This document is a mock NIH grant proposal for the course BIOE1586 - Quantative Systems Neuroscience at the University of Pittsburgh in Spring 2020.

Reconstruction of Visual Patterns From V4 Firing and Local Field Potentials

Ziyi Gong

Specific Aims

Primate ventral visual stream can be viewed as a hierarchical feature extraction process. One layer of this hierarchy, visual area V4 in macaques is sensitive to curvature and shapes of intermediate complexity [1]. Moreover, previous studies showed that both the firing rates and local field potentials (LFPs) in V4 provide retinotopic information of curvature and complex shapes [2] [3]. Bashivan *et al.* successfully synthesized patterns that selectively elevated the firing rates of subgroups of V4 neurons [4], manifesting that there exists some stable mapping from visual patterns to V4 subgroup firing activities.

It is probable that an inverse mapping also exists, given the findings stated above that firing activities and LFPs in V4 sites carry specific shape information. Previous attempts on mapping the Blood Oxygen Level-Dependent (BOLD) signals of early visual areas as well as V4 [5] [6] to visual inputs also indicate such possibility, as correlation exists between BOLD signals and electrophysiological activities [7]. I hereby propose to construct a generative framework that maps LFPs and firing patterns in V4 sites to the latent features of the corresponding visual input in their receptive fields, and decodes the latent features to generate visual patterns. Using this generative model, I plan to determine 1) if the mapping is stable when small perturbation is added to the original input, 2) if the mapping can be used to evoke intended responses in V4 subgroups, and 3) if the mapping generalizes well in the cases of novel visual patterns.

Aim 1: Does visual patterns generated from slightly perturbed input cause similar V4 responses as the original?

This is to test the model's tolerance to perturbation. I hypothesize that on average visual patterns generated from slightly perturbed input does not lead to significantly different V4 responses as the original. I will test this hypothesis by applying Gaussian noise or Poisson noise with different mean (and different variance for Gaussian noise) to the recorded V4 spiking activities and LFPs, and use the perturbed information to generate images. Then the responses evoked by the generated visual patterns will be compared with the original ones using cross-correlation. Statistical testing will be performed on the cross correlation coefficients.

Aim 2: Does visual patterns synthesized from arbitrarily-chosen, biologically-plausible firing patterns and LFPs lead to similar V4 responses as the generated responses?

I hypothesize that there is a significant similarity between the actual V4 responses and the designed, biologically-plausible responses. The hypothesis will be tested with a generator trained on spiking activities and LFPs. The generator will generate V4 responses for image synthesis, which will be used for evoking the actual responses. It is also possible to pick such input without a generator, by performing a linear combination of the data in real trials. The similarity between the actual and artificial responses will be measured using statistical testing on their cross correlation coefficients.

Aim 3: Does the mapping associate V4 responses evoked by novel patterns with the latent features of similar patterns?

I hypothesize that the mapping associates V4 responses evoked by novel patterns with the latent features of the similar patterns. To test this hypothesis, first, novel patterns will be selected from separate categories or synthesized by maximizing their dissimilarity with the images used for training. After using t-test to confirm the novelty of the synthesized patterns, the images will be presented to the macaques and the V4 responses will be used to synthesize images. Statistical testing will be run between the presented images and decoder-synthesized images.

Significance

How primate ventral stream processes visual information has been a popular topic in academia. In the 1960s, Hubel and Wiesel discovered the hierarchical structures in the early visual areas, where the simple cells have distinct bar-shaped excitatory and inhibitory regions in their receptive fields; the complex cells, on the other hand, do not have distinct receptive fields but are tuned to edge angles and widths, similar to the simple cells [8]. Since then, researchers have been investigating the receptive fields or best-tuned patterns in higher visual areas, such as V4 and MT. Though the receptive fields of V4 are not clear, Nandy *et al.* studied the shape tuning in V4 by altering the orientation, position, and shape of a structure composed of 3 linked straight bars [2]. The shape could be from straight to "C" shape, controlling the local curvature. They identified a trade-off between translation invariance and curvature selectivity, as straight and low-curvature tuned neurons are invariant to translation but not orientation, while high-curvature tuned neurons do the reverse. The study furthermore performed analysis of V4 shape tuning in-depth by fitting a predictive model. In addition to firing rates, local field potentials (LFPs) in V4 were also demonstrated to contain retinotopic components [3]. These studies established the foundation for further investigation that V4 contains the information of certain kinds of shapes.

Recent years, emerging deep learning methods assist the investigation. Bashivan *et al.* created a connection between deep convolutional neural network (DCNN) and macaque V4 sites that can both predict the responses of V4 neurons to some given visual patterns and synthesize visual patterns from population control in V4 [4]. Noticing the similarity between AlexNet, a type of DCNN, and the hierarchy from primate retina to V4, they firstly optimized a stable mapping from the Conv3 layer of AlexNet to V4 firing patterns. The mapping allowed the prediction of V4 firing patterns, which served as the feedback to find good visual patterns that can elevate V4 site firing rates selectively. The mapping can also generalize well to novel patterns, i.e. the patterns significantly different from those used for optimizing this mapping. Their achievement indicates that it is possible to find a stable mapping from visual inputs to V4 activities without fully modeling the complex connectome of primate visual stream.

Studying the system in a reverse manner, Shen *et al.* and VanRullen and Reddy used deep learning to reconstruct the visual scenes that the subjects observed [5] [6]. The former team applied a linear decoder that transformed the fMRI data from V1 to V4 after the subjects seeing some images into a pretrained DCNN's hidden layer representations. Then they essentially optimized the pixels of the reconstructed images by minimizing the reconstruction discrepancy in DCNN's hidden layer representations. To enhance the naturalness and capture nuances, they introduced a pretrained generator network $G(\mathbf{z})$ that effectively increased the quality of reconstruction. The tuned generator's weights were frozen and used to find the best latent vector $\mathbf{z}$ that minimized the discrepancies. Similarly, VanRullen and Reddy used a variational autoencoder combined with generative adversarial network (VAE-GAN) to help extract facial information from BOLD signals. They first fitted a linear function from the latent space of VAE-GAN, in which the features of seen faces are concisely represented, to the corresponding BOLD signals from several selected cerebral regions. During testing, they reconstructed the seen faces from the BOLD signals using an inverse mapping that was easily derived from the linear function. Both the novel studies show the strong generative capability of deep learning models and the possibility of finding a stable mapping from the brain to the input.

From the preliminary researches described, it is easy to notice the possibility to find such reverse mapping, i.e. reconstruction of seen patterns from V4 firing activities and LFPs. However, it is not clear what details in the original visual inputs are discarded during forward processing in the ventral stream. Thus, at least in the initial attempts, it is necessary to adapt a generative framework, such as GAN, to encode visual details that are conditioned by, or "added to," the information in V4 firing and LFPs. GAN, firstly proposed by Goodfellow *et al.*, includes two competing neural networks, a generator trying to generate fake samples and a discriminator trying to distinguish the fake from the real [9]. The two are trained together and getting better and better at their tasks, and finally the generator is able to generate fake samples with fine details. However, the vanilla GAN suffers from gradient problems, is hard to train and is incapable of conditional generation, so many variants have been developed. Instead of separately training a mapping from neuronal data to image latent features and an image generator, InfoGAN proposed by Chen *et al.* is one good end-to-end framework [10]. Like vanilla GAN, InfoGAN contains a generator and a discriminator, where the discriminator is trained on real and fake data and the generator is trained on the feedback of discriminator. But InfoGAN's generator also incorporates a condition, while the discriminator not only assesses the input images, but also performs a variational inference of the conditions.

During training, the generator need to yield fine samples whilst maximizing the mutual information between the samples and the given condition, and the discriminator should appropriately estimate the authenticity of the images and the corresponding conditions. In terms of the proposed research, the condition is the V4 eletrophysiological data. The model thus need to learn in general how to create realistic images specified by the firing and LFPs in V4. Other methods, such as the VAE-GAN used by VanRullen and Reddy, are also good candidates.

Innovation

High-resolution, realistic reconstruction of sensory inputs from various brain signals has only been emerging these years because of better hardware and the power of deep learning. Although there has been visual input reconstruction from fMRI BOLD signals, such work has yet to be done for V4 electrophysiological data, which contains information with high spatiotemporal resolution for better reconstruction results. On the other hand, several suitable deep learning frameworks' potentials in these fields have not been demonstrated. Successfully constructing such state-of-the-art model would benefit the future studies in visual neuroscience, deep learning, brain-computer interface, etc.

Moreover, the proposed study is not only for seeking such a model, as it should not be hard to find a mapping despite of its accuracy, but also an attempt to investigate the way how a good mapping works. After the model is tuned and validated to be effective and accurate to some degree, experiments on the generalizability, latent space interpolation, and tolerance to perturbation will be performed. By correctly reconstructing visual patterns from electrophysiological data, the model could have approached or possessed some components of the optimal reverse mapping. By conducting the three sets of experiments, we are able to have insight on the model's spatial organization of the visual patterns, as well as how the generated results are conditioned on V4 activities. It is also possible to discover or develop hypothesis on the characteristics of the ventral stream as a system.

Approach

Data Preparation and Model Construction

Healthy adult macaques trained on fixation will be selected to participate in the experiments. After the monkeys fixate at a small white dot on the center of a screen for 300 ms, a series of naturalistic or texture images will be shown, where each image will be shown for 100 ms and followed by a blank screen for equal amount of time. The images will be pulled from datasets of classified, preprocessed images, so that their genres, contents, and resolutions will be under control. In addition, an image should be shown more than once throughout the data collection period to avoid bias. Multiple sites in V4 in the left and right hemispheres of several macaques will be recorded using microelectrode arrays. The raw electrophysiological data will be filtered with different pass bands to separate LFP components and action potentials. The signals will be clipped, smoothed, down-sampled, and concatenated. At least ten thousand successful trials should be conducted in order to train and test a neural network.

As stated, there are multiple potential candidates, such as InfoGAN that generates images conditioned on the electrophysiological data [10], VAE-GAN with an additional mapping from electrophysiological data to image feature representations [5], etc. Because many of these models are not recurrent, to use V4 responses as data entails either a fixed window size or a preprocessing recurrent neural network such as long-short-term memory (LSTM) that can encode time series of different lengths [11]. In case of insufficient data, the networks could be trained on typical image generation task for some epochs before being trained on conditional image generation. To select the best model architecture and its hyperparameters, cross validations will be performed. After the model is well-trained, the following experiments could be conducted.

Aim 1: Does visual patterns generated from slightly perturbed input cause similar V4 responses as the original?

The generator should be tolerant to small perturbation, such as Gaussian noise and shot noise. To begin with, for signals corresponding to an image, a Gaussian noise mask will be generated directly from a Gaussian process

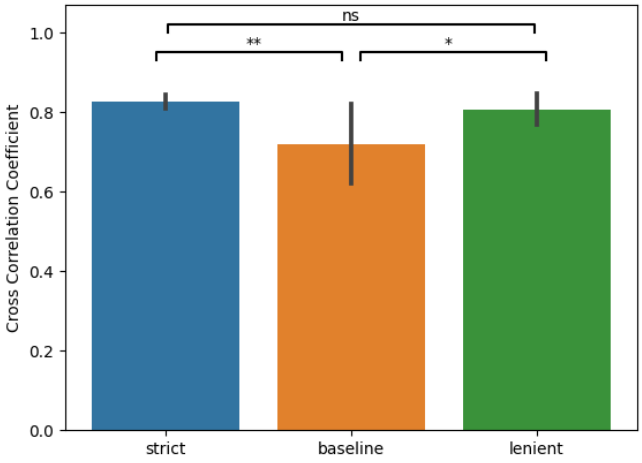

Figure 1: The mean cross correlation coefficient with respect to the average of the original signals. The difference between "strict" and "lenient" should not be significant. However, they should both be significantly higher than the baseline.

.

and a shot noise mask will be generated from a Poisson process based on the mean signal. Each parameter combination for the random processes must not lead to a significant difference between the perturbed signals and the mean signal for all time steps ($\max_t P_t < 0.05$, two-tailed one-sample t-test) to demonstrate that the perturbation does not significantly change the V4 encoding of an image. Then the perturbed signals will be used to synthesize visual patterns, which will be presented to the monkeys and V4 responses will be recorded. The experiment is named "strict" because in any time step the perturbation must not significantly change the magnitude. Since there could be some variations in the details of the images that lead to nuances in the responses, a baseline, calculated using the mean original signal and those for the other images, will be compared with the cross correlation coefficients between the mean original signal and the responses (one-tailed two-sample t-test). The coefficients should be significantly higher than the baseline.

It is also interesting to pay attention to the perturbed signals with significant difference only in some time steps ($\bar{P} < 0.05$, $\max_t P_t > 0.05$, two-tailed one-sample t-test). The experiment is named "lenient" because in some time step the perturbation can significantly change the magnitude. As the generator weights each time step differently, high perturbations in some intervals create more effects and in other intervals create less effects, so the variance should be slight larger than "strict." A "gradual effect" in the generated images is also expected in some of the samples. That is, some generated images will have changes that are slightly more noticeable than the results of the generator's randomness, but not significant to drastically change the contents of the images. In other words, the cross correlation coefficients may still be significantly higher than the baseline but not different from the group described above. The expected results can be represented in Fig. 1.

Aim 2: Does visual patterns synthesized from arbitrarily-chosen, biologically-plausible firing patterns and LFPs lead to similar V4 responses as the generated responses?

One way to have a better understanding of how the generator infers images from V4 responses is interpolation. There needs to be a thorough amount of generated LFPs and spike trains for image synthesis, and the images will be presented and the V4 responses will be recorded. I hypothesize that if, by comparing the means, the new generated LFPs and spike trains are in the domain of real samples, the paired t-test will show no significant difference between the evoked responses and the artificial ones. If the hypothesis holds, a tSNE visualization of the data on 2-dimensional space should involve no significant separation among the chosen, actual, and evoked V4 responses (Fig. 2, top). Furthermore, within each distribution, there could be multiple modes but they should be close to each other, i.e. the distribution is not sparse, if the hypothesis holds.

To generate data, the simplest approach is to arbitrarily design certain firing patterns and LFPs. However,

this may suffer from selection bias so that the interpolation is not thorough enough, or from human error which leads to biologically-implausible V4 activities. A more reliable way is to train a generative model that yields LFPs and spike trains. Molano-Mazon *et al.* trained a DCNN to generate realistic spike trains, which could be used for this task via transfer learning [12]. A recurrent neural network, combined with a discriminator, may also perform well or even better, since they are designed to deal with time series. The benefits of using a neural network generator instead of a computational model are that the former can explore some variables not included in the model and generate more diverse data.

Another less thorough but also informative approach is to take a proper linear combination of multiple LFPs and spike trains. Radford, Metz, and Chintala discovered that DCGAN learns image representation in an unsupervised manner and generating an image from the linear combination of several latent vectors leads to an intermediate result [13]. If the phenomenon also holds for this task and the synthesized images lead to similar responses (Fig. 2, bottom), then the model could be a powerful tool for investigating the V4 tuning in macaques. Specifically, one can have greater control over the contents of the reconstructed images, and learn about how the ventral stream as a system responses.

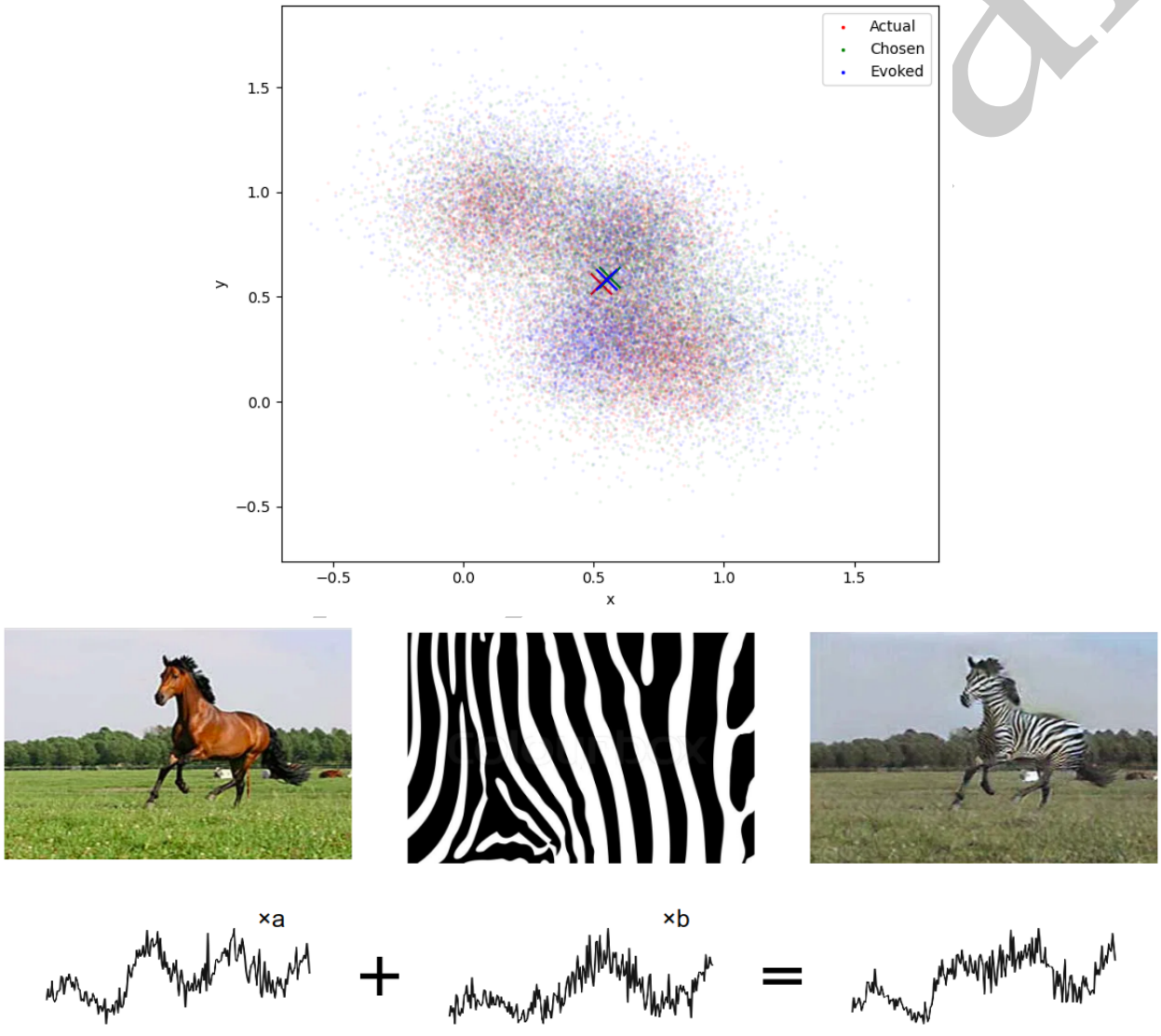

Figure 2: Top: tSNE visualization of the data. It is hypothesized that there should be no significant separations among the actual, chosen, and evoked responses. Note that the distributions could be Gaussian mixtures. Bottom: hypothesis that a reconstruction from a proper linear combination of two signals is the intermediate result of the two patterns the signals correspond to.

.

Aim 3: Does the mapping associate V4 responses evoked by novel patterns with the latent features of similar patterns?

It is also important to test the generative model's generalizability, an indicator of its potentials in future usage. To synthesize novel patterns, one straightforward method is to take a batch of images used for training and optimize the new images' pixel values such that the average distances between the new images and the used ones are maximized. To eliminate the effect of the nuances in the images, the distance is calculated using their hidden layer representation after passing the images to a pretrained DCNN on a typical classification task. However, it does not account for the variance of the training set, so an optimized image could be far away from the global mean but near a cluster center of the used images. A solution to this issue is to add a penalty of variance among the average pixel distances or the distances to the clusters, so that the synthesized images should be generally different from all images. The clustering can be done by encoding images to a lower dimension using an autoencoder, and then perform a clustering algorithm like KMeans. On the other hand, during training, a few distinctive classes from the image dataset can be reserved as the novel images, so there is no need to synthesize. Although it is simpler, the reserved images are more naturalistic, while the synthesized ones can have greater diversity. Either of the two approaches necessitates a t-test to assure that the images to be used are significantly different.

After the novel images are achieved and presented to the macaques, the V4 responses will be given to the generator to reconstruct the images. A paired t-test on the average distance will be used to assess the generalizability. Denote the distance between the reconstructed and the novel ones Δ_{rec} and the distance between the novel ones and the training set Δ_{new}. The Δ_{rec} values corresponding to p value thresholds will be used as the boundaries of the intervals where the dynamics between Δ_{rec} and Δ_{new} will be analyzed. The rate of change $d\Delta_{rec}/d\Delta_{new}$ with respect to Δ_{new} within a proper window can reflect the capacity of the model. There are multiple possible dynamics between Δ_{rec} and Δ_{new}. First, as Δ_{new} increases, Δ_{rec} experiences a sharp increase and then saturates at some high value. This may indicate that the is a clear boundary where the model cannot "comprehend" the data any more. Second, as Δ_{new} increases, Δ_{rec} experiences a plateau and then keeps increasing. This phenomenon possibly provides information about clusters of uncommon shapes, such that the distance metric works bad in this scenario. Third, $d\Delta_{rec}/d\Delta_{new}$ increases or decreases. If Δ_{rec} monotonically increases, then the generator gradually gets "unfamiliar" with the signals as their corresponding images become novel. However, ifΔ_{rec} finally decreases, it is more complex. Fourth, fluctuation occurs, indicating the representation could be complex, and the performance depends on the inputs. There could be other possibilities. In cases of all these situations, a dimension reduction and visualization technique will be used to assist in investigating how the neural network learn the mapping.

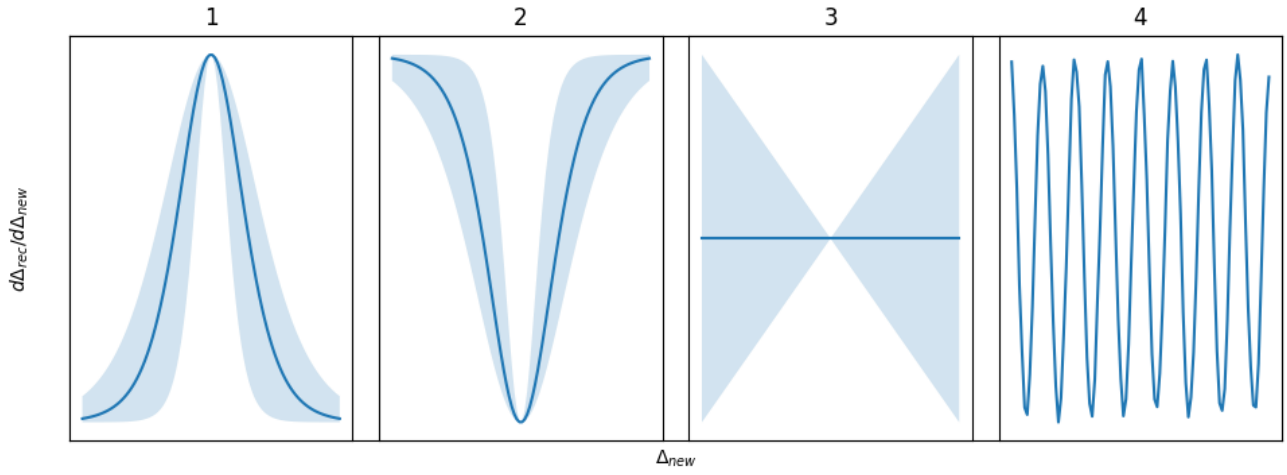

Figure 3: Multiple possibilities of Δ_{rec} and Δ_{new} dynamics. The curves manifest the trends, and the shade areas show potential variants. The variants of fluctuation are not shown due to complexity.

Summary

The study is a state-of-the-art approach to build a generative model that synthesizes seen images from V4 responses in macaques, test its tolerance to perturbation and generalizability, and investigate its learned representation through interpolation and visualization techniques. Determining whether the model is tolerant to perturbation, I will apply Gaussian and Poisson noise with different parameters to the actual signals, and use this perturbed information to generate images. Then the responses evoked by the generated visual patterns, treated as an encoding of the images, will be compared with the original ones. After testing the tolerance to perturbation, I will use a trained data generator and proper linear combination to arbitrarily generate biologically plausible signals. They will be used for image generations, and the evoked responses from those images will be compared with the generated data, and the original samples. The three groups' distributions should not be significantly separated. Finally, I will test whether the model generalizes well to novel data. I will synthesize the novel data by finding pixel combinations that maximize the distances with all image clusters used for training. The evoked responses will be used for reconstructing images, whose error will be used to assess the model's generalizability. The synthesized images will have distances with the training set, so as to see the relationship between the distance to the training samples and the reconstruction error, which could contain information about what the generator learns about the mapping.

Impact and Future Directions

The proposed model is an early attempt to explore the potentials of more complex, advanced deep learning models in brain signal understanding. The potential usage of this model can be in cortical signal decoding and theoretical research. Possible future directions can be to get higher level input into consideration, such as the interfere of cognition in cases of visual illusion, or develop equivalent models for other important higher visual cortices. It is also worth to enhance EEG understanding and decoding by using this model as the pretrained. In terms of deep learning, there could be more suitable frameworks for spiking and LFPs than the existing models. Exploring the alternatives may provide more robust models. Overall, the achievement could benefit future researches understanding the visual system or the nervous system in general, and enhances the algorithms in other relevant medical and BCI technologies.

References

[1] A. Pasupathy and C. E. Connor. Responses to contour features in macaque area V4. *Journal of Neurophysiology*, 82(5):2490–2502, 1999.

[2] A. Nandy, T. Sharpee, J. Reynolds, and J. Mitchell. The Fine Structure of Shape Tuning in Area V4. *Neuron*, 78(6):1102 – 1115, 2013.

[3] P. Mineault, T. Zanos, and C. Pack. Local field potentials reflect multiple spatial scales in V4. *Frontiers in Computational Neuroscience*, 7:21, 2013.

[4] P. Bashivan, K. Kar, and J. J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), 2019.

[5] R. VanRullen and L. Reddy. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications biology*, 2:193, 2019.

[6] G. Shen, T. Horikawa, K. Majima, and Y. Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):1–23, 01 2019.

[7] E. Tagliazucchi, F. Von Wegner, A. Morzelewski, V. Brodbeck, and H. Laufs. Dynamic BOLD functional connectivity in humans and its electrophysiological correlates. *Frontiers in Human Neuroscience*, 6:339, 2012.

[8] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks, 2014.

[10] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *CoRR*, abs/1606.03657, 2016.

[11] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[12] M. Molano-Mazon, A. Onken, E. Piasini*, and S. Panzeri*. Synthesizing realistic neural population activity patterns using Generative Adversarial Networks. In *International Conference on Learning Representations*, 2018.

[13] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2015.