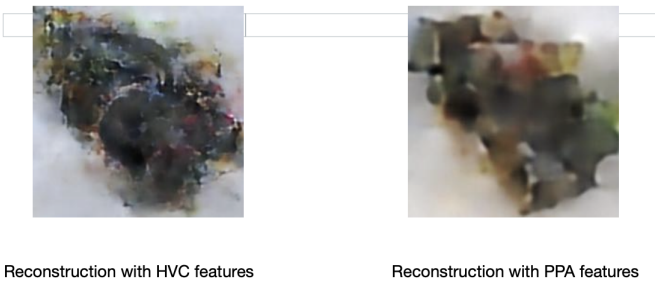

Fig. 16. HVC vs PPA Reconstruction

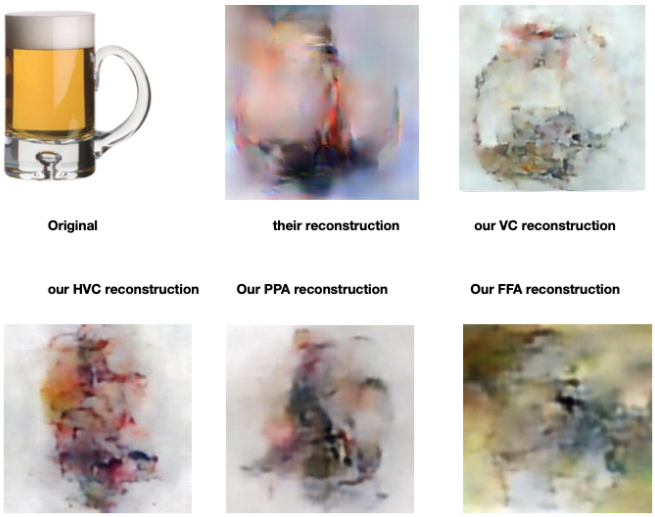

Fig. 17. Reconstruction of original image with different models

D. Ablations and Baseline

For the baseline model, we sample from a normal distribution and generate images from this instead of the fmri. As we can see in Fig 19, the images generated are random. The metrics can be seen in the Table 2. For the ablation analysis, we check the effect of each loss on the reconstruction. The metrics that we obtain on the test set for all the models we have trained are shown in the table 2. We see that the image loss (MSE) is the most important loss term in the generator's loss, as the metrics obtained after removing image loss are the worst. Also, the generated images from the image loss seem completely comprehensive and random. For the other two losses, the new metrics are still lesser than the best model. But we also notice that the reconstruction are still not very clear and definitive by the removal of these losses (Fig. 20-22).

E. Novelty

While the the original paper uses Caffenet, we tried out various models for the comparator. We found that Alexnet worked best and gave a improvement from 0.26 for their model to 0.30 for our model.

Next we tried to do a ROI wise analysis. Specifically when the reconstruction was done for V1V2, we found that there

Fig. 18. Train set image reconstruction with our best model

was considerable improvement in the colors. As we can see in the Fig 15. This is natural because the V1V2 regions are responsible for basic image processing like colors. Next, for the HVC(Higher visual cortex) we notice that the metrics are considerably low, because lower level features are all absent here. But some higher level features are clear. Like we can see the artificial objects are constructed much better as we would expect from the HVC. Even with PPA, we notice the same thing (Fig. 16). Although, we had fMRI features data for FFA, there are no images of faces to be reconstructed.

Model	Test PC	Test SSIM
Baseline	0.09	0.18
their	0.26	0.25
Our best model	0.27	0.30
V1V2	0.27	0.24
HVC	0.10	0.20
FFA	0.11	0.20
PPA	0.10	0.20
Without Feature Loss	0.193	0.215
Without Adversarial Loss	0.202	0.208
Without Image Loss	-0.04	0.04

TABLE II
TEST SET METRICS FOR DIFFERENT MODELS

Fig. 19. Baseline Model

F. Future Perspectives

We have gone through two papers for this task, which were recently published. The first one is a Nature paper titled **'Natural scenes reconstruction from Generative Latent Diffusion'**. It presents a two-stage scene reconstruction framework called "Brain-Diffuser". In the first stage, starting from fMRI signals, they reconstruct images that capture low-level properties and overall layout using a VDVAE (Very Deep Variational Autoencoder) model. In the second stage, they use the image-to-image framework of a latent diffusion model (Versatile Diffusion) conditioned on predicted multi-model (text and visual) features, to generate final reconstructed images. The other one it titled **'Instance Conditioned GAN'**. It takes inspiration from kernel density estimation techniques and introduce a non-parametric approach to modeling distributions of complex datasets.

Fig. 20. Removing Feature Loss

Fig. 21. Removing adversarial Loss

III. REFERENCES

- End-to-End Deep Image Reconstruction from Human Brain Activity
- Miyawaki et. al. (2008)
- Deep Image Reconstruction from Human Brain Activity
- Natural scene reconstruction from fMRI signals using generative latent diffusion
- Instance-Conditioned GAN

Fig. 22. Removing Image Loss

Fig. 23. Our weight adjusted (best) model on natural image

Fig. 24. Our weight adjusted (best) model on artificial object image

Fig. 25. Our weight adjusted (best) model on artificial object image